THESE

pour obtenir

le grade de DOCTEUR

UNIVERSITÉ DE SAVOIE

Spécialité : *Electronique - Electrotechnique - Automatique*

et

UNIVERSITATEA „POLITEHNICA" BUCUREŞTI

Spécialité : *Ingénierie Electronique et Télécommunications*

M. Bogdan-Emanuel IONESCU

Caractérisation Symbolique de Séquences d'Images : Application aux Films d'Animation

---

Soutenue le 7 mai 2007 devant le jury composé de :

M. Teodor PETRESCU	Président
Mme Cornelia GORDAN	Rapporteur
Mme Michèle ROMBAUT	Rapporteur
Mme Christine FERNANDEZ-MALOIGNE	Examinateur
M. Constantin VERTAN	Examinateur
M. Patrick LAMBERT	Directeur de thèse
M. Vasile BUZULOIU	Directeur de thèse
M. Didier COQUIN	Co-Directeur de thèse

Thèse préparée au sein du Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (Annecy) et du Laboratoire d'Analyse et Traitement des Images (Bucarest).

Je dédie cette thèse à la mémoire de mon père prof.dr.ing. Dumitru-Cezar Ionescu qui nous a quitté trop tôt. C'est grâce à lui que j'ai pu m'engager dans ce domaine fascinant qu'est la recherche. Je dédie également mon travail à ma famille, à la meilleure mère du monde, Rodica, et à mon frère, Razvan.

Remerciements

Je tiens à exprimer toute ma reconnaissance à M. Vasile Buzuloiu, directeur du laboratoire LAPI de l'Université "Politehnica" de Bucarest, pour m'avoir accepté en thèse dans son équipe de Traitement des Images, pour son amitié et son encadrement précieux pendant la thèse. Je veux également adresser tous mes remerciements à M. Patrick Lambert et M. Didier Coquin pour leur encadrement, leurs nombreux conseils, leur soutien constant tout au long de ma thèse et leurs efforts pour corriger ce manuscrit.

J'exprime toute ma reconnaissance à M. Philippe Bolon, directeur du laboratoire LISTIC de l'Université de Savoie, pour m'avoir accueilli au sein de l'équipe de Traitement de l'Information et pour son soutien.

M. Teodor Petrescu m'a fait l'honneur d'accepter d'être président du jury. Je remercie Mmes Cornelia Gordan et Michèle Rombaut d'avoir accepté d'être rapporteur de ce travail, ainsi que pour leurs jugements très pertinents et très constructifs. Je remercie également Mme Christine Fernandez-Maloigne et M. Constantin Vertan, examinateurs de cette thèse, de leurs remarques et leurs nombreux commentaires très enrichissants.

Je remercie particulièrement M. Patrick Lambert de m'avoir aidé à résoudre mes divers problèmes administratifs et personnels rencontrés pendant mes différents séjours en France. Son implication dans ma formation a largement dépassé ses responsabilités en tant qu'encadrant et je lui en suis très reconnaissant.

Je remercie M. Corneliu Burileanu pour son amitié et son soutien professionnel et moral pendant toute ma formation scientifique.

Je tiens aussi à exprimer mes remerciements à tous les chercheurs et enseignants qui m'ont aidé pendant ces trois années de thèse, particulièrement :

- Mihai Ciuc pour son amitié et son soutien et Constantin Vertan pour son aide précieuse,
- Mrs. Adrian Badea et Ilie Prisecaru pour leur soutien,
- Daniel Beauchêne et Françoise Deloule les spécialistes des ontologies,
- Jean-Jacques Curtelin et Michel Cuny pour leur aide efficace en enseignement, Pascal Mouille pour son aide particulièrement précieuse pour résoudre les problèmes de visa,
- Sylvie Galichet, responsable de l'équipe Traitement de l'Information, Emmanuel Trouvé et Lionel Valet, expert en programmation en C,
- Eric Benoît pour sa collaboration et Gilles Mauris pour ses compétences en logique floue.

Je n'oublierai pas la disponibilité des personnels admnistratifs du LISTIC : Joelle Pellet, Valérie Braesch et Olivier Iund. Que soient également associés à ces remerciements Catherine Browne et le personnel du BLS d'Annecy ainsi qu'Eugenia Burcea de l'Université "Politehnica" de Bucarest.

Je remercie également la rédaction de la revue de "Politehnica" de Bucarest, Sci.Bull.

pour m'avoir aidé à la publication de quatre articles.

Je tiens aussi à remercier le personnel de CITIA, la Cité de l'Image en Mouvement d'Annecy, et plus particulièrement Daniel Bouillot, et la société Folimage pour nous avoir permis d'utiliser leurs films dans le cadre de ce travail.

J'exprime toute ma reconnaissance à la direction et au personnel de la résidence Pré Saint Jean pour leur aide et la bonne ambiance qu'ils ont su créer.

Je tiens aussi exprimer mes remerciements à tous mes amis ou collègues qui ont partagé mon quotidien pendant ces années : Lavinia, Filip, Gabi, Florentin, Andreea, Anthony, Ivan, Mishu, Andrei et Andreea, Alexandra, Puiu et Dana, Cristi, Laetitia, Charles, Greg, Serban Oprisescu, Ciubotaru Matei, Vladimir Popescu, Stefan Stancu, Laurent Ott, Hervé Combe, Sylvie Jullien, Jean Claude Jouffre.

Je remercie tout particulièrement Monica pour son aide et pour être constamment proche de moi.

Je termine par un grand merci à ma famille pour leur soutien : ma mère Rodica, mon frère Razvan, mes oncles Sorin et Eugen, mes deux tantes Silvia et ma tante Olga, et ma grand-mère Maria Ciurea.

Tous mes remerciements à ceux qui m'ont aidé et m'ont soutenu pendant cette thèse et que, sans le vouloir, j'ai oublié de mentioner ici.

Table des matières

Première partie

Introduction

Chapitre 1

Les systèmes d'indexation

Résumé : *Nous vivons dans un monde d'informations dans lequel le texte, l'image, le son, et la vidéo sont présents et font partie de notre quotidien. Le volume des informations et donc des données est devenu très important et touche de nombreux domaines. Accéder au contenu de ces informations est devenu un problème crucial. Une solution est un système basé sur l'indexation de ces données. Le but de ce chapitre introductif est d'abord de présenter un état de l'art sur les différentes directions utilisées par les systèmes d'indexation multimédia existants. Nous allons décrire les systèmes d'indexation, dans lesquels on trouve des techniques d'annotation du contenu, des moteurs de navigation et de recherche. Ensuite, nous présenterons la description globale du système d'analyse sémantique de séquences d'images que nous proposons dans cette thèse et conclurons sur son utilité dans différents domaines d'application.*

Nous vivons dans un monde où l'accès aux informations multimédia est devenu indispensable et fait partie de la vie quotidienne. Le volume des données : texte, image, son, vidéo est de plus en plus important suite à l'évolution des dispositifs de prise de vue (appareil photo et camescope numérique), de stockage (ordinateurs, serveurs) et au développement d'Internet. De ce fait, il existe maintenant de nombreuses bases qui regroupent des données généralement de même type mais dont le contenu est très varié. A cause du volume très important des données d'une base (plusieurs millions de Mo), retrouver une information ou accéder au contenu de ces informations sont des tâches difficiles. La vision par ordinateur et le traitement des images proposent des solutions pour résoudre ce problème, par le moyen des systèmes d'indexation de type *"content-based retrieval"* ou CBR (recherche par le contenu).

En étudiant le sens du concept d'indexation nous avons trouvé que le mot *"indexer"* est défini par : *lier les variations d'une valeur à celle d'un élément de référence, d'un indice déterminé* (utilisé en économie) [Robert 88]. Ce concept a été utilisé dans le domaine du traitement d'images pour définir la notion d'annotation ou d'indexation des données. Dans ce cas le mot "indexation" *définit le processus qui consiste à associer aux données des labels liés à leur contenu* [University 06]. L'indexation des données est ainsi une étape incontournable pour accéder aux informations dans une base de données. En conséquence, les informations

de la base qui ne sont pas indexées seront très difficiles à retrouver car les systèmes de recherche s'appuient sur les index.

L'indexation du contenu des données peut prendre deux formes principales : *l'annotation manuelle* ou *l'annotation automatique*. L'effort nécessaire pour annoter le contenu est directement proportionnel au niveau de détail désiré dans la recherche d'information. En effet, pour un niveau de détail élevé il faut une analyse plus importante du contenu des données. En général, mais cela varie selon la spécificité de l'application, les annotations manuelles sont lourdes et demandent beaucoup de ressources humaines. Les méthodes automatiques assistées par l'ordinateur sont beaucoup plus rapides car l'analyse du contenu ne demande pas l'intervention humaine. Cependant, elles ne pas toujours capables de fournir les mêmes informations que les méthodes manuelles, en particulier en ce qui concerne les index de haut niveau sémantique.

Pour accéder au contenu des données d'une base, l'étape d'indexation n'est pas suffisante. L'utilisateur doit avoir accès à des outils logiciels lui permettant d'interagir avec les données à l'aide d'une interface graphique ergonomique et agréable qui répondra principalement à deux besoins : *la navigation* et *la recherche*. Dans un système de navigation l'utilisateur peut facilement visualiser les données directement dans la base ou à travers des aperçus efficaces des contenus. D'autre part, dans le système de recherche, qui est souvent inclus dans le système de navigation, l'utilisateur peut rechercher les informations désirées dans la base en utilisant différents critères comme par exemple des informations sémantiques proches du langage humain. Le diagramme de la structure globale d'un système d'indexation est présenté dans la Figure 1.1.

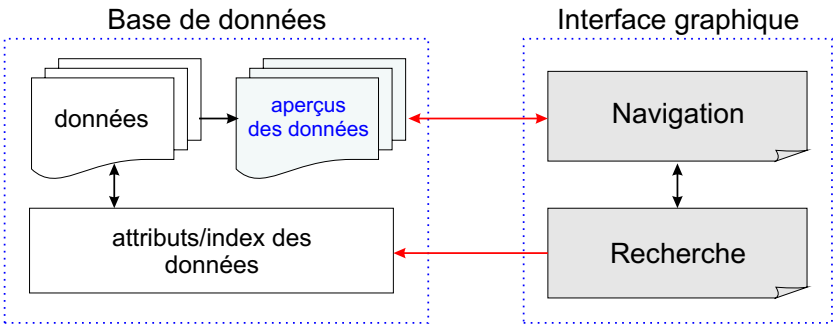


FIG. 1.1 – Le principe de fonctionnement d'un système d'indexation.

Le principe de fonctionnement est le suivant :

- **dans un premier temps** les données de la base sont analysées, hors-ligne, pour calculer des index qui sont typiquement des mesures numériques ou symboliques du contenu des données. En parallèle à cette étape, en fonction du type de données utilisées, une description compacte de chaque donnée peut être construite. Cette description compacte pourra servir comme aperçu du contenu, facilitant ainsi la tâche de navigation.
- **le deuxième temps** concerne la navigation et la recherche. La navigation se fait grâce à un outil logiciel qui permet l'accès rapide et efficace au contenu de la base à l'aide des aperçus déjà construits ou, selon la volonté de l'utilisateur, directement sur les données de la base. La recherche a pour but de retrouver dans la base des informations à partir d'une requête de l'utilisateur. De la même façon que pour la construction des index de la base, la requête est d'abord convertie par le système en index/attributs. La

recherche est effectuée ensuite par l'analyse de la similarité des index de la requête avec ceux déjà disponibles dans la base de données. Les données de la base les plus proches de la requête, au sens de la similarité des index, sont proposées à l'utilisateur comme résultat de la recherche. Pour améliorer les performances du système, une interaction de l'utilisateur sur les résultats obtenus est souvent utilisée. L'utilisateur donne son avis sur la qualité de la réponse du système, et une nouvelle recherche est effectuée prenant en compte l'avis de l'utilisateur. Cette technique d'apprentissage du système est appelée "boucle de pertinence" ("relevance feedback").

En conclusion, les systèmes d'indexation sont le résultat de travaux entrepris sur la manière d'accéder au contenu de données multimédia, données de plus en plus nombreuses et diversifiées.

Selon le type de données utilisées les systèmes d'indexation se divisent en :

- **CBIR** : systèmes d'indexation d'images où les données à traiter sont des images statiques ;
- **CBISR** : systèmes d'indexation de séquences d'images où les données à traiter sont des séries temporelles d'images ;
- **CBAR** : systèmes d'indexation du son où les données à traiter sont des documents audio : musique, parole, etc. ;
- **CBVR** : systèmes d'indexation vidéo où les données à traiter sont les documents audiovisuels. Typiquement un document vidéo est une séquence d'images associée à un contenu sonore (par exemple les films, les documentaires, etc.).

Pour un état de l'art sur les différents systèmes d'indexation existants et sur les techniques utilisées, on pourra se rapporter aux travaux proposés dans [Naphade 02] ou [Maillet 03]. Dans la suite nous allons présenter les caractéristiques générales de chaque catégorie de systèmes.

1.1 Les systèmes d'indexation d'images

Les premiers systèmes d'indexation multimédia sont les systèmes d'indexation de bases d'images statiques ou CBIR ("Content-Based Image Retrieval"). Dans ces systèmes l'analyse du contenu des images, nécessaire à l'indexation, est généralement orientée vers trois axes principaux : *la couleur*, *la forme* et *la texture* [Smeulders 00].

L'analyse des couleurs est une des directions les plus utilisées car, dans le système visuel humain, la couleur est une caractéristique fondamentale. On trouve ainsi des systèmes CBIR basés sur l'analyse des couleurs dans des applications pratiques telles que l'indexation des peintures [Lay 04] ou des photographies [Flickner 95]. Les couleurs sont analysées en utilisant différents espaces, en commençant par le classique espace RVB et en passant à des espaces plus complexes comme par exemple les espaces perceptuels (HSV, Lab, etc. [Mojsilovic 00]).

L'analyse des formes utilise les propriétés géométriques des objets contenus dans l'image pour caractériser la scène. Ceci demande la détection préalable des objets, le plus souvent par des techniques de segmentation par approche contours ou régions. Ces approches sont donc fortement dépendantes des performances des techniques de segmentation. Il faut

ensuite définir des caractéristiques des objets ne dépendant pas du point de vue sous lequel ces objets sont observés. Les démarches existantes proposent différents descripteurs de formes invariants aux transformations géométriques de l'image [Rivlin 95] et permettant également de solutionner le problème d'occlusion entre différents objets survenu suite à la projection de l'espace réel 3D dans l'espace 2D de l'image [Schmid 97].

L'analyse des textures est également très utilisée car ces informations permettent de caractériser les propriétés des matériaux présents dans l'image. Les démarches existantes utilisent des paramètres classiques de texture [Gimel'farb 96] ou des approches plus complexes comme l'analyse Markovienne [Choi 98].

Les systèmes actuels d'indexation d'images utilisent la collaboration naturelle entre ces trois informations : couleur, forme et texture. Pour un état de l'art détaillé sur les systèmes CBIR existants on pourra se rapporter aux travaux proposés dans [Smeulders 00].

L'utilisation des systèmes CBIR a fait apparaître deux problèmes. Le premier concerne l'écart entre les informations extraites automatiquement des images et la signification que l'on peut leur attribuer ("fossé sémantique" ou "semantic gap"). Le deuxième, connu dans la littérature comme le paradigme "sensor gap" se définit comme *la différence entre la scène 3D réelle et les informations reproduites par la représentation discrète 2D obtenue lors de l'enregistrement de la scène dans l'image* [Smeulders 00]. De ce fait, les systèmes d'indexation d'images donnent des résultats qui ne sont pas toujours en conformité avec les informations de la scène réelle représentée par l'image.

1.2 Les systèmes d'indexation de séquences d'images

Le passage à la vidéo numérique a orienté les systèmes d'indexation vers *les séquences d'images*. Les systèmes d'indexation des séquences d'images ou CBISR ("Content-Based Image Sequence Retrieval") sont à la base l'extension temporelle des systèmes CBIR. Dans ce cas le traitement n'est pas effectué sur des images statiques indépendantes les unes des autres, mais sur des séquences qui sont des suites temporelles d'images ou des images en mouvement. Souvent les séquences d'images sont nommées à tort films ou vidéos. La différence entre les deux notions vient du fait que les documents vidéo (films, documentaires, etc.) contiennent également l'information audio. Donc, une séquence d'images peut être définie comme étant la partie image d'un document vidéo[1]

Dans les systèmes CBISR le paradigme "sensor gap" est moins sensible grâce à la richesse des informations, en particulier temporelles, présentes dans les séquences d'images. Néanmoins, les systèmes d'indexation de séquences d'images vont poser de nouvelles difficultés.

Le premier problème est *la taille des données*. Par exemple, à une cadence de 25 images par seconde, une séquence d'images de 10 minutes contient *15000 images*. Un film à lui seul est ainsi équivalent, du point de vue de la taille, à une base contenant plusieurs dizaines de milliers d'images. Pour une base contenant plusieurs milliers de séquences d'images, les volumes atteints en nombre d'images sont absolument gigantesques. Ce volume très important des données entraîne une grosse difficulté pour accéder au contenu de ces données.

[1] les systèmes présentés dans ce chapitre mettent l'accent sur la partie image mais dans certaines situations le son est également pris en compte, les deux modalités étant parfois indissociables.

D'autre part, il s'ajoute à l'information spatiale fournie par l'image une nouvelle information à traiter : *l'information temporelle*. Si dans un système CBIR deux images qui contiennent les mêmes objets sont considérées comme similaires du point de vue de leur contenu, dans un système CBISR deux séquences d'images contenant les mêmes objets peuvent avoir des contenus très différents si l'on prend en compte l'aspect temporel. Ainsi, le comportement des objets et l'évolution temporelle de la scène sont des informations essentielles pour la compréhension du contenu des séquences et donc pour la tâche d'indexation.

Un autre aspect spécifique aux séquences d'images est *la structure hiérarchique* des données. Dans une séquence les images sont groupées *en plans vidéo* qui constituent l'entité de base de la séquence. Un plan vidéo est caractérisé par un ensemble d'images correspondant à une unité temporelle, spatiale et d'action. Le contenu de la séquence peut être aussi représenté en utilisant des informations sémantiques de plus haut niveau que les plans vidéo, comme par exemple *les scènes* (ensembles de plans similaires du point de vue sémantique), *les épisodes* (groupes de scènes), etc. (la structure temporelle d'une séquence d'images est abordée dans le Chapitre 2).

Dans la suite nous allons détailler les différentes étapes de traitements utilisées par les systèmes d'indexation de données en se focalisant sur les systèmes d'indexation de séquences d'images. Généralement, comme nous l'avons déjà mentionné, un système d'indexation comporte trois parties distinctes : *l'annotation du contenu* des données (syntaxiques ou sémantiques), *la navigation* et *la recherche*.

1.2.1 Le principe de l'annotation du contenu

L'étape *d'annotation du contenu* d'un système d'indexation correspond à la création d'attributs, ou d'index, permettant la compréhension automatique par l'ordinateur du contenu des données. Ces informations sont typiquement des propriétés basées sur les pixels, sur les régions de pixels, sur l'image entière, sur un groupe d'images ou sur la séquence entière. Les méthodes existantes d'analyse de séquences d'images exploitent deux directions fondamentales qui sont : l'information *spatiale* et l'information *temporelle*. Les différentes catégories d'informations utilisées pour la tâche d'annotation du contenu sont présentées en Figure 1.2.

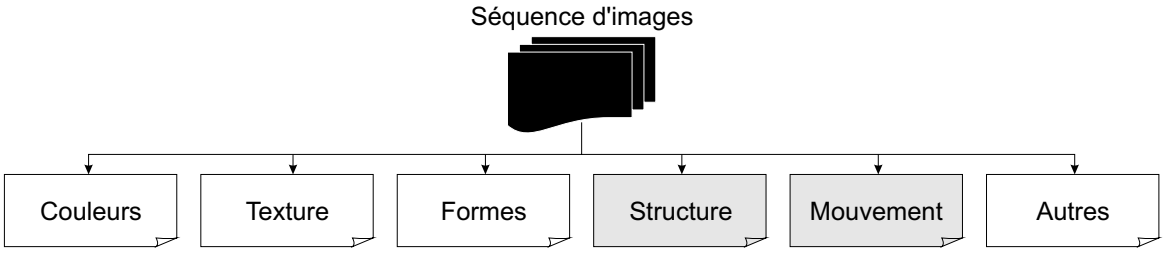


FIG. 1.2 – Les directions d'analyse utilisées par les systèmes d'indexation d'images (les éléments en grisé correspondent aux spécificités apportées par la dynamique des séquences).

L'indexation d'images étant un cas particulier simplifié de l'indexation des séquences d'images, on retrouve dans ce schéma les méthodes utilisées pour les images statiques (couleur, texture et formes) caractérisant les propriétés spatiales de l'image. La nouveauté liée aux séquences se situe d'une part dans l'analyse de l'évolution de ces attributs au cours du temps et d'autre part dans l'étude de la structure temporelle de la séquence et du mouvement

présent dans le succession des images.

Les différents types d'attributs utilisés pour l'annotation

La couleur. La couleur est un des attributs les plus importants pour la caractérisation du contenu de l'image. L'utilisation de la couleur permet de retrouver des ressemblances visuelles entre séquences d'images. En particulier, des caractéristiques statistiques telles que les histogrammes couleurs sont très souvent utilisées pour apprécier des similarités globales ou locales.

Une approche multi-résolution est proposée dans [Calic 02] où des histogrammes couleurs calculés sur différentes échelles de détails sont utilisés pour l'indexation. Le système propose un nombre variable de niveaux de détails et une mesure de pertinence est calculée en fonction de la dégradation de l'image. Cependant les mesures statistiques basées sur les histogrammes ne contiennent pas d'information spatiale. Différentes solutions ont été proposées pour combler ce manque. Par exemple [Chen 99] utilise des histogrammes "augmentés" qui sont calculés en ajoutant des informations statistiques sur les distances entre les pixels comme la moyenne, l'entropie, la variance, etc. L'approche basée sur l'histogramme ne permet pas non plus d'analyser l'évolution temporelle de la séquence.

Pour éviter ces inconvénients, d'autres méthodes proposent de caractériser la séquence à travers des vecteurs de caractéristiques locales des couleurs de l'image et en étudiant leurs évolutions temporelles. Un exemple d'une telle approche est proposé dans [Zhong 97] où les séquences sont caractérisées par des propriétés liées aux objets : couleurs spécifiques, dimensions, position et trajectoire dans la séquence.

Comme autres approches sur l'analyse des couleurs nous pouvons mentionner les arbres de décisions flous utilisés dans [Detyniecki 03] qui permettent d'extraire des règles d'indexation ou l'approche présentée dans [Adjeroh 01] fondée sur des modèles de la distribution des rapports entre les couleurs ou "color ratio models" (ces rapports sont calculés sur les contours de l'image).

La texture. L'analyse de la texture dans les séquences d'images est utilisée plutôt pour caractériser les propriétés des matériaux de la scène ou des objets d'intérêt. Les approches existantes utilisent généralement les mêmes paramètres de texture que ceux utilisés pour l'indexation des images fixes (voir la Section 1.1). Dans la littérature nous retrouvons très peu de techniques d'annotation du contenu qui utilisent seulement l'analyse de la texture, mais plutôt des méthodes qui utilisent la collaboration de plusieurs modalités de la séquence parmi lesquelles nous retrouvons la texture. Nous pouvons mentionner l'analyse des textures utilisées dans [Chang 99] pour la segmentation des images ou l'approche proposée dans [Bouthemy 98] dans laquelle le champ vectoriel du mouvement de la séquence est vue comme une texture et ses propriétés temporelles sont caractérisées à l'aide des matrices de co-occurrences.

Les formes. Les paramètres caractérisant les formes des objets, comme dans le cas des systèmes d'indexation d'images, sont aussi analysés dans le domaine spatial de l'image. La spécificité des séquences d'images est le déplacement des objets dans la scène qui se traduit dans l'espace de l'image par des transformations géométriques progressives de l'objet. Dans la construction des descripteurs de forme qui serviront d'index, l'invariance aux transformations géométriques est une propriété fondamentale. Les descripteurs les plus utilisés sont donc les moments invariants et les descripteurs de Fourier. Ainsi, dans [Mehtre 97], l'efficacité

des descripteurs basés sur les contours (Fourier, UFF - "UNL Fourier Features", etc.) est comparée à l'efficacité des descripteurs basés sur les régions de pixels (moments invariants, moments de Zernike). De plus, pour améliorer l'invariance, une collaboration entre différents types de descripteurs est proposée et testée : moments invariants et descripteurs de Fourier, ou moments invariants et caractéristiques UFF, etc. L'analyse de l'évolution temporelle des formes est souvent abordée pour le suivi d'objets d'intérêt dans la scène. Comme exemple nous pouvons mentionner l'approche proposée dans [Mazière 00] qui utilise un modèle multi-résolution des contours actifs pour la caractérisation et le suivi des objets.

La structure temporelle. Si l'analyse de la couleur, de la texture et des formes est commune aux images statiques et aux séquences, la structure temporelle est un paramètre spécifique aux séquences d'images. Dans l'étape de montage, les plans vidéo sont liés les uns aux autres pour définir la suite temporelle d'événements de la séquence (voir le Chapitre 2). Les méthodes d'annotation du contenu utilisées par les systèmes CBISR sont basées sur la segmentation temporelle en plans, sur l'extraction d'images clés (images représentatives du contenu) et sur l'analyse des similarités entre les unités structurelles de la séquence : les scènes, les épisodes, etc. (voir le Chapitre 5). La façon dont la structure d'une séquence a été conçue est souvent analysée à l'aide de modèles probabilistes tels que les chaînes de Markov cachées [Ferman 99]. Un état de l'art sur l'annotation basée sur l'analyse et l'évaluation des scènes est proposé dans [Vendriga 01].

Le mouvement. L'analyse du mouvement intervient naturellement de manière très importante dans l'annotation du contenu de séquences d'images car le mouvement représente l'évolution temporelle de la séquence. Dans les systèmes CBISR la caractérisation du mouvement est souvent réalisée par l'analyse du champ vectoriel du mouvement. Les vecteurs de déplacement sont soit estimés dans l'image soit directement récupérés du flux vidéo des séquences MPEG (voir le Chapitre 3).

Une première direction d'étude est l'annotation spatio-temporelle qui inclut la segmentation, le suivi d'objets et la caractérisation du mouvement à l'intérieur de certains passages importants de la séquence. Comme exemples nous pouvons mentionner les travaux proposés dans [Dagtas 00] où le mouvement spatio-temporel des objets est utilisé pour la caractérisation des événements importants de la séquence, ou le système VideoQ proposé dans [Chang 98] dédié uniquement à la caractérisation globale du mouvement des objets.

Le mouvement de la caméra est une autre direction importante d'étude pour l'annotation du contenu des séquences d'images. Les méthodes existantes caractérisent les propriétés des différents mouvements globaux présents dans la séquence : mouvements de translation, de rotation, zoom in/out, etc. (voir la Section 3.2). On peut citer par exemple l'approche proposée dans [Lee 01] où les différents mouvements de la caméra sont classés en utilisant des réseaux neuronaux et un certain nombre de modèles prédéfinis, ou encore l'approche basée sur la segmentation du mouvement proposée dans [Fablet 02] qui utilise des modèles de Gibbs pour représenter les dérivées du signal spatio-temporelles de l'image.

D'autres approches caractérisent l'information du mouvement en utilisant les trajectoires des objets ou des régions d'intérêt [Hsu 02]. [Zeng 02] propose une méthode basée sur l'information temporelle qui est transposée dans le domaine spatial de l'image à l'aide des cartes de mouvement.

Les autres axes d'étude. Les autres axes d'étude utilisent différentes sources d'informations issues de la séquence. Une des approches très souvent utilisée pour la tâche d'annotation

est la détection et l'analyse de la présence des personnes dans les scènes. Elle est réalisée par la détection des caractéristiques spécifiques : la couleur de la peau, le visage, les yeux, etc. [Acosta 02]. Les méthodes de localisation de visages sont basées sur des techniques supervisées comme les réseaux neuronaux ou les modèles de Markov cachés [Ben-Yacoub 99]. D'autres approches d'annotation incluent la détection de certains objets d'intérêt comme par exemple la présence de voitures dans [Schneiderman 00].

Enfin, une autre information utilisée pour l'annotation est le texte incrusté dans les images : des annotations textuelles, génériques, sous-titres, les indications de scores dans les séquences sportives, etc. Ces méthodes s'appuient sur la reconnaissance automatique de lettres (OCR - "optical character recognition"). Par exemple, dans [Kim 00b] les régions de l'image contenant du texte sont d'abord extraites lors d'une classification par des réseaux neuronaux et ensuite les lettres sont segmentées et reconnues.

1.2.2 L'annotation sémantique du contenu

Si nous analysons les différentes approches existantes d'annotation du contenu nous retrouvons deux axes d'étude privilégiés :

- **l'annotation syntaxique**, utilisée plutôt par les systèmes cités dans la section ci-dessus,
- **l'annotation sémantique**, qui est la nouvelle direction d'analyse utilisée par la plupart des systèmes CBISR actuels.

L'annotation *"syntaxique"*, est définie dans le dictionnaire par *ce qui concerne les relations entre les unités linguistiques et la construction grammaticale* [Robert 88]. Elle peut s'effectuer à partir des informations de bas niveau de la séquence comme par exemple des paramètres statistiques caractérisant les pixels ou des régions de pixels de l'image, les propriétés géométriques des objets, la structure temporelle de la séquence ou le mouvement. Les index obtenus sont des valeurs numériques, décrivant les attributs évoqués ci-dessus et aussi les relations syntaxiques existant entre ces attributs. Généralement ces index ne sont compréhensibles que pour des utilisateurs avisés. Par exemple, chercher une séquence d'images contenant 30% de mouvement de translation, 20% de rotation, etc., cela n'est pas très explicite.

D'autre part l'annotation *sémantique* permet de proposer une description perceptuelle du contenu des données. Les informations de bas niveau obtenues lors de l'analyse syntaxique peuvent être converties en concepts proches de la perception humaine. Une description sémantique du contenu demande la compréhension de l'ensemble du contenu de la séquence (image-son-texte). Des méthodes multi-modales seront donc envisagées.

En analysant la provenance du mot "sémantique" nous avons trouvé qu'il vient du domaine linguistique et est défini par : *étude du langage considéré du point de vue du sens* [Robert 88]. De plus, un "système sémantique" est défini par : *tout système comportant un ensemble de symboles (son vocabulaire), des lois de formation ou règles permettant de former des propositions, des lois de désignation et des lois de vérité* [Robert 88]. Dans les systèmes d'indexation le mot sémantique conserve ce sens. Il se traduit par le *codage de l'interprétation des données pour servir à une application* [Smeulders 00]. Les systèmes d'indexation sémantique contiennent alors des ensembles de règles et symboles permettant de réaliser l'interprétation linguistique de certains événements ou de certaines propriétés de la séquence.

L'annotation sémantique du contenu est également abordée dans les systèmes CBIR mais elle est rendue difficile car les caractéristiques sémantiques sont moins simples à extraire d'une seule image. Grâce à la richesse des informations présentes dans les séquences d'images (informations spatio-temporelles et de mouvement) l'analyse sémantique devient plus naturelle. Par exemple, si nous prenons le cas d'une image d'un joueur de foot, les seules caractéristiques que l'on peut faire ressortir de l'image sont liées à sa physionomie et à sa présence dans la scène. En revanche, si nous disposons de la séquence entière nous pourrons déterminer s'il est le joueur qui va marquer le but, quel est son style de jeu etc., autant d'informations sémantiques essentielles. Nous illustrons, dans la Figure 1.3, la différence entre les deux concepts d'annotation, l'annotation syntaxique et sémantique.

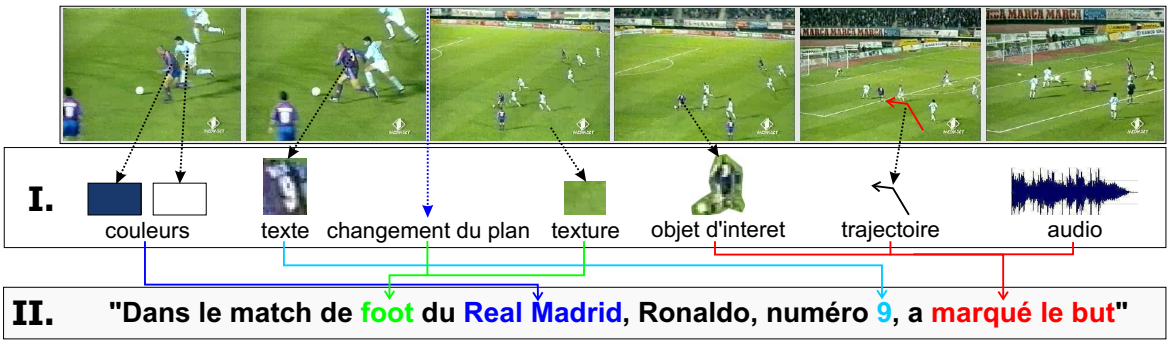


FIG. 1.3 – La différence entre l'annotation syntaxique et sémantique : I. annotations syntaxiques, II. annotation sémantique (séquence représentant un but de Ronaldo).

Comme nous l'avons déjà dit, le paradigme "sensor gap" des systèmes d'indexation d'images est moins sensible dans les systèmes d'indexation de séquences d'images car une meilleure compréhension du contenu est possible grâce au volume d'informations fourni par les séquences d'images. Par contre, le "fossé sémantique" ("semantic gap"), déjà présent dans les systèmes CBIR, prend encore plus d'ampleur dans les systèmes CBISR. C'est : *le manque de correspondance entre l'information que l'on peut récupérer du contenu des données et l'interprétation que l'on peut en faire* [Smeulders 00]. Dans les systèmes CBISR, différentes solutions ont été proposées pour tenter de combler ce fossé sémantique [Naphade 02][Snoek 05b]. Dans la suite nous allons présenter les différents travaux proposés dans cette direction.

En fin de compte, un système d'indexation sémantique efficace doit réunir plusieurs caractéristiques importantes [Naphade 02] :

- la première est sa *capacité d'analyse sémantique* à partir des requêtes formulées par l'utilisateur (voir la Section 1.2.4 de ce chapitre),
- un système CBISR efficace doit être *multi-modal* en réunissant et homogénéisant l'analyse de plusieurs modalités de la séquence comme par exemple l'image, le son et le texte,
- enfin, les relations entre les caractéristiques de bas niveau et leur perception sémantique doivent être *résumées d'une manière efficace* pour que le système soit capable de mettre à la disposition de l'utilisateur des descripteurs sémantiques.

L'évolution actuelle des systèmes d'indexation des séquences d'images vers l'analyse sémantique du contenu a été également motivée par l'attention accordée au nouveau standard de compression vidéo MPEG-7 [Wang 00]. Le nouveau format vidéo introduit dans le flux de données des informations sémantiques sur le contenu de la séquence (voir le Chapitre

3). Pour plus de détails sur les systèmes d'indexation sémantiques on pourra se rapporter aux travaux proposés dans [Naphade 02] et [Snoek 05b].

Les objectifs d'un système d'indexation sémantique

Globalement, les objectifs des systèmes d'indexation CBISR sémantique peuvent être divisés selon quatre directions différentes [Naphade 02] :

- l'analyse des **structures de haut niveau** de la séquence,
- la classification des **genres**,
- l'analyse **dépendante du domaine** d'application,
- l'analyse **indépendante du domaine** d'application.

L'analyse des structures de haut niveau. L'étude des *structures de haut niveau* de la séquence comporte la détection et l'analyse de différents passages sémantiques, comme les dialogues, les publicités, etc.

Par exemple, dans [Hauptmann 98] les taux de changements de plans et la présence d'images noires sont utilisés pour la détection des passages publicitaires dans des séquences d'informations. Une approche similaire appliquée à la détection des passages publicitaires dans les films utilise la vitesse de changements des contours et l'analyse de la taille des vecteurs de mouvement pour retrouver l'action [Lienhart 97]. L'approche proposée dans [Alatan 01] détecte les scènes de dialogue par l'analyse du son, la détection et la localisation des visages. Située au plus bas niveau de l'analyse sémantique, la détection d'événements d'intérêt est aussi abordée par les systèmes CBISR. Par exemple la méthode proposée dans [Haering 00] classifie les passages de chasse d'animaux en utilisant des réseaux neuronaux. Les informations analysées sont la couleur, la texture et le mouvement.

La classification des genres. Un autre objectif des systèmes d'indexation sémantique est la *classification des genres* ou *du type de la séquence*. Les catégories les plus analysées sont les nouvelles, les reportages sportifs, les films et les films publicitaires.

Dans [Truong 00a] la taille des plans vidéo, le pourcentage d'apparition de différentes transitions vidéo et certaines caractéristiques des couleurs sont utilisées conjointement pour classer les séquences d'images en séquences d'animation, publicitaires, musicales, de nouvelles ou sportives. Une autre approche sur la classification des genres dans des séquences génériques (pour lesquelles le domaine d'application n'est pas a priori connu) est proposée dans [Kobla 00]. Elle utilise comme informations la répétition du mouvement, la présence du texte et des mouvements spécifiques de la caméra ou des objets. Les genres sont déterminés en utilisant des arbres de décisions. Une autre approche intéressante est présentée dans [Colombo 99] où les séquences publicitaires sont classées selon la perception du contenu en quatre sous-genres définis a priori : le type pratique, critique, utopique et animé. Les paramètres utilisés sont la saturation des couleurs, la présence des lignes horizontales dans les images, le mouvement et la statistique des transitions vidéo.

L'analyse dépendante du domaine d'application. Dans *l'analyse dépendante du contexte* les connaissances sont extraites à l'aide de l'expertise de spécialistes du domaine cible et sont plutôt spécifiques au genre des séquences analysées (voir l'état de l'art proposé dans [Snoek 05b]). Comme exemples nous pouvons mentionner l'approche proposée dans [Saur 97] pour les séquences sportives (séquences de basket-ball) ou dans [Fan 04] pour les

séquences de reportages médicaux.

L'analyse indépendante du domaine d'application. *L'analyse indépendante du domaine* est la direction la plus difficile à aborder pour les systèmes d'indexation sémantique.

Actuellement, il n'y a pas beaucoup de travaux proposés dans la littérature. Les techniques existantes *essaient* de mettre en place des techniques d'annotation et de classification automatique du contenu de séquences génériques, sans utiliser de connaissances a priori sur le domaine ou sur le contenu des séquences utilisées. Comme exemples nous pouvons mentionner l'utilisation de représentations probabilistes du contenu multimédia proposée dans [Naphade 01a] ou l'utilisation de différents chemins de classification sémantique proposée dans [Qian 99]. Cette direction d'étude reste le défi le plus important des systèmes d'indexation sémantique.

Les difficultés de l'analyse sémantique

Les systèmes d'indexation sémantique existants répondent plus ou moins aux besoins actuels de l'indexation de séquences d'images. Leur succès dépend de la façon dont les difficultés de l'annotation sémantique du contenu ont été surmontées. Dans la suite nous allons présenter les différentes contraintes de l'annotation sémantique qui ont été bien mises en évidence par les travaux proposés dans [Fan 04].

La concordance. La première difficulté de l'annotation sémantique du contenu est *la concordance* entre l'analyse de bas niveau et la description sémantique envisagée. Elle dépend de la qualité des paramètres numériques de bas niveau utilisés pour l'inférence sémantique. Pour bien faire la différence entre les différents concepts sémantiques possibles la diversité des paramètres utilisés doit être suffisamment importante. La plupart des approches existantes utilisent comme information de départ le découpage en plans ou en objets sémantiques (par exemple, les scènes, certains passages d'intérêt, etc.) pour l'extraction d'attributs [Fan 01]. Cependant, ce type de caractéristiques de bas niveau peut difficilement être associé à des concepts sémantiques, donc une approche automatique est peu envisageable [Erol 00].

La modélisation. Un deuxième problème est *la modélisation de concepts sémantiques*. A cause du "fossé sémantique" ("semantic gap", évoqué en début de section) les systèmes actuels ne sont pas capables de fournir un accès au contenu des séquences d'images qui soit complètement sémantique. Différentes solutions ont été proposées pour tenter de résoudre ce problème. On peut citer les approches utilisant des connaissances a priori sur le domaine pour générer des règles perceptuelles de la description sémantique [Adames 02], celles exploitant l'interaction entre l'utilisateur et le système au niveau du résultat ("boucle de pertinence") pour l'amélioration des performances de l'indexation [Cox 00], ou enfin les approches basées sur des méthodes statistiques d'apprentissage par ordinateur ("machine learning") des corrélations cachées existant entre les données multi-modales [Barnard 03].

La classification. *La classification sémantique* est aussi une des difficultés rencontrées par les systèmes d'indexation sémantique. Les méthodes proposées se divisent entre les approches basées sur des systèmes de règles déterminées à travers des connaissances a priori sur le domaine [Alatan 01] et les approches statistiques [Wang 01]. La première catégorie se limite à l'utilisation unique de règles perceptuelles sans l'exploitation des relations cachées entre les modalités de la séquence. A l'inverse, les approches statistiques permettent d'explo-

rer les liens existant entre les différentes sources d'informations, mais leurs performances sont très dépendantes de l'efficacité des paramètres choisis et de l'entraînement des classificateurs utilisés.

La sélection. Un autre problème est *la sélection des caractéristiques* et *la réduction de la dimension de l'espace des paramètres*. Intuitivement on sent bien que l'augmentation du nombre d'attributs de bas niveau utilisés pour l'annotation sémantique entraînera une meilleure puissance discriminatoire et par conséquent une annotation sémantique plus efficace. Mais, dans ce cas, le temps de calcul nécessaire à l'entraînement des classificateurs utilisés pour l'annotation augmente très fortement avec la dimension de l'espace des caractéristiques utilisées. Il est donc nécessaire de sélectionner les attributs les plus "rentables" vis-à-vis de l'indexation.

L'organisation. *L'organisation* de la base de données et *l'accès au contenu* sont deux aspects aussi importants pour les systèmes d'indexation sémantique. Malheureusement, le domaine des bases de données et celui de la vision par ordinateur n'ont encore actuellement qu'une faible interaction et donc un faible échange de connaissances [Smeulders 00]. Dans le cas idéal, la base de données devrait être structurée au départ de telle manière que l'accès aux données soit facilement effectué à un degré sémantique perceptuelle proche de la perception humaine et accessible aussi aux utilisateurs non avisés [Benitez 01].

1.2.3 Le système de navigation dans la base

L'accès au contenu des données dans une base de séquences d'images est très difficile à cause du volume élevé d'informations disponibles (nombre trop élevé d'images à passer en revue). Regarder chaque séquence est une tâche presque impossible car une base comporte des milliers de séquences. Pour faciliter l'analyse des données, les systèmes d'indexation de séquences d'images[2] utilisent des aperçus compacts des contenus des séquences. Ils sont mis à la disposition de l'utilisateur à travers des outils logiciels permettant la visualisation efficace des données de la base, outils qui constituent le *système de navigation* de la base.

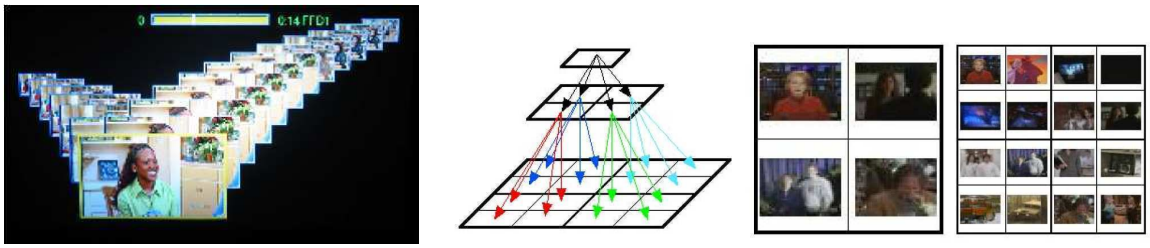
Une première catégorie de construction d'aperçus est de *résumer le contenu* de la séquence en utilisant sa structure temporelle (comme par exemple le découpage en plans). Deux types de résumés sont possibles : le résumé en images (statique) qui est à la base une collection d'images représentatives de la séquence et le résumé en mouvement (dynamique) qui est une collection de sous séquences. Nous reviendrons plus en détails sur les techniques de construction de résumés dans le Chapitre 6. Les résumés nous permettent de nous faire rapidement une idée globale du contenu de la séquence. Le résumé statique permet d'avoir un résumé du contenu visuel en seulement quelques images, facilement accessible dans l'outil de navigation. Le résumé dynamique, quant à lui, apporte plus d'informations sur le contenu au niveau de l'action de la séquence, informations qui sont perdues dans le résumé en images.

Comme exemple de systèmes d'indexation de séquences d'images qui utilisent les résumés pour la navigation, nous pouvons mentionner le système proposé dans [Zhu 05] dans lequel le contenu des séquences est résumé en utilisant des images clés. Les images ainsi retenues sont ensuite modifiées selon un ensemble prédéfini de planches présentées sous la forme d'une brochure. L'approche proposée dans [van Houten 03] représente le contenu de la séquence sous la forme d'un ensemble de fragments ("patches"). Les fragments sont définis comme des

[2]mais pas seulement, c'est aussi le cas des systèmes d'indexation audio et vidéo.

passages de la séquence de même nature, par exemple les interviews, les dialogues, etc.

Un cas particulier de résumé est la représentation d'images clés du résumé dans un espace 3D formé en utilisant comme troisième dimension l'axe temporel de la séquence. Par exemple dans [S. Vogl 99] les séquences d'images sont résumées par des suites temporelles d'images clés qui sont visualisées dans un monde virtuel permettant l'interaction de l'utilisateur. Un système similaire, appelé le "tunnel temporel", est proposé dans [Laboratoires 05]. Les images clés sont visualisées sous la forme de couches superposées représentant l'évolution temporelle des images dans la séquence (voir la Figure 1.4.a).



(a) (b)

FIG. 1.4 – Exemples de méthodes de visualisation du contenu de séquences d'images : (a) Timetunnel, (b) ViBE.

Une autre catégorie de visualisation du contenu de la séquence est la représentation par des structures hiérarchiques. Le système proposé dans [Eidenberger 04] utilise la navigation interactive à l'aide de descripteurs du contenu extraits du standard MPEG-7 (voir [Jeannin 01]). Deux représentations arborescentes sont possibles : une représentation du contenu des plans et une représentation de la structure temporelle. Dans chaque niveau hiérarchique les données sont structurées sous la forme de cartes auto-organisables ("self-organizing maps") d'objets MPEG-7. Un autre système de représentation hiérarchique est le système ViBE proposé dans [Chen 06]. Les plans sont visualisés sous la forme de structures arborescentes d'images clés classées selon des catégories pseudo-sémantiques a priori connues. Les données sont représentées comme une structure pyramidale permettant plusieurs niveaux de détails (voir la Figure 1.4.b).

Une catégorie à part de systèmes de navigation concerne les systèmes de navigation disponibles en ligne sur Internet qui utilisent comme interface de navigation le "web browser". Comme exemple nous pouvons mentionner le système Vimix proposé dans [Yao 01] qui utilise une structure hiérarchique de données vidéo basée sur le langage XML ou le système BIBS [Rowe 01] qui propose une organisation structurelle linéaire de la séquence et synchronise la visualisation des segments vidéo avec l'annotation textuelle du contenu.

1.2.4 Le mécanisme de recherche dans la base

Le but d'un système d'indexation est d'offrir à l'utilisateur un accès facile et efficace au contenu des données. Pour atteindre ce but le système comporte plusieurs étapes. Dans un premier temps, différentes annotations contextuelles sont construites servant au tri des données selon la similarité de leurs contenus (voir la Section 1.2.1). L'accès aux informations est ensuite effectué à l'aide du système de navigation (voir la Section 1.2.3). A cause du nombre élevé de données disponibles dans la base, ces deux premières étapes ne suffisent

pas pour accéder d'une manière efficace au contenu. Le dernier outil exigé est le *système de recherche* permettant la recherche des données désirées en fonction de différents critères formulés habituellement sous la forme de requêtes. La recherche doit être *intuitive et naturelle*, accessible aux utilisateurs non avisés du domaine.

Globalement, un système de recherche fonctionne de la manière suivante :

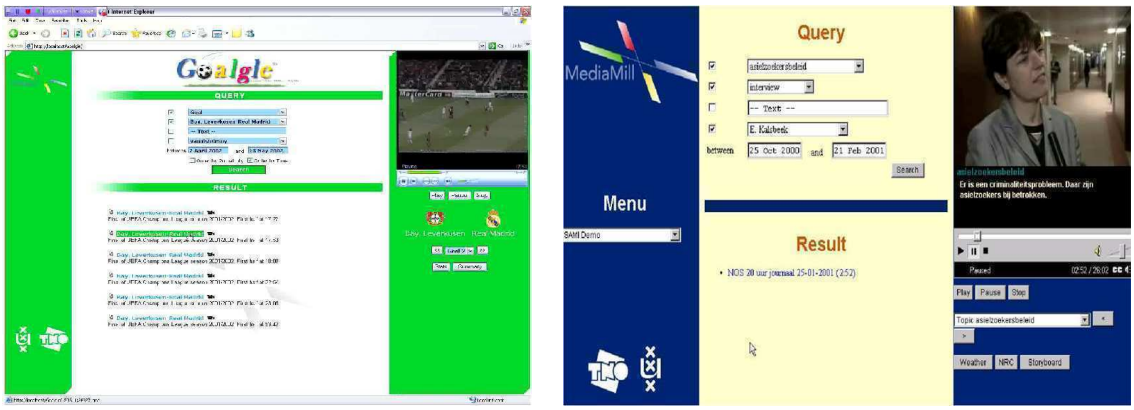
- **la requête** : d'abord l'utilisateur rédige sa requête. Plusieurs méthodes sont possibles : en proposant *un exemple* de ce que l'on recherche, par une *description textuelle* du contenu désiré ou enfin en utilisant des *outils graphiques* permettant une description rapide de la requête.
- **conversion en descripteurs de bas niveau** : ensuite le système de recherche traduit la requête en attributs syntaxiques de bas niveau, ceux utilisés pour l'annotation du contenu de la base.
- **la recherche** : enfin, différentes mesures de similarité entre les attributs des données sont utilisées pour retrouver dans la base les données correspondant au mieux à la requête faite par l'utilisateur. Les résultats ainsi obtenus sont visualisés dans le système de navigation.
- **le "feedback" du système** : optionnelle, cette étape d'interaction de l'utilisateur sur les résultats obtenus, appelée "boucle de pertinence" ou "feedback", est utilisée pour améliorer l'algorithme de recherche. L'évaluation des résultats par l'utilisateur servira comme apprentissage pour le système de recherche.

La qualité de la recherche dépend de plusieurs facteurs. Bien sûr, la qualité des attributs caractérisant les données et les mesures de similarité influenceront les résultats de la recherche. Mais ces résultats sont aussi dépendant de la façon dont la requête est formulée. Selon la formulation de cette requête, nous retrouvons plusieurs techniques de recherche [Fan 04] :

- **l'utilisation d'un exemple** : la demande est formulée en utilisant un modèle de données [Tong 01]. Par exemple, l'utilisateur cherchera toutes les séquences qui ressemblent le plus à une certaine séquence dont il dispose. La ressemblance est traduite en utilisant le contenu multi-modal de la séquence : similarité des couleurs, du mouvement, de l'action, etc. Mais, si l'utilisateur ne dispose pas de bons exemples pour sa recherche (cas d'un utilisateur non avisé) cette méthode de recherche ne sera pas du tout efficace.
- **représentation textuelle** : la demande est formulée en utilisant une représentation textuelle du contenu des données souhaitées [Benitez 01]. Ce type de représentation permet la recherche en utilisant des concepts sémantiques formulés à l'aide d'un texte. Par exemple, l'utilisateur peut chercher toutes les "séries TV" ou, à un niveau sémantique supérieur, tous les "films tristes". Le principal inconvénient est le manque de sens de certaines annotations textuelles automatiques dans des bases de données très volumineuses.
- **l'outil de navigation** : l'utilisateur peut utiliser directement le système de navigation pour effectuer la recherche [Smith 99]. Ce type de recherche est approprié pour les utilisateurs non avisés qui n'ont pas de connaissance sur le contenu des données et sur les critères de recherche. La contrainte de ce type de recherche vient du fait que les données de la base ne sont habituellement pas structurées selon des critères sémantiques basés sur le contenu.

Nous allons présenter quelques uns des principaux systèmes de recherche utilisés par les systèmes d'indexation de séquences d'images. La plupart de ces systèmes analysent les caractéristiques définissant le contenu de la séquence au moyen de relations temporelles entre attributs [Allen 83].

Comme exemples de systèmes de recherche nous pouvons d'abord mentionner les systèmes SMOOTH [Kosch 01], GOALGLE et News RePortal [Snoek 05a] (voir la Figure 1.5) dans lesquels la recherche est effectuée par rapport à des critères sémantiques ou temporels. Un autre exemple est le système SoccerQ [Chen 05] qui permet la recherche sémantique d'événements, ayant une certaine spécificité dans les séquences sportives. La recherche est effectuée par rapport à trois niveaux structurels : niveau séquence, niveau segment ou niveau variable (par exemple les noms des équipes). Les demandes sont formulées dans un langage naturel, comme par exemple : "trouver toutes les séquences contenant un coup-franc". D'autres approches utilisent la définition d'un langage de recherche, comme par exemple les travaux proposés dans [Donderler 04]. Une approche différente est proposée dans [Liu 02a] dans laquelle les relations spatiales et temporelles entre les objets de la scène sont modélisées par des relations textuelles entre des symboles à travers des "3D-strings". Le problème de recherche est transformé en un problème de ressemblance entre des symboles.



(a) (b)

FIG. 1.5 – Exemples de moteurs de recherche par le contenu : (a) le système GOALGLE de recherche des séquences de foot, (b) le système News RePortal de recherche des séquences d'informations.

D'une manière générale les systèmes de recherche sont dépendants du domaine d'application : séquences sportives, films, dessins animés, etc. Les critères de recherche sont construits en utilisant des connaissances a priori sur le domaine. Si l'on désire disposer d'un système de recherche général indépendant de tout domaine, une solution pourrait être de réunir "l'expérience" des systèmes de recherche développés dans chaque domaine d'application. De cette manière nous profitons des critères de recherche qui sont les meilleurs de chaque domaine.

1.3 Les systèmes d'indexation du son

Une autre catégorie de systèmes d'indexation concerne les systèmes d'indexation du contenu audio ou CBAR ("content-based audio retrieval"). Dans ce cas les données à in-

dexer sont des *documents audio*, comme par exemple la parole ou la musique. Généralement, pour l'annotation du contenu, les informations audio sont analysées de façon chronologique sous deux niveaux différents de détails : le niveau des cadres qui sont des fenêtres temporelles réduites ("short-term frame level") ou le niveau des segments ("long-term clip level"). Les attributs spécifiques aux données audio sont calculés dans le domaine temporel mais aussi dans le domaine fréquentiel. Les méthodes utilisées sont plutôt des méthodes classiques issues du domaine du traitement du signal et de la parole. Les propriétés des données sont exprimées par des paramètres de bas niveau spécifiques, comme par exemple : le volume, les taux de passages par zéro du signal, le "pitch", les paramètres spectraux, etc. Pour plus de détails on pourra se rapporter aux travaux présentés dans [Wang 00].

Une direction largement abordée par les systèmes CBAR est l'indexation de la parole [Naphade 02]. Dans ce cas, pour faciliter l'analyse, les méthodes d'annotation sont appliquées dans des conditions particulières, comme par exemple l'absence de bruit ou sur un vocabulaire de mots a priori connu. Comme exemples, nous pouvons mentionner ceux présentés dans [Lab. 05] ou [Nam 97]. D'autres systèmes CBAR font l'indexation du son en catégories définies a priori (c'est en quelque sorte une classification), comme par exemple : la parole, la musique, la violence, etc. Les méthodes utilisées se divisent en deux axes : l'analyse à partir de règles et l'analyse à partir de modèles [Naphade 02].

Une autre direction d'étude importante est la classification des genres de musique, demande forte pour la recherche dans des bases de données musicales au format électronique mp3, ogg, etc. Les paramètres définissant un document musical dans la base sont le genre musical (par exemple blues, classique, jazz, etc.), le nom de l'artiste et le titre de la chanson. Les méthodes d'annotation utilisent des algorithmes de classification comme le k-means, les réseaux neuronaux ou les systèmes experts. La classification est effectuée sur un certain nombre de paramètres spécifiques : le timbre musical, l'harmonie/mélodicité et le rythme [Scaringella 06].

La plupart des systèmes d'indexation actuels ne se limitent pas à l'utilisation d'une seule modalité ou d'une seule information (comme par exemple le son ou l'image) mais envisagent la collaboration des deux sources d'information. C'est le cas des systèmes d'indexation de documents vidéo présentés dans la suite.

1.4 Les systèmes d'indexation vidéo

Les systèmes d'indexation vidéo ou CBVR ("content-based video retrieval") sont l'extension naturelle des systèmes d'indexation de séquences d'images et d'indexation du son, présentés ci-dessus, car les données à traiter sont des documents audio-visuels. Un document vidéo est défini comme une séquence d'images pour laquelle le son est disponible. Souvent dans la littérature spécialisée le mot vidéo est utilisé abusivement pour désigner des séquences d'images. Aussi, définir une frontière entre ces deux types de systèmes (indexation de séquences d'images et de documents vidéo) n'est pas toujours facile. Dans la suite, nous désignerons par système d'indexation vidéo tout système qui inclut l'analyse du son, analyse qui n'est pas présente dans les systèmes d'indexation de séquences d'images.

Les systèmes d'indexation de séquences d'images peuvent être considérés comme un cas particulier de systèmes d'indexation vidéo (absence du son), toutes les méthodes utilisées par les premiers (voir la Section 1.2) sont réutilisables par les seconds. Dans la suite nous n'allons pas reprendre les techniques déjà discutées dans le cas de systèmes d'indexation de

séquences d'images, mais nous allons présenter la spécificité des systèmes d'indexation vidéo existants.

La plupart des systèmes d'indexation vidéo CBVR n'utilisent pas de vrais approches multi-modales où la partie image et la partie son sont analysées conjointement. Les deux modalités, image et son, sont d'abord traitées séparément et les résultats sont ensuite fusionnés permettant une description des documents vidéo [Naphade 02]. Par exemple, l'analyse des images peut être utilisée pour la segmentation temporelle en plans et le son peut être utilisé plus tard pour classer le contenu vidéo, comme c'est le cas du système proposé dans [Wang 00].

D'une manière générale, les systèmes utilisant l'intégration de l'analyse multi-modale image-son sont divisés en deux catégories [Snoek 05a] : les approches basées sur des *règles* (définies à partir de connaissances a priori) [Babaguchi 02] et les approches *statistiques* (utilisant des algorithmes de classification de données) [Han 02]. La première catégorie de méthodes analyse indépendamment chaque modalité des données vidéo (par exemple l'image, le son et le texte) et les résultats sont fusionnés à la fin en utilisant une classification basée sur des règles. Par exemple, nous pouvons mentionner l'approche proposée dans [Babaguchi 02] où le son est utilisé d'abord pour l'identification de mots spécifiques aux événements qui ont trait au football (par exemple le moment d'un but) et ensuite l'information visuelle est utilisée pour la classification du contenu vidéo. Parmi les approches statistiques les plus utilisées, nous pouvons mentionner : les réseaux dynamiques Bayesiens [Naphade 01b], les arbres de décision [Zhou 02] ou les "support vector machines" [Lin 02].

L'analyse multi-modale image-son utilisée par les systèmes CBVR demande des méthodes efficaces de fusion entre les différents canaux d'informations. Les contraintes de l'analyse multi-modale image-son peuvent être résumées de la manière suivante :

- **la synchronisation des données** : nécessaire pour homogénéiser les informations issues des différents canaux d'informations utilisés par le système. La solution la plus souvent utilisée est la conversion de toutes les données en un système unique de référence [Snoek 05b].
- **le choix du modèle adéquat** : l'introduction d'informations supplémentaires qui ne sont pas disponibles au moment exact de l'événement sémantique analysé et qui peuvent être prélevées en amont ou en aval de l'événement.
- **la redondance des paramètres utilisés** : l'annotation multi-modale du contenu utilise différents paramètres qui sont calculés pour chaque modalité de la séquence. Le problème est la forte corrélation des données ainsi obtenues. Une étape de décorrélation sera souvent nécessaire [Wang 00].

En conclusion, les systèmes d'indexation de documents vidéo actuels utilisent la fusion de différents types d'attributs extraits des images, du son et du texte incrusté dans les images. Le but est d'annoter le contenu dans l'objectif de fournir un niveau sémantique d'accès aux données.

1.5 Présentation du système d'indexation proposé

Les travaux présentés dans cette thèse s'appuient sur la construction d'un système d'annotation et de caractérisation du contenu des séquences d'images servant à l'indexation et à l'analyse d'une base de données de séquences d'images. Les séquences sont analysées en

utilisant plusieurs sources d'information : *l'image*, *la structure temporelle*, *le mouvement*, et, dans une moindre mesure, *les synopsis* ou *les informations textuelles* qui se rapportent aux séquences. Les annotations proposées sont des descriptions symboliques et sémantiques proches de la perception humaine du contenu qui peuvent servir comme index sémantiques de recherche dans la base. De plus nous avons proposé et étudié différentes techniques de construction automatique de résumés du contenu facilitant la navigation dans la base de données.

La structure du système proposé

Le système d'annotation du contenu de séquences d'images proposé comporte plusieurs étapes d'analyse, correspondant au diagramme présenté dans la Figure 1.6. Les séquences sont analysées en deux étapes :

- **bas-niveau** : d'abord c'est *l'analyse de bas-niveau* du contenu où un certain nombre de paramètres statistiques sont calculés servant par la suite comme attributs syntaxiques du contenu,
- **sémantique** : ensuite une *analyse symbolique/sémantique* du contenu est réalisée. Différentes descriptions sémantiques du contenu sont extraites des paramètres de bas-niveau en utilisant des connaissances a priori du domaine et une représentation des données par des ensembles flous.

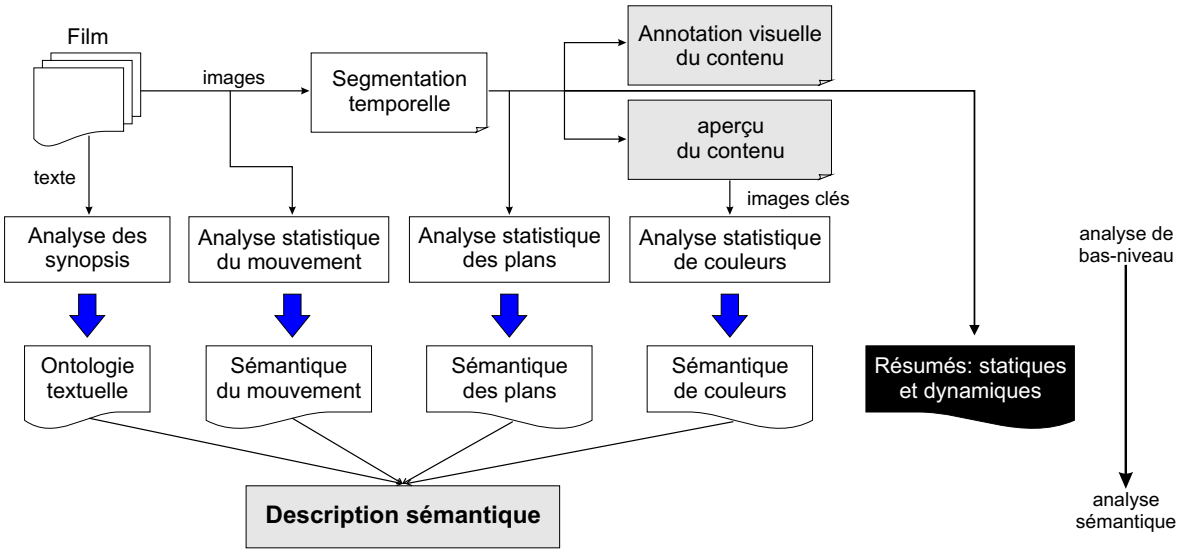


FIG. 1.6 – Le diagramme du système d'annotation proposé.

Dans l'étape d'analyse de bas-niveau la séquence est *découpée en plans vidéo* par la détection de transitions vidéo (voir la Section 2.1). Nous avons construit une *annotation visuelle* des transitions et des plans (voir la Section 2.8) pouvant servir à analyser la structure temporelle de la séquence. De plus, différentes techniques de construction de *résumés automatiques* du contenu sont proposées : résumés en images, résumés dynamiques et résumés de type bande-annonce (voir le Chapitre 6). L'utilité des résumés est double : premièrement c'est une représentation compacte du contenu, donc cela facilite la tâche de navigation dans

la base de données, et ensuite cela permet de diminuer la redondance temporelle des données. Les attributs de bas-niveau sont calculés sur les différentes modalités de la séquence :

- **le mouvement** : une analyse du mouvement de la caméra et des objets est réalisée en utilisant l'estimation des vecteurs de mouvement. Cette analyse nous a permis de calculer un certain nombre de paramètres statistiques spécifiques (voir Chapitre 3).
- **la structure temporelle** est analysée à partir des transitions vidéo (voir Section 2.9).
- **la couleur** : un aperçu compact du contenu de la séquence est utilisé pour la caractérisation globale de la distribution des couleurs de la séquence (voir Chapitre 4).
- **le texte** : une décomposition lexicale en termes est effectuée en utilisant les synopsis de la séquence, qui sont des fiches textuelles du contenu attachées à chaque séquence (voir l'approche envisagée dans le Chapitre 9 sur les perspectives[3]).

Dans l'étape d'analyse symbolique/sémantique du contenu, les paramètres de bas-niveau obtenus lors de l'étape précédente sont utilisés pour construire des descriptions sémantiques du contenu de la séquence. L'ontologie textuelle est obtenue en étudiant les concepts linguistiques recherchés à partir des synopsis (voir Chapitre 9). Les paramètres de bas-niveau du mouvement, de la structure temporelle et de la distribution des couleurs sont transformés en concepts linguistiques à partir d'une représentation à base d'ensembles flous et en utilisant des connaissances a priori sur le domaine. Différentes descriptions symboliques/sémantiques sont proposées, comme par exemple : le rythme, l'action, les techniques d'utilisation des couleurs, etc. (voir Chapitre 7). Les descriptions proposées ont été utilisées pour la classification d'une base de séquences d'images, montrant ainsi leur pouvoir discriminant et leur utilité en tant qu'index sémantiques lors d'une recherche (voir Chapitre 8).

L'application aux films d'animation

Le système d'analyse proposé, a été appliqué au cas particulier des films d'animation dans le contexte du "Festival International du Film d'Animation" (FIFA) [CICA 06]. Le FIFA est un événement international qui se déroule depuis 1960 à Annecy, faisant suite aux premières rencontres au "Festival de Cannes" en 1956. Chaque année des centaines de films d'animation provenant de différents pays du monde rentrent en concurrence lors de ce festival. Actuellement, après 46 années d'existence, le FIFA est devenu l'un des événements les plus importants du domaine du cinéma d'animation.

Le "Centre International du Cinéma d'Animation" ou CICA [CICA 06], qui organise le festival, est en train de mettre en place une base numérique des films d'animation (voir Animaquid [CICA 06]) qui sera disponible en ligne pour une utilisation publique ou semi-publique. Compte tenu de la quantité de films, des outils logiciels permettant l'accès au contenu artistique des films d'animation sont nécessaires. Ces outils peuvent par exemple être des outils d'indexation sémantique du contenu proches de la perception humaine. Les seules informations disponibles pour le moment sont les fiches textuelles, fournies pas les auteurs, qui peuvent être consultées en ligne sur le site du CICA. Cette représentation n'est pas totalement appropriée car la richesse du contenu artistique se trouve plutôt dans les images de la séquence.

Dans cette collaboration entre le LISTIC de l'Université de Savoie et le CICA, les objectifs peuvent être résumés par :

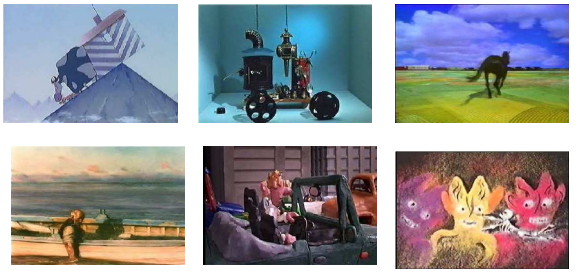
[3]les travaux sur l'analyse des synopsis ont été réalisés en collaboration avec l'équipe Ingénierie des Connaissances (Condillac) du LISTIC [Condillac 05]

- **un objectif de conservation du "patrimoine"** : conservation et mise à disposition de l'ensemble des films disponibles. La constitution d'une base numérisée permet d'obtenir un support de sauvegarde fiable et un accès facile. Des outils informatiques peuvent alors être mis en place pour permettre la recherche ou la navigation dans la base constituée.
- **un objectif d'exploitation** : permettre une utilisation et une exploitation nouvelle de cette base par la construction d'outils de caractérisation et d'analyse des films. Cela présente un intérêt fondamental à la fois pour les professionnels, les cinémathèques, les enseignants ou même le grand public.

Dans ce contexte, le système proposé a pour but de caractériser la perception des films d'animation et de mettre à la disposition des artistes des outils permettant une analyse détaillée du contenu artistique de la séquence (techniques utilisées, structure, etc., voir [Lambert 07]). Les films d'animation du festival FIFA sont différents des films traditionnels d'animation, communément appelés dessins animés, par leur contenu qui relève souvent d'une intention artistique plus que d'une recherche de divertissement. Chaque film traite d'un sujet particulier et essaye de transmettre certaines idées ou certains sentiments de l'artiste (voir des exemples dans l'Annexe F).

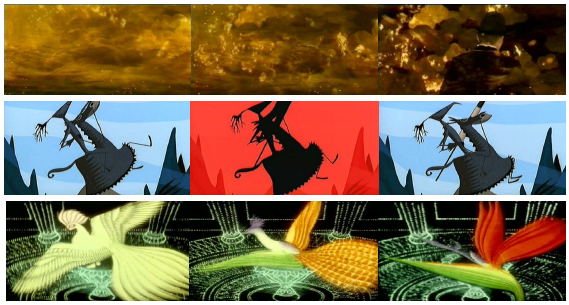
Les caractéristiques les plus importantes des films d'animation peuvent se résumer de la manière suivante (voir Figure 1.7) :

- **la durée** : les films d'animation utilisés sont plutôt du type court métrage (durée moyenne de 10 minutes).
- **les événements** : ils ne suivent pas forcément une chronologie bien établie : des objets peuvent apparaître ou disparaître de la scène, tout est possible et ne dépend que de l'imagination de l'artiste.

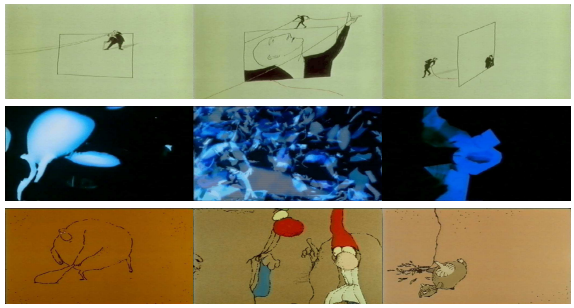


(a) Variété de techniques d'animation

(b) Palettes de couleurs différentes



(c) Effets visuels



(d) Complexité du contenu

FIG. 1.7 – Les particularités des films d'animation (source : les films de CICA [CICA 06]).

- **les personnages** : quand il y en a, ils peuvent avoir n'importe quelle forme, couleur

et caractéristiques.
- **le mouvement** : il peut être discontinu (en particulier dans le cas de certaines techniques d'animation comme par exemple la pâte à modeler). Le mouvement prédominant dans la plupart des situations rencontrées est le mouvement d'objets [Snoek 05b].
- **les techniques d'animation** : différentes techniques d'animation sont utilisées, comme par exemple : le dessin sur celluloïd, l'animation d'objets, les images de synthèse, la pâte à modeler, etc.
- **les effets visuels** : il sont souvent utilisés, mais correspondent à un jeu de techniques bien identifiées et très spécifiques.
- **des concepts artistiques** : ces œuvres sont souvent des projets artistiques reposant sur des volontés artistiques, comme par exemple le choix de la distribution des couleurs du film.
- **palettes couleurs particulières** : la majorité des films d'animation utilisent une palette de couleurs particulière contenant un nombre réduit de couleurs.
- **le contenu** : le contenu des films d'animation présente une variété extrême. Les spécialistes de l'animation disent que 30% des films d'animation du CICA [CICA 06] ne peuvent pas être résumés tant leur contenu est singulier.

Ainsi, la particularité des films d'animation rend la tâche d'analyse difficile. Les méthodes d'analyse utilisées ont donc été adaptées à la spécificité de ces films.

1.6 Conclusions générales

Dans ce chapitre nous avons introduit le concept d'indexation du contenu des données multimédia. Les différents systèmes existants ont été présentés : les systèmes d'indexation d'images CBIR, de séquences d'images CBISR, du son CBAR et de documents vidéo CBVR. Globalement nous retrouvons deux directions distinctes d'analyse :

- **l'analyse syntaxique du contenu** où l'annotation est effectuée avec des descripteurs statistiques de bas-niveau calculés en utilisant differentes sources d'information (la couleur, la forme, la texture, le son, le mouvement, le texte, etc.),
- **l'analyse sémantique** qui propose des descriptions perceptuelles du contenu des données.

L'orientation des systèmes d'indexation actuels se tourne plutôt vers la description sémantique automatique du contenu. Ces systèmes essayent d'optimiser la tâche d'indexation en fournissant aux utilisateurs non-avisés des outils permettant un accès efficace et rapide dans la base en utilisant des critères proches du langage et de la perception humaine. Cependant, cette direction d'analyse pose de nouvelles difficultés. Par exemple, comment sélectionner efficacement parmi la grande variété de paramètres de bas-niveau ceux à utiliser pour réduire la redondance entre les différents canaux d'information disponibles (son, texte, image) ? Un autre aspect qui apparaît très vite est le manque de correspondance entre toutes ces informations dont nous disposons et leur interprétation sémantique, problème connu sous le nom de "fossé sémantique" ("semantic gap").

Les systèmes d'indexation indépendants du domaine d'application sont encore rares. La plupart des systèmes d'indexation sémantique existants utilisent des connaissances a priori sur le contenu des données et sont donc dédiés à un seul type d'application, comme par exemple le domaine sportif, les journaux télévisés, les séries télévisées, etc. Notons que les

approches multi-modales son-image-texte sont préférées aux approches classiques qui utilisent une seule modalité de données.

Les travaux que nous avons développés propose une annotation sémantique du contenu structurel, du mouvement et de la couleur des séquences d'images dans le but d'obtenir un système d'indexation sémantique. Ils sont appliqués et adaptés au domaine particulier des films d'animation. La suite de ce document sera structurée de la manière suivante :

- **Chapitre 2** : s'intéresse à la segmentation temporelle par la détection de transitions vidéo, l'agrégation en plans vidéo, l'annotation visuelle des transitions et le calcul de paramètres de bas niveau des plans.
- **Chapitre 3** : traite de l'analyse du mouvement de caméra et d'objets et du calcul de paramètres de bas niveau du mouvement.
- **Chapitre 4** : propose l'analyse de la distribution des couleurs de la séquence à travers l'histogramme global pondéré de la séquence.
- **Chapitre 5** : présente la problématique du découpage en unités structurelles de plus haut niveau que les plans et discute du découpage en scènes.
- **Chapitre 6** : étudie la construction de résumés statiques et dynamiques et des méthodes d'évaluation de ces résumés.
- **Chapitre 7** : s'intéresse à la description sémantique du contenu de la séquence, particulièrement sur la caractérisation du contenu des plans, des couleurs et du mouvement.
- **Chapitre 8** : étudie l'applicabilité des descripteurs du contenu proposés (tests de classification) dans le moteur de recherche d'un système d'indexation.
- **Chapitre 9** : présente les conclusions finales et les perspectives de nos travaux.

Deuxième partie

La description bas-niveau du contenu

Chapitre 2

Segmentation temporelle

Résumé : *La structure d'une séquence d'images est similaire à celle d'un livre où les différents chapitres sont liés les uns aux autres permettant de définir ainsi le contenu. Dans ce chapitre nous présentons la segmentation temporelle appelée également le découpage en plans d'une séquence d'images. Dans un premier temps, nous analysons les différentes méthodes de détection des transitions vidéo qui existent dans la littérature. Parmi les différentes transitions que nous rencontrons, nous avons approfondi les plus fréquentes, à savoir les "cuts", les "fades" et les "dissolves". Dans un deuxième temps, nous proposons une adaptation de certaines méthodes au cas des films d'animation. Puis, le contenu de la séquence est caractérisé en utilisant un certain nombre de paramètres bas-niveau calculés à partir des statistiques appliquées à la distribution des plans. Enfin, nous proposons une annotation visuelle de la distribution des transitions, ce qui nous aide dans l'analyse de la structure globale de la séquence.*

La segmentation temporelle d'une séquence d'images est définie comme étant le découpage en *unités structurelles* de base de la séquence appelées *plans vidéo*. Toutes les techniques existantes d'analyse sémantique ou syntaxique de documents vidéo ou de séquences d'images utilisent comme point de départ cette segmentation en plans [Lienhart 01b].

Le découpage en plans peut également être vu comme le processus inverse de l'étape de montage, effectuée dans les studios, au montage du film. Les différents plans vidéo sont collés les uns aux autres en utilisant différentes opérations d'édition ou différents types de transitions, pour en définitif, aboutir au film final, processus appelé "final cut".

2.1 La structure temporelle

Du point de vue de la structure temporelle, une séquence d'images peut être représentée sur plusieurs niveaux hiérarchiques illustrés dans la Figure 2.1 :

– **image** : le plus petit niveau de granularité de la séquence qui est représentée par toutes les images contenues dans la séquence.

- **plan vidéo** : correspond à l'ensemble des images qui ont été enregistrées (filmées) entre le moment où la caméra a démarré et le moment où elle s'est arrêtée. La séquence d'images ainsi obtenue présente une continuité visuelle [J.M.Corridoni 95].
- **scène** : comprend un ensemble de plans qui sont liés du point de vue sémantique. Le contenu d'une scène doit respecter la règle des trois unités : unité de lieu, unité de temps et unité d'action [J.M.Corridoni 95].
- **episode** : correspond à l'ensemble des scènes qui sont similaires du point de vue de l'action globale (par exemple les épisodes d'une série télévisée) [Bimbo 99].
- **séquence** : le plus haut niveau hiérarchique représenté par la séquence elle même.

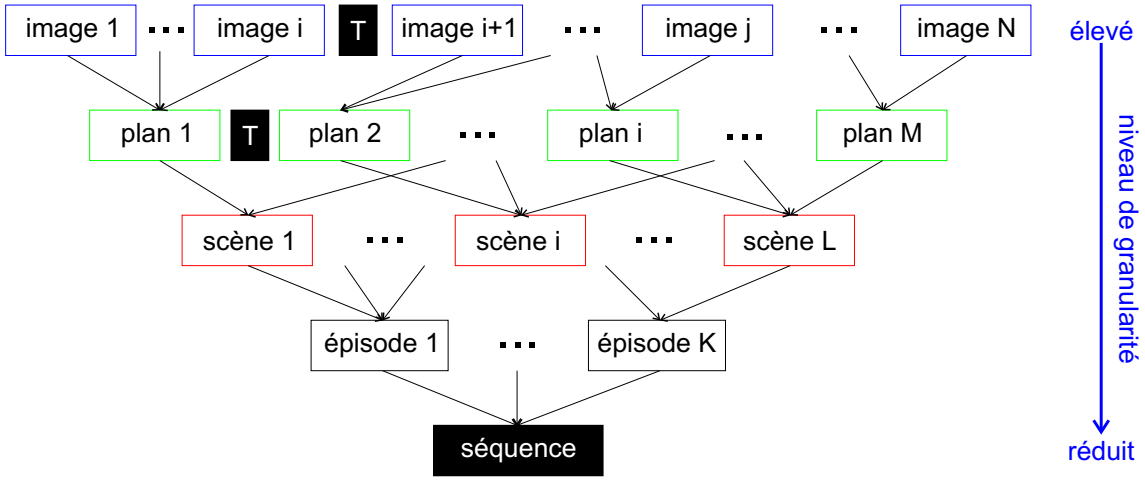


FIG. 2.1 – La structure hiérarchique d'une séquence d'images (T est une transition vidéo).

La plupart des méthodes d'analyse de séquences d'images utilisent comme niveau d'analyse le niveau des plans. Les autres niveaux hiérarchiques, comme les scènes ou les épisodes, nécessitent une analyse sémantique de la perception du contenu et sont plutôt utilisés dans les systèmes d'indexation sémantique. Cela permet une représentation de la séquence à un niveau plus élevé. Nous donnons une description de ces différentes approches sur la détection de scènes dans le Chapitre 5.

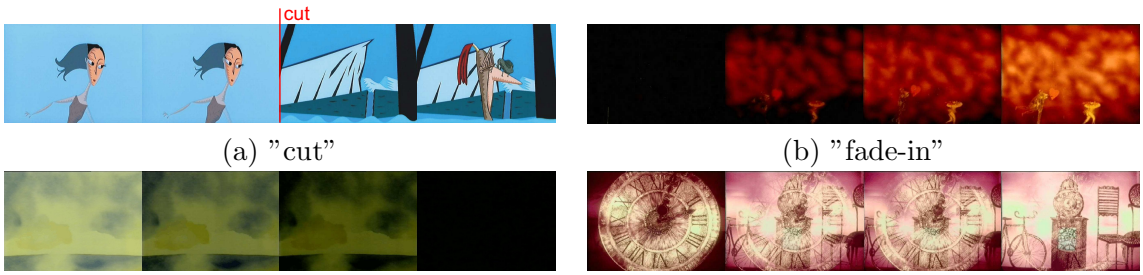
2.2 Les transitions vidéo

Dans une séquence vidéo, les plans sont assemblés les uns aux autres en utilisant des *transitions vidéo* (voir la Figure 2.1). Les transitions vidéo sont des effets visuels qui font le lien entre des plans différents. En fonction des transformations 2D de l'image utilisées, les transitions vidéo peuvent se diviser en 5 classes :

- **la classe d'identité** : les plans ne sont pas modifiés par la transition et il n'y aucune image supplémentaire ajoutée [Lienhart 01b]. Seuls les "cuts" qui sont des transitions abruptes, font partie de cette catégorie. Un "cut" donne une discontinuité visuelle dans la scène car les plans adjacents sont directement assemblés les uns aux autres (voir la Figure 2.2).
- **la classe spatiale** : les transitions appartenant à cette catégorie font subir aux plans des transformations spatiales [Lienhart 01b]. Comme exemple, nous pouvons mention-

ner les "wipes", "mattes", "page turns", "slides", etc.

- **la classe chromatique** : les plans sont modifiés en utilisant des transformations de couleurs [Lienhart 01b], comme les "fades" et les "dissolves" (voir la Figure 2.2). Un "fade" est une transition qui permet de passer d'une image de la séquence à une image qui s'assombrit progressivement jusqu'à ce qu'elle devienne complètement noire (ou plus généralement d'une couleur constante). Ce type de transition est appelée "fade-out". La transition inverse est appelée "fade-in". Un "dissolve" est le plus souvent la superposition d'un "fade-out" et d'un "fade-in".
- **la classe spatio-chromatique** : les transitions appartenant à cette catégorie combinent les deux transformations précédentes, c'est à dire les transformations spatiales et les transformations chromatiques appliquées aux deux plans [Lienhart 01b]. Dans cette catégorie on trouve tous les effets de morphing. Mais, certaines variantes des techniques de la classe chromatique peuvent être considérées aussi comme des transformations spatio-chromatiques (par exemple un "dissolve" comportant un mouvement de caméra).
- **la classe temporelle** : les mouvements de translation de caméra qui font la transition entre deux plans différents peuvent être vus comme un cas particulier de transitions vidéo. Par exemple, nous pouvons mentionner les mouvements particuliers de la caméra de type "tilt" ou "pan" (voir la Section 3.2).



(a) "cut" (b) "fade-in"

(c) "fade-out" (d) "dissolve"

FIG. 2.2 – Exemples de transitions vidéo : (a) film "François le Vaillant" [Folimage 06b], (b) et (d) film "Coeur de Secours" [CICA 06], (c) film "Le Moine et le Poisson" [Folimage 06b].

Mais si nous prenons en compte leur durée, les transitions vidéo se divisent en deux catégories : les *transitions abruptes* comme les "cuts" et les *transitions progressives* comme les "fades" ou les "dissolves". Les "cuts" sont les plus utilisés. Parmi les transitions graduelles, ce sont les "fades" et les "dissolves" qui sont les plus courants. La répartition des différentes transitions dans une séquence n'est pas aléatoire. Chaque transition a une signification sémantique adaptée au contenu de la séquence. Par exemple les "cuts" augmentent le dynamisme de la séquence [Colombo 99], les "dissolves" sont souvent utilisés pour changer le temps de l'action [Lienhart 01b], les "fades" introduisent un changement de lieu ou de temps [J.M.Corridoni 95] et l'ensemble "fade out" suivi d'un "fade-in" introduit une pause avant un changement de l'action.

Dans les séquences que nous avons traitées (voir les films d'animation), les principales transitions vidéo que nous avons rencontrées sont les "cuts", les "fades" ("fade-in" et "fade out") et les "dissolves". Les différentes techniques de détection des "cuts" et des "fades" sont présentées dans les sections suivantes. Pour la détection des "dissolves" nous avons utilisé la méthode proposée dans [Lienhart 01b] basée sur l'analyse des contours.

Nous ne nous attarderons pas sur les autres types de transitions pour lesquels il existe différentes techniques de détection décrites dans [Bimbo 99] pour les "wipes" et "mattes", dans [Song 02] pour les "wipes", dans [Ren 03] pour les mouvements de caméra de type "tilt" et "pan", et pour d'autres transitions vidéo. Une approche statistique globale permettant une détection générique des transitions est également présentée dans [Hanjalic 02].

2.3 L'évaluation de la détection des transitions

L'évaluation de la performance des méthodes de détection des transitions vidéo est réalisée en utilisant les taux de détection. Dans un premier temps une vérité terrain est construite pour servir de référence. Elle consiste à étiqueter manuellement des transitions vidéo dans la séquence. Afin d'analyser finement les résultats de la détection, il faut prendre en compte les transitions qui n'ont pas été détectées, appelées *erreurs de non-détection* et les transitions qui ont été détectées mais qui ne sont pas présentes dans la séquence, appelées *erreurs de fausse détection*.

Une approche consiste à utiliser le taux d'erreurs de détection, noté E_D, correspondant au pourcentage de transitions non-détectées, et le taux d'erreurs de fausse détection, noté E_{FD}, correspondant au pourcentage de fausses détections. Ces taux d'erreurs sont définis par :

$$E_D = \frac{N_t - BD}{N_t}, \quad E_{FD} = \frac{FD}{N_t} \tag{2.1}$$

où N_t est le nombre total de transitions dans la séquence, BD est le nombre de transitions détectées et FD est le nombre de fausses détections. Ainsi, la précision de la détection des transitions est importante si les deux taux d'erreurs sont faibles.

L'approche la plus fréquemment utilisée consiste à calculer des taux de *précision* et de *rappel*. Ces deux taux sont définis par :

$$Précision = \frac{BD}{BD + FD}, \quad Rappel = \frac{BD}{N_t} \tag{2.2}$$

La *précision* est maximale et vaut 100% pour $FD = 0$ (c'est à dire lorsqu'il n'y a pas de fausses détections), c'est en quelque sorte une mesure de la quantité de fausses détections. D'autre part, le *rappel* est une mesure de la quantité de bonnes détections, et est maximum si $BD = N_t$.

2.4 La détection des "cuts"

Les *"cuts"* sont les transitions les plus fréquemment rencontrées dans les séquences d'images. Ils sont définies comme la concaténation directe de deux plans adjacents, $P_1(x, y, t)$ et $P_2(x, y, t)$, et produisent une discontinuité visuelle dans la séquence. La séquence résultante, $S(x, y, t)$ peut être définie formellement par [Lienhart 01b] :

$$S(x, y, t) = (1 - u(t - t_{cut})) \cdot P_1(x, y, t) + u(t - t_{cut}) \cdot P_2(x, y, t) \tag{2.3}$$

où t_{cut} est l'instant correspondant à la première image qui suit le "cut", et $u()$ est la fonction échelon unité définie par $u(t) = 1$ pour $t \geq 0$ et 0 ailleurs.

2.4.1 État de l'art

Plusieurs approches ont été proposées pour mesurer la discontinuité visuelle produite par un "cut". Ces approches font cependant apparaître quelques étapes communes. La première étape de la détection est l'extraction des images d'un certain nombre de *paramètres caractéristiques* des cuts. Dans un deuxième temps certaines *mesures de similarité* sont utilisées pour caractériser les variations des paramètres extraits, calculés entre les images au instants k et $k+l$, $l \geq 1$ étant l'écart temporel ou le pas d'analyse. Ensuite pour détecter la transition, la valeur de discontinuité obtenue est comparée à un certain *seuil* T. Si cette valeur est supérieure au seuil T, alors le "cut" est détecté entre les images k et $k+l$.

Les principaux problèmes liés à ce type d'approche ont été bien mis en évidence dans [Hanjalic 02] :

- **la puissance de discrimination** : tout d'abord la performance de détection est liée à *la puissance de discrimination* des paramètres choisis,
- **la mesure de similarité** : un deuxième problème est lié à *la mesure de similarité* qui doit être importante pour des images similaires faisant partie d'un même plan et qui doit être faible pour des images séparées par un cut.
- **le seuil de détection** : un troisième problème est lié au *choix du seuil* de détection. Un seuil trop bas va augmenter les fausses détections et un seuil trop haut va augmenter le nombre de non détections.

Ces situations sont des sources d'erreurs de détection. Pour améliorer la précision de la détection il faudrait bien connaître les causes qui les produisent.

Les causes de dissimilarité entre images d'un même plan sont les mouvements de caméra, les mouvements des objets ou les changements de l'intensité lumineuse dans les images. Pour se soustraire à ces situations, une solution consiste à utiliser d'autres informations que les mesures de similarité. Par exemple, nous pouvons utiliser la compensation du mouvement pour réduire l'influence du mouvement global de caméra, des mesures statistiques (basées sur les histogrammes) pour réduire l'influence du mouvement des objets, une analyse du voisinage de l'image où nous avons détecté la discontinuité, une analyse des variations de l'intensité lumineuse (par exemple la détection de flashs [Heng 99]) ou d'informations à priori sur l'occurrence de transitions (par exemple la relation entre deux changements de plans consécutifs, qui ne peuvent se produire qu'après un certain laps de temps).

Dans la littérature, de nombreuses méthodes sont proposées pour la détection de "cuts", améliorant ou corrigeant l'influence des différents facteurs énumérés ci-dessus, comme par exemple dans [Bimbo 99], [Lienhart 01b], [W.A.C.Fernando 01], [Hanjalic 02] ou [Ren 03]. Selon le type de caractéristiques utilisées pour mesurer la discontinuité visuelle dans la séquence, les approches existantes se divisent en :

- des méthodes qui utilisent l'analyse de *l'intensité* des pixels ou leur histogramme,
- des méthodes basées sur l'analyse des *contours*,
- des méthodes basées sur l'analyse du *mouvement*,
- des méthodes qui utilisent une analyse dans le format d'origine des films (comme par exemple le format *compressé* MPEG).

Méthodes utilisant l'intensité des pixels

La méthode la plus facile pour mesurer la discontinuité visuelle des "cuts" est de calculer les différences entre les intensités des pixels dans les images aux instants k et $k+l$. Par exemple dans [Otsuji 91] un "cut" est détecté si le nombre de pixels qui ont changé est plus élevé qu'un certain seuil T, c'est-à-dire $N_{pixels} \geq T$ où N_{pixels} est défini de la manière suivante :

$$N_{pixels} = \frac{1}{NM} \sum_{x=1}^{X} \sum_{y=1}^{Y} D_{k,k+l}(x,y) \tag{2.4}$$

où $X \times Y$ est la taille de l'image, et $D_{k,k+l}(x,y)$ est défini par :

$$D_{k,k+l}(x,y) = \begin{cases} 1 & \text{si } |I_k(x,y) - I_{k+l}(x,y)| > T_1 \\ 0 & sinon \end{cases} \tag{2.5}$$

où $I_k(x,y)$ est l'image à l'instant k et T_1 est le seuil lié à l'importance du changement de l'intensité d'un pixel.

Le problème majeur de cette méthode de détection est sa forte sensibilité à la présence de bruit dans l'image ou de mouvements de caméra. Des techniques similaires ont été proposées comme par exemple :

- [Zhang 93] propose d'utiliser un filtrage médian avant de calculer les différences entre les pixels,
- [Boreczky 98] classifie les distances entre pixels et entre intervalles audio en utilisant des modèles de Markov cachés,
- [Kobla 99] utilise comme mesure de dissimilarité la distance Euclidienne calculée dans l'espace couleur YUV ou RGB.

Pour être moins sensible aux transformations géométriques dans l'image, les méthodes de détection de "cuts" utilisent les histogrammes de l'intensité des pixels de l'image. Ceux ci sont calculés soit en niveaux gris, soit en utilisant l'information sur les couleurs présentes dans les images. La technique la plus courante est de mesurer la discontinuité visuelle entre les images k et $k+l$ en utilisant la somme des écarts entre les "bins"[1] des histogrammes couleurs de deux images voisines, selon l'approche proposée dans [Yeo 95]. Pour réduire l'influence des changements d'intensité lumineuse dans l'image, [Furht 95] propose l'utilisation de mesures de similarité entre les histogrammes calculés dans l'espace HVC (H-teinte, V-intensité et C-saturation), pour séparer l'information d'intensité lumineuse de celle de teinte des couleurs. [Arman 93a] propose de calculer les histogrammes en utilisant seulement les composantes H et C, formant une surface 2D, HC, qui est ensuite utilisée pour calculer la discontinuité définie par :

$$D(k,k+l) = \sum_{x=1}^{X} \sum_{y=1}^{Y} |d_{k,k+l}(x,y)| \times \Delta_H \times \Delta_C \tag{2.6}$$

où $d_{k,k+l}(x,y)$ est la différence entre les bins de coordonnées (x,y) appartenant à la surface HC pour les images k et $k+l$, et Δ_H, Δ_C sont les pas de discrétisation des composantes H et C utilisées pour la construction de la surface HC.

Certaines approches de détection des "cuts", basées sur l'analyse des histogrammes utilisent différentes mesures de distances calculées dans différents espaces couleurs : RGB, HSV,

[1] dans toute la suite, on utilisera le terme "bin" pour désigner les différentes valeurs résultant de la quantification des données.

YIQ, Lab, Luv, etc. [Lienhart 01b]. Par exemple [Shen 97] propose l'utilisation de la distance de Hausdorff multi-niveau entre les histogrammes, [Drew 00] propose comme mesure de similarité l'intersection entre les histogrammes en utilisant les distances entre les couleurs dans les espaces Cb-Cr et r-b, [Kim 02] propose le calcul des histogrammes dans l'espace YUV, et [Ma 01] utilise l'intersection entre des histogrammes et la différence entre les couleurs moyennes des blocs de pixels.

D'autres méthodes calculent les histogrammes sur des blocs de pixels pour réduire l'influence du mouvement des objets ou du bruit. Dans [Nagasaka 92] les images k et $k+l$ sont divisées en 16 blocs de pixels et les histogrammes $H_{k,i}$ et $H_{k+l,i}$ sont respectivement calculés pour les blocs $b_i(k)$ et $b_i(k+l)$. Ensuite un test du χ^2 est utilisé pour comparer les histogrammes obtenus :

$$D(i) = \sum_{j=0}^{63} \frac{(H_{k,i}(j) - H_{k+l,i}(j))^2}{H_{k+l,i}(j)} \tag{2.7}$$

où j est l'indice des "bins" de l'histogramme, $j = 0, ..., 63$.

[Gargi 00] a fait une étude comparative des différentes méthodes basées sur les histogrammes couleurs, sur le flux MPEG et sur l'estimation du mouvement par bloc. La meilleure détection a été obtenue en utilisant une méthode basée sur l'histogramme : l'intersection des histogrammes calculés dans l'espace de Munsell (MTM). Du point de vue du temps de calcul, les méthodes utilisant l'analyse d'histogrammes ont une complexité modérée par rapport aux autres méthodes testées.

Méthodes basées sur l'analyse des contours

Ces méthodes utilisent l'analyse des contours pour détecter un "cut". En effet, un "cut" produit une discontinuité structurelle de l'image. Les contours des objets présents dans l'image précédant un "cut" ne se retrouvent pas dans l'image suivant ce "cut". De nombreuses méthodes de détection des "cuts" utilisent donc cette propriété.

Les méthodes proposées dans [Zabih 95] et [Zabih 99] utilisent le rapport des changements des contours, noté ECR ("Edge Change Ratio"), pour la détection d'un "cut". Ce rapport ECR est défini entre l'image k et $k+l$ de la manière suivante :

$$ECR_{k+l} = max(\frac{X_k^{out}}{\sigma_k}, \frac{X_{k+l}^{in}}{\sigma_{k+l}}) \tag{2.8}$$

où σ_k est le nombre des points de contour dans l'image k et X_k^{out}, X_{k+l}^{in} représentent respectivement le nombre de points de contour qui disparaissent dans l'image k et qui apparaissent dans l'image $k+l$.

Pour améliorer l'invariance du rapport ECR à la présence de mouvement dans l'image, [Zabih 95] propose une compensation du mouvement, effectuée avant le calcul du rapport ECR, qui est réalisée en utilisant la distance de Hausdorff. Ainsi, les points de contour analysés dans l'image courante qui sont proches des points de contour dans l'image suivante ne sont pris en compte qu'à partir d'une distance de Hausdorff supérieure ou égale à 6 pixels. D'autres approches ont été proposées dans [Kim 02] utilisant les histogrammes couleurs dans l'espace YUV et le rapport de similarité des contours EMR ("Edge Matching Rate"). Dans [Lienhart 00] plusieurs approches sont développées pour la segmentation de séquences utilisant les points de contours, les histogrammes et le mouvement.

Une comparaison des performances entre les méthodes basées sur l'analyse des contours et celles qui utilisent les histogrammes est présentée dans [Lienhart 01a] et [Lupatini 98]. Les approches contours sont moins efficaces et nécessitent un temps de calcul plus important que les autres méthodes basées sur les histogrammes. Leur principal atout vient du fait que l'approche contours peut être utilisée pour la détection d'autres transitions comme les "fades" ou les "dissolves" [Lienhart 01b].

Méthodes basées sur l'analyse du mouvement

Ces méthodes de détection des "cuts" utilisent l'analyse du mouvement dans la séquence, puisqu'aux abords d'un "cut" il se produit une discontinuité du mouvement. Ces méthodes utilisent la technique du "block matching", méthode employée pour la compensation du mouvement, qui permet de calculer la différence entre blocs de pixels. La procédure consiste pour chaque bloc i de l'image k noté $b_i(k)$, à rechercher le bloc de l'image $k+l$ d'indice j noté $b_{i,j}(k+l)$, qui lui est le plus similaire. La recherche de similarité entre les blocs de pixels se traduit par la minimisation d'une fonction de coût, notée $F_c()$, qui peut être par exemple l'écart entre les valeurs des pixels, une mesure d'erreur absolue quadratique, etc.. Soit $D_{k,k+l}(i)$ l'erreur minimale entre le bloc i de l'image k et le bloc le plus similaire dans l'image $k+l$, alors $D_{k,k+l}(i)$ est définie par :

$$D_{k,k+l}(i) = min|_{j=1,..,N_{cand}} F_c(b_i(k), b_{i,j}(k+l)) \tag{2.9}$$

où N_{cand} est le nombre de blocs $b_{i,j}(k+l)$ de l'image $k+l$, qui peuvent être similaires au bloc $b_i(k)$. Si les images k et $k+l$ sont des images voisines à l'intérieur d'un même plan, les valeurs de l'erreur $D_{k,k+l}()$ sont plutôt faibles. Mais en présence d'un "cut" les valeurs sont très grandes, puisqu'il y a une différence visuelle très importante entre les deux images. Des détails supplémentaires seront présentés sur l'estimation du mouvement dans la Section 3.1.

Dans [Shahraray 95] les images sont divisées en 12 blocs disjoints, et la compensation du mouvement est effectuée en utilisant comme fonction de coût la différence entre les intensités de pixels. Les valeurs de $D_{k,k+l}()$ ainsi obtenues sont triées et normalisées entre 0 et 1, obtenant de nouvelles valeurs notées $d^s_{k,k+l}()$. La mesure de discontinuité entre les images k et $k+l$ est calculée en utilisant des coefficients de pondération c_i, de la façon suivante :

$$D(k, k+l) = \sum_{i=1}^{12} c_i \cdot d^s_{k,k+l}(i) \tag{2.10}$$

Les "cuts" sont ensuite détectés par le seuillage des valeurs de $D(k, k+l)$.

D'autres approches basées sur l'analyse du mouvement, comme celle proposée dans [Porter 00], utilisent la corrélation entre les blocs de pixels comme fonction de coût pour la compensation du mouvement, corrélation qui est calculée dans le domaine fréquentiel. Une autre approche proposée dans [Hanjalic 02] utilise des connaissances a priori sur la distribution de la durée des transitions, sur la compensation du mouvement et sur les amplitudes des changements temporels dans la séquence. Des approches plus complexes, comme les travaux proposés dans [Zhong 96] et [Lupatini 98], utilisent l'estimation du flux optique. Les mesures de similarité entre les images sont calculées à partir de l'estimation des vecteurs de mouvement ou des déplacements dans l'image.

D'une manière générale, les méthodes de détection de "cuts" basées sur l'analyse du mouvement sont moins efficaces que celles basées sur les histogrammes [Gargi 00]. De plus le

temps de traitement nécessaire pour effectuer une estimation du mouvement est important, car cette opération est plus complexe que la détection d'un "cut" à partir des méthodes basées sur les histogrammes [Lienhart 01b]. Elles restent néanmoins intéressantes pour les séquences où le mouvement est prépondérant.

Méthodes développées dans le domaine compressé

Ces approches exploitent directement l'information dans le domaine compressé du flux MPEG, comme par exemple l'analyse des coefficients de la Transformée en Cosinus Discrète (DCT). Un état de l'art est présenté dans [W.A.C.Fernando 01] qui propose également une méthode de détection des "cuts" pour le flux MPEG-2 en utilisant l'analyse du nombre des prédictions des macro-blocs dans des images de type B (compression bidirectionnelle en utilisant des informations sur les images précédentes et suivantes).

La méthode proposée dans [Arman 93b] utilise des sous-blocs de l'image, de taille 8×8 pixels codés par la DCT, choisis à partir de n régions connexes dans l'image courante d'indice k. De plus, pour tous les blocs retenus, seulement 64 coefficients de la DCT sont conservés, choisis aléatoirement parmi l'ensemble des coefficients autres que la composante continue. Chaque image est représentée dans le domaine compressé par un vecteur de coefficients, $V_k = (c_1, c_2, ..., c_{64})$. Les similarités entre les images aux instants k et $k + l$ sont calculées à partir du produit scalaire normalisé des vecteurs par :

$$\Psi_{k,k+l} = \frac{V_k \cdot V_{k+l}}{|V_k| \cdot |V_{k+l}|} \tag{2.11}$$

Un "cut" est détecté si $1 - |\Psi| > T$, où T est le seuil de discontinuité. Des méthodes similaires proposant une détection des "cuts" à partir du flux MPEG sont décrites dans [Zhang 94] et [Meng 95].

L'avantage des méthodes qui utilisent directement le flux MPEG est de ne pas avoir besoin de décompresser les données pour la détection, étape qui est nécessaire dans toutes les autres méthodes utilisant les images. Certains coefficients du flux MPEG contiennent des informations suffisantes pour la détection des discontinuités où des "cuts". La détection peut alors être implantée facilement en temps réel. Notons que la précision de la détection est parfois moins bonne que celle obtenue avec les méthodes basées sur l'image, malgré la mise en place de corrections éliminant les vecteurs mouvement incohérents (voir Section 3.1). Une solution efficace aboutissant à un bon compromis performance/rapidité consiste à travailler sur des données limitées à un certain niveau de détail.

Notons également que le nouveau standard de compression MPEG-7 pourra englober des informations sur la structure temporelle de la séquence, donc les plans, les scènes, etc. Ainsi, l'étape de segmentation ne sera plus nécessaire [Wang 00].

Autres méthodes

Certaines méthodes utilisent différentes informations pour mesurer la similarité des images. La méthode proposée dans [Boreczky 98] transforme le problème de détection en un problème de classification. Il propose d'utiliser les modèles de Markov cachés ou HMM pour la segmentation en plans et en même temps pour la classification. Les différents états du HMM sont utilisés pour modéliser les différents types de segments de la séquence. L'utilisation

de HMM pour la segmentation en plans est détaillée dans [Wang 00]. Une méthode statistique indépendante de la séquence, est proposée dans [Hanjalic 02], où la minimisation de la probabilité de l'erreur moyenne de détection est utilisée pour la segmentation en plans.

Une approche différente est proposée dans [Guimaraes 03]. Premièrement chaque image de la séquence est résumée par une seule ligne de pixels correspondant à la diagonale principale de l'image. Puis, une image est construite en juxtaposant verticalement l'ensemble des lignes retenues pour la séquence entière. Les "cuts" sont représentés dans cette nouvelle image illustrant le rythme visuel comme des transitions verticales. Ensuite, des méthodes de traitement d'images, comme la détection de contours, et des opérations de morphologie mathématique, sont appliquées pour faciliter la détection de ces transitions représentant les "cuts" de la séquence.

Ces méthodes, relativement récentes, ont généralement été appliquées et testées dans des applications bien spécifiques. Actuellement, on ne dispose pas de suffisamment de tests comparatifs pour dire que ces méthodes sont meilleures que les méthodes classiques (voir les approches basées sur l'intensité des pixels, sur les contours et le mouvement).

Les méthodes de calcul du seuil de détection

Les plupart des mesures de discontinuité d'un "cut" utilisent un ou plusieurs seuils de détection. Le choix du seuil est essentiel pour la qualité de la détection. Un seuil trop bas va augmenter le nombre de fausses détections et un seuil trop haut va augmenter le nombre de non-détections. Un état de l'art concernant le calcul du seuil de détection est présenté dans [Lienhart 01b] et [Hanjalic 02].

Les premières approches sur le calcul du seuil étaient basées sur des *approches heuristiques* comme celles de [Otsuji 91], [Nagasaka 92] ou [Arman 93b]. D'autres méthodes proposent une *analyse statistique* de la distribution des valeurs des discontinuités. [Zhang 93] propose de modéliser cette distribution par des fonctions Gaussiennes de moyenne μ et de variance σ^2. Le seuil de détection T est défini par :

$$T = \mu + r \cdot \sigma \tag{2.12}$$

où le paramètre r est lié à une probabilité de fausses détections a priori fixée.

Les deux approches précédentes (l'approche heuristique et statistique) proposent un seuil global pour la séquence. Une autre méthode est de calculer le seuil d'une manière *adaptative*. [Yeo 95] calcule la valeur du seuil T en fonction de l'information temporelle dans la séquence. L'analyse est faite avec un pas $l = 1$ dans des fenêtres temporelles de N valeurs. Un "cut" est détecté dans le milieu de la fenêtre courante analysée si la discontinuité $D(k, k+1)$ est maximale :

$$D(k, k+1) = max|_{i=-\frac{N}{2},\ldots,\frac{N}{2}}\{D(k+i, k+1+i)\} \tag{2.13}$$

et

$$D(k, k+1) \geq \alpha \cdot D_{smax} \tag{2.14}$$

où $D(k, k+1)$ est la discontinuité entre l'image aux instants k et $k+1$, D_{smax} est la deuxième valeur maximale de la discontinuité dans la fenêtre N et α est un paramètre lié à la forme du signal de discontinuité. Pour plus de détails sur les méthodes adaptatives voir [Gargi 00] et [Truong 00b].

D'autres approches sont les *approches mixtes* qui combinent les méthodes adaptatives avec les approches statistiques. [Hanjalic 97] propose une analyse dans des fenêtres temporelles à l'aide de la modélisation Gaussienne de la distribution des valeurs de discontinuité.

Le paramètre α est déterminé en utilisant la méthode proposée dans [Zhang 93] basée sur l'analyse de la probabilité a priori de fausses détections.

Le calcul d'un *seuil optimal* est une autre direction d'étude. Ce type d'approche s'appuie sur la théorie de la détection statistique. Des connaissances statistiques a priori sur la distribution des "cuts" ont été déterminées à partir de l'analyse d'un nombre suffisamment important de séquences. Puis, celles-ci sont utilisées pour la détection. La loi de détection de la discontinuité est calculée par la minimisation de l'erreur de détection [Vasconcelos 00].

2.4.2 Les pré-traitements utilisés pour les méthodes de détection des "cuts" développées

Nous avons développé des algorithmes de détection de "cut" adaptés aux particularités des films d'animation (voir la Section 1.5). Les méthodes proposées sont basées sur l'analyse des histogrammes couleurs, ces méthodes étant plus efficaces que les approches basées sur l'analyse du mouvement ou l'analyse des contours (voir la Section 2.4.1 sur l'état de l'art). La dissimilarité visuelle introduite par les "cuts" est transformée en une distance entre histogrammes couleurs. Avant que la détection soit effectuée, un certain nombre de prétraitements doivent être réalisés : les *sous-échantillonnages temporel et spatial* et une *réduction des couleurs*.

Le sous-échantillonage

Un premier traitement consiste à faire un *sous-échantillonnage temporel* de la séquence pour réduire la redondance temporelle et par conséquent le temps de traitement. Cette étape est motivée par le fait qu'une seconde de séquence correspond typiquement à 25 images et que les "cuts" se produisent généralement dans des intervalles temporels de plus de 3 à 4 secondes dans le cas des films d'animation, et des intervalles plus longs dans le cas des films naturels. Donc, les images sont analysées avec un pas d'analyse, noté l comme dans l'état de l'art (voir la Section 2.4.1), défini comme étant l'écart temporel entre deux images successives retenues.

Dans un deuxième temps les images sont *sous-échantillonnées spatialement* pour réduire le temps de traitement. Comme les méthodes proposées de détection des "cuts" utilisent des mesures statistiques, la qualité de la détection n'est pas sensiblement détériorée par l'utilisation des images de plus basse résolution. Ainsi, dans chaque image analysée, pour chaque bloc de pixels sans chevauchement, de taille $n \times n$ (avec $n \in \{2, 3, 4\}$ selon l'application) seul le pixel central est conservé. Ceci permet d'avoir une résolution de l'image d'environ 100×100 pixels.

La réduction des couleurs

Classiquement, représentée sur 3x8 bits (représentation RVB), la couleur peut donc prendre plus de 16 millions de valeurs. Dans la plupart des applications, *la réduction des couleurs* est une étape préalable indispensable. En général, les méthodes de réduction des couleurs diminuent le nombre des couleurs utilisées tout en minimisant la perte de qualité visuelle. Ces méthodes sont basées sur le fait que l'œil humain ne perçoit pas les petites variations de couleur. On peut ainsi modifier la couleur de certains pixels sans modification majeure de la perception visuelle. Les méthodes existantes utilisent des approches basées sur

la logique floue, les réseaux neuronaux ou les algorithmes génétiques de façon à obtenir des images quantifiées de très bonne qualité [Kanjanawanishkul 05]. Il faut également noter que ces méthodes, habituellement, cherchent un compromis entre la qualité de la préservation des couleurs et le temps de calcul, compromis qui dépend du type d'application.

En général, la réduction couleur se déroule en deux étapes : la construction d'une palette puis l'attribution d'une des couleurs de la palette pour chaque couleur de l'image ce qu'on appelle le "pixel mapping". Il y a deux catégories de quantification des couleurs : les approches qui utilisent une *palette fixe* (universelle) et les approches *adaptatives*. Les approches basées sur des palettes fixes nécessitent un temps de calcul réduit mais la qualité visuelle est moyenne puisqu'elle est liée à la taille et à la diversité des couleurs de la palette utilisée, et qu'elle n'est pas choisie en fonction du contenu de l'image. Les approches adaptatives, quant à elles, déterminent une palette optimale de couleurs en minimisant la perte de la qualité visuelle. Un état de l'art sur les méthodes de réduction des couleurs est présenté dans [Trémeau 04] et [Kanjanawanishkul 05].

Les approches adaptatives déterminent pour chaque image une palette optimale particulière, opération qui nécessite un temps de calcul souvent important. De plus, chaque image étant représentée par une palette spécifique, le nombre total de couleurs obtenu sur l'ensemble des images différentes peut être très élevé et contient probablement des variations faibles de la même couleur. Comparer les couleurs devient alors une tâche difficile qui demande des mesures efficaces de similarité entre les couleurs comme par exemple la mesure "Earth Mover's Distance" [Rubner 97] ou "quadratic form distance" [Hafner 95].

Généralement dans les méthodes de détection des "cuts" le soucis n'est pas de préserver fidèlement la qualité visuelle de l'image mais de conserver les différences de couleurs pouvant exister entre les images. Nous nous sommes donc limités à l'utilisation des approches basées sur une palette fixe permettant une comparaison plus rapide entre les histogrammes. De plus, chaque film d'animation utilisant généralement une petite quantité de couleurs différentes (voir la Section 1.5) l'utilisation d'une palette fixe n'altère pas de trop la qualité visuelle.

Nous avons envisagé plusieurs méthodes de réduction des couleurs selon la qualité visuelle de la quantification et le temps de calcul nécessaire à cette réduction :

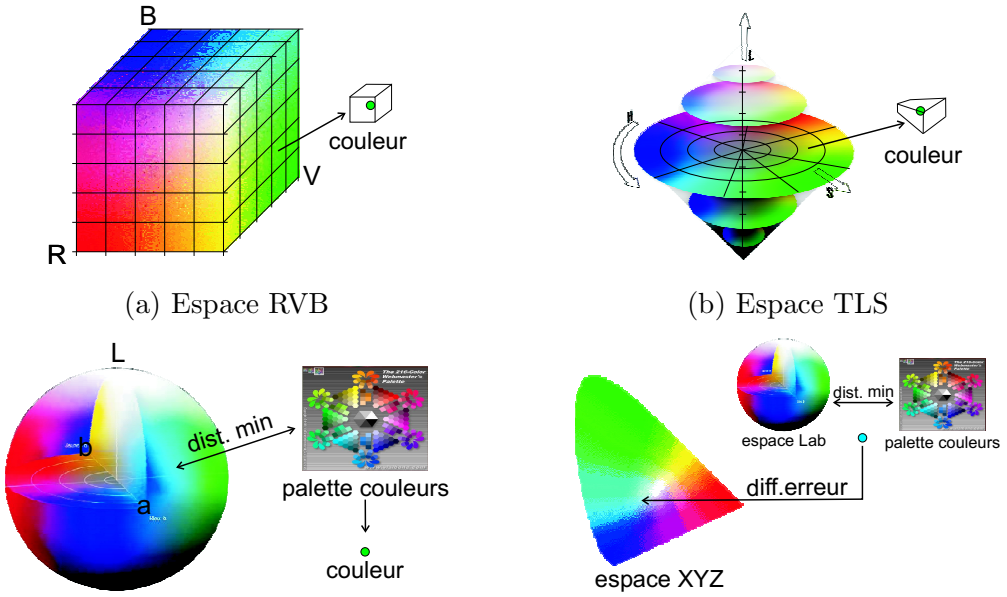
A. La réduction des couleurs dans l'espace RVB. Une quantification uniforme dans l'espace RVB est proposée. En gardant 5 couleurs sur chaque composante (R-rouge, V-verte et B-bleu) nous disposons de 125 couleurs. Le cube RVB est divisé en $5 \times 5 \times 5$ cubes de couleur qui seront approximés par la couleur centrale ou le centroïde (voir la Figure 2.3.a). C'est la stratégie la plus rapide du point de vue du temps de calcul, mais il existe des différences perceptuelles avec les images d'origine, certaines couleurs n'étant pas toujours bien restituées avec cette palette.

B. La réduction des couleurs dans l'espace TLS. Une quantification de l'espace TLS (T-teinte, L-luminance et S-saturation) est proposée. L'espace TLS, représenté sous la forme d'un double cône, est basé sur la perception humaine de la couleur. Les axes de la teinte et de la luminance sont divisés en 7 intervalles et celui de la saturation en 4 intervalles, l'œil humain étant plus sensible à la teinte et à la luminance, qu'à la saturation. De plus, 9 niveaux de gris, incluant le noir et le blanc, sont ajoutés, ce qui permet d'obtenir une palette fixe à 205 couleurs (voir la Figure 2.3.b). Pour la réduction des couleurs on utilise l'Algorithme 1, où les valeurs de T, L et S sont normalisées entre 0 et 1.

Algorithm 1 Réduction couleur dans l'espace TLS

si $(L < 0.08)$ **alors**
  $c \leftarrow Noire$ {c est la couleur à quantifier, L, S sont les composantes de luminance et de saturation}
sinon si $(L > 0.9)$ **alors**
  $c \leftarrow Blanc$
sinon si $(S < 0.08)$ **alors**
  $c \leftarrow NiveauGris$ {on se retrouve dans la situation des niveaux gris et on utilise un des 7 niveaux de quantification}
sinon
  $c \leftarrow Couleur$ {on se retrouve dans la situation où c est une couleur pertinente. Elle est alors approximée par la couleur du milieu du secteur 3D (défini par la division des axes TLS mentionnés ci-dessus) auquel elle appartient}
fin si

Les valeurs des seuils ont été déterminées par une analyse empirique précise de l'espace TLS. Nous nous sommes limités à un nombre faible de couleurs pour garder un bon compromis entre la qualité visuelle et le nombre de couleurs utilisées, même s'il est préconisé d'utiliser une palette de plus de 561 couleurs ($10 \times 11 \times 5$ plus 11 niveaux gris) pour conserver un haut niveau de qualité visuelle (voir [Bimbo 99]).



(a) Espace RVB (b) Espace TLS

(c) Espace Lab (d) Diffusion d'erreur

FIG. 2.3 – Les différentes méthodes de réduction des couleurs proposées.

C. La réduction des couleurs dans l'espace Lab. On utilise une palette appelée "Webmaster", définie a priori [Visibone 06]. Cette palette comporte 216 couleurs. Pour plus de détails sur cette palette voir la Section 4.2.3. Chaque couleur de l'image est remplacée par la couleur la plus proche de la palette, mesurée avec la distance Euclidienne calculée dans

(a) Image originale (b) Espace RVB (c) Espace TSL

(d) Espace Lab (e) Diffusion d'erreur

FIG. 2.4 – Exemples de réduction des couleurs (image du film "A Viagem" [CICA 06]).

l'espace Lab (voir la Figure 2.3.c). L'espace Lab présente une meilleure uniformité de la perception visuelle que les autres espaces. En particulier, il donne une bonne concordance entre les écarts perceptuels entre deux couleurs et leur distance Euclidienne.

D. La diffusion d'erreur dans l'espace XYZ. On utilise la même palette que pour la réduction dans l'espace Lab présentée ci-dessus. Les couleurs sont modifiées en tenant compte de leur voisinage spatial en utilisant l'algorithme de Floyd and Stenberg appliqué dans l'espace XYZ [Evans 03] :

- d'abord la couleur c de chaque pixel est remplacée par la couleur c_{min} la plus proche dans la palette "Webmaster" en utilisant l'approche présentée dans le paragraphe précédent.
- ensuite, chaque couleur est légèrement modifiée en tenant compte d'une part du voisinage du pixel courant et d'autre part de l'écart mesuré dans l'espace XYZ entre la couleur initiale et la couleur sélectionnée dans la palette "Webmaster". Le détail de l'algorithme de la diffusion d'erreur dans l'espace XYZ est donné en Annexe A.

Il faut noter que le traitement effectué assure, au bout du compte, de n'obtenir que des couleurs de la palette choisie, même si provisoirement d'autres couleurs sont utilisées dans l'algorithme. La diffusion permet d'atteindre une meilleure qualité visuelle au prix d'une plus grande complexité.

Dans le Tableau 2.1 nous avons indiqué la qualité visuelle des méthodes proposées, évaluée de manière subjective par une note allant de 1 (mauvais) à 4 (bon), ainsi que le temps de calcul mesuré sur un processeur PentiumM 1.6 GHz pour des images de taille 178×81 pixels. Un exemple de la qualité visuelle des images ayant subi une réduction couleur par chacune des méthodes proposées est présenté dans la Figure 2.4.

Du point de vue du temps de calcul, les deux dernières méthodes sont beaucoup moins rapides puisqu'elles demandent, pour chaque couleur de l'image, la recherche de la couleur de la palette la plus proche, mais la qualité visuelle est bien meilleure. L'influence de la réduction des couleurs sur la détection des "cuts" est abordée dans la Section 2.4.4.

Réduction	Espace RVB	Espace TSL	Espace Lab	Diffusion d'erreur
Qualité	2	1	3	4
Tps. calcul	0.007s	0.017s	4.1s	4.2s

TAB. 2.1 – Les performances des méthodes de réduction des couleurs proposées.

2.4.3 Les méthodes de détection des "cuts" développées

Dans la suite nous allons détailler les méthodes de détection des "cuts" que nous avons proposées. Nous avons développé deux algorithmes de détection : la méthode *"4histogrammes"* qui est une approche classique basée sur l'analyse des distances entre histogrammes et la méthode *"2dérivées"* qui propose une amélioration de la première méthode à l'aide de la dérivée seconde. Les deux méthodes sont adaptées aux particularités des films d'animation.

La méthode "4histogrammes"

La première méthode de détection des "cuts" que nous avons développée, appelée la méthode *4histogrammes* (voir [Ionescu 05f]), utilise l'approche classique qui mesure la discontinuité visuelle d'un "cut" en utilisant la distance entre les histogrammes couleurs issus des images de la séquence. Elle servira plutôt de point de référence pour les autres méthodes proposées.

Avant la détection, la séquence est dans un premier temps sous-échantillonnée temporellement et spatialement et les couleurs sont réduites avec l'une des méthodes proposées dans la section précédente.

Un problème important qui survient dans les films d'animation, le mouvement des objets étant prédominant [Snoek 05b], est le déplacement (apparition/disparition) d'objets dans la scène. Ces objets ont typiquement une taille suffisamment élevée pour produire des différences importantes entre les histogrammes couleurs des images successives. Pour réduire l'influence de ces objets en mouvement sur les performances de la méthode proposée, les images sont divisées en 4 quadrants.

L'algorithme de détection proposé est le suivant (voir la Figure 2.5) :

- pour chaque image courante analysée, $\widehat{image}_{k\cdot l}$, à l'instant $k \cdot l$, k étant un nombre entier, l le pas d'analyse (le chapeau dénote le sous-échantillonnage spatial et la réduction des couleurs), et l'image voisine suivante, $\widehat{image}_{(k+1)\cdot l}$, on calcule quatre histogrammes couleurs correspondant aux 4 quadrants de chacune des images, notés respectivement $H^j_{k\cdot l}$ et $H^j_{(k+1)\cdot l}$, avec j l'indice du quadrant, $j \in \{1, 2, 3, 4\}$,
- ensuite, on calcule les 4 distances Euclidiennes entre les histogrammes des mêmes quadrants de l'image courante $\widehat{image}_{k\cdot l}$ et de l'image suivante $\widehat{image}_{(k+1)\cdot l}$:

$$d^j_E(k) = \left(\sum_{c=1}^{N_c} \left[H^j_{(k+1)\cdot l}(c) - H^j_{k\cdot l}(c) \right]^2 \right)^{1/2} \tag{2.15}$$

où N_c est le nombre des couleurs et c l'indice des couleurs.

En utilisant le même principe pour toutes les images retenues de la séquence, on obtiendra 4 ensembles de distances, $d^j_E(k)$, avec $j \in \{1, 2, 3, 4\}$ et $k = 0, ..., [\frac{N_{seq}}{l}]$ avec N_{seq} le nombre

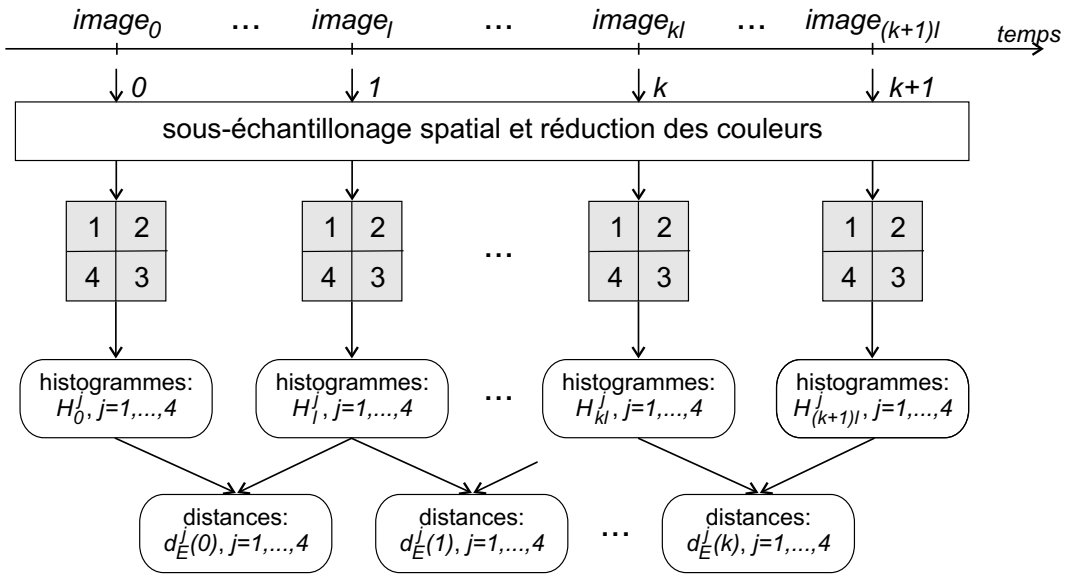


FIG. 2.5 – Principe de détection de la méthode *4histogrammes* (les quadrants sont notés de 1 à 4).

total d'images de la séquence.

La détection d'un "cut" est effectuée en analysant ces ensembles. Un "cut" sera détecté au moment $k \cdot l$ si la condition :

$$d_E^j(k) > \tau_{cut} \quad \textbf{et} \quad d_E^j(k+1) < \tau_{cut} \tag{2.16}$$

est vérifiée sur au moins 3 des 4 quadrants, c'est-à-dire au moins pour 3 des 4 valeurs de j. τ_{cut} est le seuil de similarité entre histogrammes, il sera défini plus tard dans cette section. La condition exposée se traduit par : un "cut" est détecté si l'image courante au moment $k \cdot l$ et l'image suivante au moment $(k+1) \cdot l$ sont différentes, et aussi si l'image au moment $(k+1) \cdot l$ et l'image au moment $(k+2) \cdot l$ sont similaires. Cette condition est motivée par le fait qu'un "cut" commence avec une différence forte entre deux image voisines et continue avec des images similaires, au moins sur une petite période de temps. Des résultats expérimentaux sont présentés dans la Section 2.4.4.

Nous avons effectué une étude permettant de connaître l'influence de la taille d'un objet sur l'histogramme global. Le test a été effectué en utilisant deux situations : l'image contient un objet dont la taille correspond à peu près à un quadrant de l'image ($image_{1/4}$), et l'image contient un objet plus petit, d'une taille voisine du 1/16 de la taille de l'image, ($image_{1/16}$). L'image de référence est l'image sans objet, $image_{ref}$ (voir la Figure 2.6). Pour comparer les images on calcule des distances Euclidiennes entre les histogrammes des images $image_{1/4}$ et $image_{1/16}$, donc $H_{1/4}()$ et respectivement $H_{1/16}()$, et l'histogramme de l'image de référence, $H_{ref}()$. Les résultats obtenus sont présentés dans le Tableau 2.2.

Seuls les objets d'une taille égale ou supérieure à la taille d'un quadrant introduisent des changements importants entre les histogrammes et donc de fausses détections. La distance obtenue avec l'image $image_{1/4}$, 0.3, est plus élevée que le seuil de détection utilisé, dont la valeur est typiquement inférieure à 0.2 (voir la Section 2.4.3). Nous avons donc choisi de diviser les images en 4 quadrants.

(a) $image_{ref}$ (b) $image_{1/4}$ (c) $image_{1/16}$

FIG. 2.6 – Images avec des objets de différente taille (les lignes blanches délimitent les quadrants).

image	$image_{ref}$	$image_{1/4}$	$image_{1/16}$
$d_E(H_n, H_{ref})$	0	**0.3**	0.04

TAB. 2.2 – Les distances Euclidiennes entre les histogrammes correspondent à des images comportant des objets de différentes tailles.

La méthode "2dérivées"

La deuxième méthode que nous avons développée, est appelée la méthode *2derivées* (voir [Ionescu 06a]), qui comporte certaines améliorations de la méthode *4histogrammes*. Dans un premier temps les 4 ensembles de distances, $d_E^j()$, avec $j \in \{1, 2, 3, 4\}$, sont remplacés par l'ensemble correspondant à la moyenne des 4 distances, que nous notons $\bar{d}_E()$, et qui est calculée de la manière suivante :

$$\bar{d}_E(k) = \frac{1}{4} \sum_{j=1}^{4} d_E^j(k) \tag{2.17}$$

où $k = 0, ..., [\frac{N_{seq}}{l}]$ avec N_{seq} le nombre total d'images de la séquence.

Plusieurs facteurs nous ont conduit à utiliser un seul vecteur de distances :

- **décision simplifiée** : premièrement, la méthode de décision de détection d'un "cut" est *simplifiée* puisqu'il n'y a plus qu'une valeur à comparer à un seuil,
- **décision nuancée** : deuxièmement, la décision sur les 4 valeurs de distances, utilisée par la méthode *4histogrammes*, est remplacée par *une décision nuancée*. Par exemple, si 2 des 4 valeurs de distance sont fortes (supérieures à τ_{cut}) et 2 sont faibles (inférieures à τ_{cut}), la décision majoritaire impose qu'il n'y a pas de "cut". En utilisant la valeur moyenne des 4 distances, selon ces valeurs, on peut avoir un écart moyen supérieur au seuil de détection et détecter le "cut".
- **meilleurs résultats** : troisièmement, le calcul du seuil de détection, τ_{cut}, sur la valeur moyenne des distances a donné de *meilleurs résultats* de détection que les autres approches (comme nous le verrons dans la présentation des résultats expérimentaux de la Section 2.4.4).

La présence de mouvements répétitifs d'objets ou de caméra dans la scène est une des sources de fausses détections dans la séquence. Cela introduit des différences significatives, et ceci de façon répétitive. Comme nous l'avons déjà mentionné dans la méthode précédente, un "cut" est représenté par une forte dissimilitude entre deux images consécutives, donc une

valeur élevée de $\bar{d}_E(k)$, suivie par une forte similarité entre images, donc une valeur faible de $\bar{d}_E(k+1)$.

Ces observations nous ont donné l'idée d'utiliser *la dérivée* du vecteur $\bar{d}_E()$ pour mieux localiser les "cuts" dans la séquence et donc pour réduire l'influence du mouvement sur la détection. La dérivée est estimée par une simple différence. Elle permet de préserver les différences importantes de $\bar{d}_E()$ (présence d'un "cut") et a l'avantage de donner un résultat faible dans le cas de valeurs successives élévées de $\bar{d}_E()$ (mouvement important à l'intérieur d'un plan). L'effet de la dérivée appliquée au vecteur $\bar{d}_E()$ est présenté par la Figure 2.7.

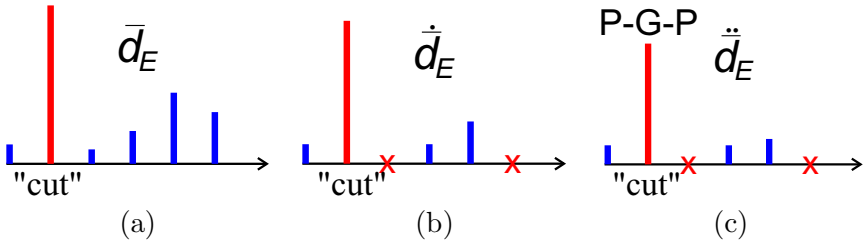


FIG. 2.7 – L'amélioration de la détection de "cuts" obtenue par l'utilisation de la dérivée seconde : (a) $\bar{d}_E()$, (b) la dérivée première $\dot{\bar{d}}_E(k)$, (c) la dérivée seconde $\ddot{\bar{d}}_E(k)$. L'axe oX correspond au temps, les valeurs négatives (marquées avec le $\times$ rouge) sont mises à 0. P-G-P est la signature d'un "cut", valeurs Petite-Grande-Petite.

Pour les mêmes raisons, si on itère ce processus de dérivation on améliore à nouveau la détection. Cependant, la dérivation amplifiant généralement le bruit, l'ordre de dérivation doit être limité. Des tests ont montré que l'utilisation de la dérivée seconde ($n = 2$) est le meilleur compromis permettant d'obtenir un taux de détections correct et un taux de fausses détections faible.

La dérivée première, $\dot{\bar{d}}_E()$, est calculée par :

$$\dot{\bar{d}}_E(k+1) = \begin{cases} \bar{d}_E(k+1) - \bar{d}_E(k) & \text{si } \bar{d}_E(k+1) \geq \bar{d}_E(k) \\ 0 & \text{sinon} \end{cases} \tag{2.18}$$

où k est l'indice temporel. La dérivée seconde, $\ddot{\bar{d}}_E()$, est obtenue de la même manière à partir de la dérivée première. Les valeurs négatives de la dérivée sont mises en 0, puisqu'elles contiennent des informations redondantes pour la détection (voir la Figure 2.7). Les "cuts" sont vus comme des maxima dans la séquence de $\ddot{\bar{d}}_E(k)$.

Dans la Figure 2.8 nous présentons un exemple de réduction de l'influence du mouvement de la caméra sur la délimitation des "cuts" à l'aide de la dérivée seconde. Dans l'ensemble $\ddot{\bar{d}}_E()$ (l'image du bas) les différences répétitives causées par le mouvement de la caméra ont été réduites tout en préservant les transitions abruptes ("cuts").

La détection d'un "cut" est effectuée par le seuillage de l'ensemble $\ddot{\bar{d}}_E()$. De façon similaire à la méthode *"4histogrammes"*, un "cut" est détecté à l'instant $(k+2) \cdot l$ dans la séquence si la condition suivante est vérifiée :

$$\ddot{\bar{d}}_E(k) > \tau_{cut} \quad \textbf{et} \quad \ddot{\bar{d}}_E(k+1) < \tau_{cut} \tag{2.19}$$

où τ_{cut} est le seuil de détection. Le décalage $k+2$, au lieu de k, s'explique par le fait que dans l'ensemble des valeurs des distances l'indice de la position du "cut" se décale de 1 dans

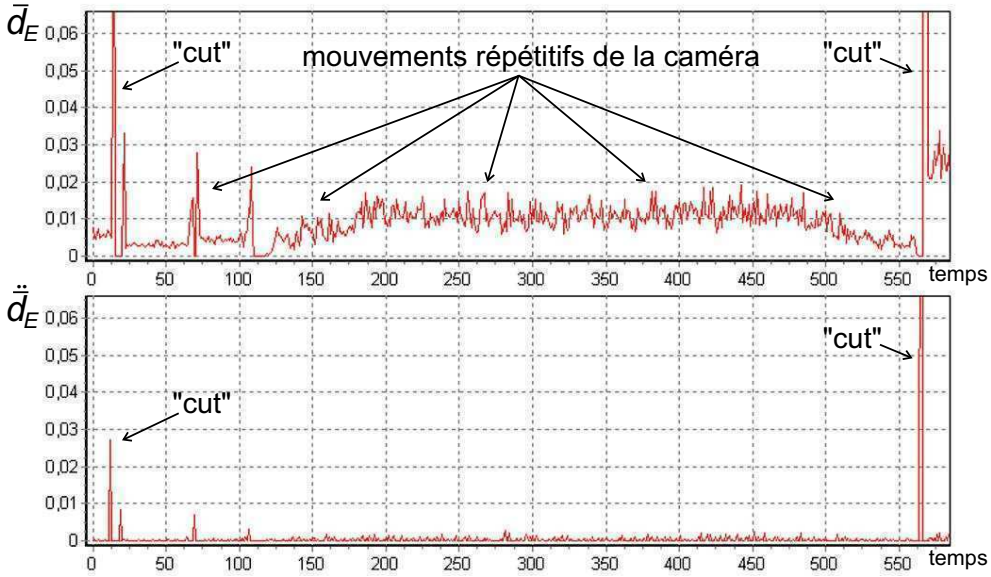


FIG. 2.8 – L'amélioration de la délimitation des "cuts" obtenue en utilisant la dérivée seconde : l'ensemble $\bar{d}_E()$ (en haut) et l'ensemble $\ddot{\bar{d}}_E()$ (en bas). On peut remarquer la robustesse de la méthode même en présence de mouvements répétitifs de la caméra.

le sens négatif pour chaque dérivée. Donc si la discontinuité visuelle (le "cut") se trouve en k dans l'ensemble $\ddot{\bar{d}}_E()$, dans l'ensemble $\dot{\bar{d}}_E()$ elle se retrouve en $k+1$ et, par conséquent, en $k+2$ dans l'ensemble $\bar{d}_E()$.

Le calcul du seuil de détection

Une procédure d'estimation automatique du seuil de détection, τ_{cut}, a été mise en place. L'utilisation d'un seuil global pour l'ensemble des séquences est pratiquement impossible à déterminer [Lienhart 99a], car chaque film d'animation a sa propre palette de couleurs. Seule une approche adaptative est envisageable. La méthode proposée (voir [Ionescu 06a]) est inspirée de l'approche basée sur la modélisation Gaussienne de la distribution des discontinuités visuelles dans la séquence, présentée dans l'équation 2.12.

Dans un premier temps on calcule la valeur moyenne du vecteur $\bar{d}_E()$, par :

$$m_{\bar{d}} = \frac{1}{N_k} \sum_{k=0}^{N_k-1} \bar{d}_E(k) \tag{2.20}$$

où $N_k = [\frac{N_{seq}}{l}] + 1$ est le nombre de valeurs de l'ensemble $\bar{d}_E()$, avec N_{seq} le nombre total d'images de la séquence et l le pas d'analyse. En calculant le seuil de la façon proposée dans l'équation 2.12 nous obtenons :

$$T^G_{cut} = m_{\bar{d}_E()} + \alpha \cdot \sigma_{\bar{d}_E()} \tag{2.21}$$

où $\sigma_{\bar{d}_E()}$ est l'écart type de $\bar{d}_E()$.

Comme les "cuts" (discontinuités) sont moins probables que les autres variations, le seuil T^G_{cut} est en général trop bas (voir la Figure 2.9) et donc le nombre de fausses détections est

élevé. De plus, le réglage du seuil T_{cut}^{G} est lié au bon réglage du paramètre α qui pose les mêmes difficultés que le choix du seuil lui même.

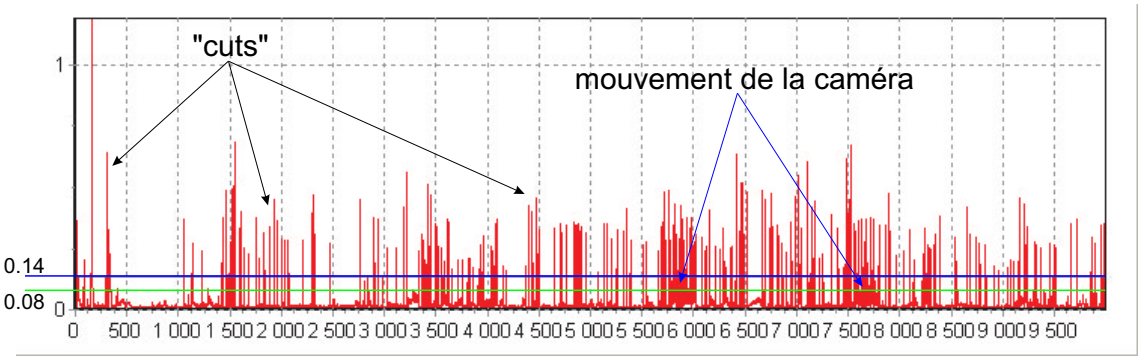


FIG. 2.9 – Exemple d'estimation du seuil calculé sur les valeurs de $\bar{d}_E()$: les maxima locaux correspondent à des "cuts", le seuil proposé T_{cut} est marqué avec la ligne bleue, la ligne verte correspond au seuil T_{cut}^{G} pour $\alpha = 1$ (extrait du film "A Bug's Life").

Notre approche est différente. Le seuil τ_{cut} est estimé en analysant les maxima locaux du vecteur $\bar{d}_E()$. On définit un maximum local comme étant la valeur $\bar{d}_E(k)$ qui vérifie les conditions suivantes :

$$\bar{d}_E(k) > m_{\bar{d}} \quad \textbf{et} \quad \bar{d}_E(k-1) < \bar{d}_E(k) \quad \textbf{et} \quad \bar{d}_E(k+1) < \bar{d}_E(k) \tag{2.22}$$

où k est l'indice des valeurs de l'ensemble $\bar{d}_E()$.

Ensuite le seuil τ_{cut} *est déterminé comme la valeur moyenne de tous les maxima locaux* (voir la Figure 2.9). Pour la méthode *2dérivées* le seuil est calculé de la même façon, mais en utilisant le vecteur $\ddot{\bar{d}}_E()$ à la place du vecteur $\bar{d}_E()$. Le seuil proposé a donné de bons résultats sur plusieurs séquences. Les résultats expérimentaux sont présentés dans la Section 2.4.4.

Amélioration par la détection de SCC

Une des particularités des films d'animation est la présence d'effets de couleurs. Un effet particulier, appelé SCC ou "short color change" correspond à un changement très brutal de la couleur, comme un flash, un éclair, etc. (voir la Section 2.5). Leur présence dans la séquence est souvent détectée comme un "cut".

Une amélioration possible de la méthode *2dérivées* de détection des "cuts" est donc la détection des SCC par l'algorithme proposé dans la Section 2.5. Une fois qu'un "cut" est détecté, on vérifie si c'est un SCC, sinon on l'assimile à un "cut".

Dans les films naturels une situation similaire existe, en présence de flashs. Généralement, c'est l'analyse des variations de l'intensité lumineuse dans la scène qui permet de corriger la détection des transitions vidéo en détectant la présence de tels effets, comme par exemple dans les méthodes proposées par [Heng 99] ou [Truong 01].

2.4.4 Résultats expérimentaux

Pour valider les méthodes proposées, nous avons effectué plusieurs tests sur des séquences spécifiques du domaine de l'animation mais aussi sur des séquences naturelles.

Comparaison entre le parcours séquentiel et le parcours adaptatif

Dans un premier temps nous avons essayé d'améliorer le parcours de la séquence. Le sous-échantillonnage temporel avec un pas d'analyse l a été comparé avec un parcours adaptatif (voir le rapport [Ionescu 03]), méthode inspirée des travaux proposés dans [Lee 01] qui utilise la méthode "divide et impera", un principe connu en techniques de programmation. L'algorithme est présenté ci-dessous (Algorithme 2) et la décision de discontinuité est effectuée en utilisant la décision majoritaire proposée par l'équation 2.16 de la méthode *4histogrammes*.

En ce qui concerne la détection des "cuts", les parcours séquentiel et adaptatif aboutissent pratiquement aux mêmes résultats. Le Tableau 2.3 présente les temps de calcul de la détection des "cuts" en utilisant les deux méthodes de parcours de la séquence[2] : le parcours avec un pas d'analyse fixe en utilisant à chaque fois le principe de la méthode *4histogrammes* et le parcours avec la fenêtre adaptative. Nous avons utilisé la réduction des couleurs dans l'espace Lab (voir la Section 2.4.3). Les valeurs des paramètres sont : $l = l_{min} = 10$ et $l_{max} = 40$ (où l_{min} et l_{max} sont le pas d'analyse minimal et maximal, voir l'Algorithme 2).

Séquence	Nb.images	Nb."cuts"	"cuts"/10s	T_{fixe}	T_{adapt}
The Buddy System	9385 (6min15s)	60	1.6	125.13min	97min
Gazoon	3927 (2min37s)	18	1.2	52min	60min

TAB. 2.3 – Les temps de calcul pour la détection des plans en utilisant le parcours avec un pas fixe (T_{fixe}) et adaptatif (T_{adapt}). Les films proviennent du CICA [CICA 06].

Le parcours séquentiel est plus rapide sur des séquences courtes. Par contre, le parcours adaptatif est plus efficace pour de longues séquences. Nous avons constaté qu'en augmentant la précision de détection, donc en diminuant la valeur de l_{min}, le temps de calcul de l'algorithme adaptatif augmente car plusieurs étapes de division de l'intervalle d'analyse sont requises. De plus, les mouvements répétitifs de la caméra ou les effets de couleurs augmentent également le temps de traitement car ils introduisent dans chaque fenêtre d'analyse des dissimilarités entre les images qui sont alors traitées comme d'éventuels "cuts". Les principales contraintes de ce type de parcours sont les suivantes :

- d'abord il ne peut pas être implanté d'une manière efficace pour la méthode *2dérivées* (qui est plus performante que la méthode *4histogrammes*) car dans ce cas la détection des "cuts" est effectuée après le parcours complet de la séquence. Le parcours adaptatif impose que la détection soit effectuée progressivement pour chaque paires d'images analysées.
- ensuite le calcul automatique du seuil de détection proposé, τ_{cut}, est impossible car la détection est effectuée en même temps que le calcul de la similarité entre les images.

[2]les temps de calculs sont donnés à titre indicatif car les méthodes testées n'ont pas été optimisées. Les tests ont été réalisés sur une machine Sun UltraSparc III à 333MHz en utilisant des images sans sous-échantillonnage spatial, taille moyenne de 720×480 pixels.

Algorithm 2 La détection des "cuts" avec un parcours adaptatif

$l \leftarrow l_{max}$ {on commence avec un pas l de taille maximale l_{max} (valeur élevée)}
$index_d \leftarrow 0$
$index_f \leftarrow index_d + l$
faire
détection_cut($index_d$, $index_f$, l) {on exécute la détection récursive des "cuts" pour les images aux indices $index_d$ et $index_f$}
$l \leftarrow l_{max}$
$index_d \leftarrow index_d + l$
$index_f \leftarrow index_f + l$ {positionnement sur les images suivantes}
tant que ($index_f + l < index_{max}$) {$index_{max}$ est l'indice de la dernière image de la séquence}

procedure détection_cut($index_d$, $index_f$, l)
si [dissim($image_{index_d}$, $image_{index_f}$) et sim($image_{index_f}$, $image_{index_f+l_{min}}$)
et ($l \leq l_{min}$)] **alors**
"cut" détecté à l'instant $index_f$ {un "cut" est détecté si les images aux instants $index_d$ et $index_f$ sont dissimilaires, les images aux instants $index_f$ et $index_f + l_{min}$ sont similaires et le pas d'analyse est inférieur à la valeur minimale, l_{min}}
sinon
détection_cut($index_d$, $index_d + l/2$, $l/2$)
détection_cut($index_d + l/2$, $index_f$, $l/2$) {le pas l est divisé par 2 et l'analyse se fait récursivement une fois entre les images aux instants $index_d$ et $index_d + l/2$ et en deuxième temps entre les images aux instants $index_d + l/2$ et $index_f$}
fin si
fin procedure

L'influence de la réduction des couleurs

Nous avons étudié l'influence du choix de la méthode de réduction des couleurs sur les résultats de la détection des "cuts" (voir [Ionescu 05f]). La méthode de détection des "cuts" *4histogrammes* a été testée sur deux films d'animation de longue durée : "A Bug's Life" 84min46s et 1597 "cuts", et "Toy Story" 73mn18s et 1569 "cuts", ce qui fait une durée totale de 158min et comportant 3166 "cuts". Nous avons utilisé les quatres techniques de réduction couleur proposées dans le Section 2.4.3. Après avoir réalisé une étape de segmentation manuelle pour bénéficier d'une vérité terrain, nous avons pu calculer les taux de détections globaux pour les deux films.

Les résultats sont présentés dans le Tableau 2.4 (le pas d'analyse a été fixé à $l = 2$). Pour mesurer les performances nous avons utilisé les taux de *précision* et de *rappel* définis dans l'équation 2.2 de la Section 2.3.

Réd.couleurs	Espace RVB	Espace TLS	Espace Lab	Diff.d'erreur
Précision	**92.83%**	**92.8%**	**92.94%**	**94.32%**
Rappel	**86.67%**	**92.1%**	**90%**	**88.63%**

TAB. 2.4 – Les taux de détection pour différentes méthodes de réduction des couleurs.

On constate que la méthode de réduction couleur utilisée joue un rôle important sur le nombre de bonnes détections : variation du *rappel* allant de 86.67% à 92.1%. C'est la réduction des couleurs dans l'espace TLS qui fournit globalement les meilleurs résultats, avec des valeurs des deux taux de détection élevées : $précision = 92.8\%$ et $rappel = 92.1\%$. Ce résultat peut s'expliquer par le fait que cette méthode a tendance à préserver, voire accentuer, les différences entre les couleurs. Par contre elle a également pour effet d'augmenter légèrement les fausses détections par rapport aux autres réduction couleurs (voir la *précision* dans le Tableau 2.4).

Cependant les erreurs de fausses détections semblent moins sensibles à la méthode de réduction couleur employée. Les taux de *précision* restent élevés compris entre 92.8% et 94.32%, la diffusion d'erreur donnant le moins de fausses détections, à savoir $précision = 94.32\%$. On notera toutefois que cette méthode demande un temps de calcul bien supérieur aux autres méthodes testées (voir la Section 2.4.3).

Test comparatif

Nous avons proposé un test comparatif entre des différentes méthodes. La méthode *2dérivées* a été comparée premièrement avec la méthode *4histogrammes* et deuxièmement avec la méthode basée sur l'analyse de la discontinuité du mouvement, *mdiscont*, présentée dans la Section 3.2.4 (voir [Ionescu 06a]). Les tests ont été réalisés sur les deux films d'animation "A Bug's Life" et "Toy Story" utilisés dans le test précédent.

Méthode	*Précision*	*Rappel*
4histogrammes	**93.37%**	**88.63%**
mdiscont	**89.39%**	**94.53%**
2dérivées	**94.92%**	**92.6%**
2dérivées$^+$	**95.97%**	**92.6%**

TAB. 2.5 – Les taux de détections obtenus ($^+$ représente l'amélioration de la détection en utilisant la détection des SCC).

En ce qui concerne le réglage des paramètres, pour les méthodes *4histogrammes* et *2dérivées* le pas d'analyse a été fixé à $l = 2$, nous avons utilisé la méthode de réduction des couleurs qui donnait le plus petit taux de fausses détections, à savoir la méthode utilisant la diffusion d'erreur, et pour le sous-échantillonnage spatial nous avons utilisé $n = 4$. Pour la méthode *mdiscont* l'analyse a été réalisée pour chaque image du film en utilisant un seuil de discontinuité de $\tau_{discont} = 10000$. Les résultats obtenus sont synthétisés dans le Tableau 2.5.

La méthode *4histogrammes* a obtenu le taux de rappel le plus faible 88.63% et donc le plus faible taux de bonnes détections. La méthode *mdiscont* a obtenu le meilleur taux de bonnes détections, $Rappel = 94.53\%$, mais avec un taux plus élevé de fausses détections, $Précision = 89.39\%$.

La méthode *2dérivées* a obtenu des taux élevés de précision et rappel, supérieurs à 92%, ce qui correspond à une amélioration du rappel de 4% par rapport à la méthode *4histogrammes* (soit 125 "cuts" détectés en plus) et une amélioration de la précision par rapport à la méthode *mdiscont* de 5.5% (soit 178 "cuts" détectés en plus). Ajoutons également qu'en utilisant la méthode de détection des SCC (voir la Section 2.5) nous améliorons de 1% la précision de

la méthode *2dérivées*, si bien qu'elle passe à 96%.

La plupart des fausses détections surviennent à cause de mouvements très rapides de la caméra (que l'on rencontre très fréquemment dans les deux films testés). Nous pouvons également remarquer que les non-détections sont liés à des similarités entre les couleurs et à la présence de transitions graduelles dans la séquence, comme par exemple avec l'utilisation fréquente de "fades" et de "dissolves". Dans la Figure 2.10 nous présentons un comparatif de détection pour les trois méthodes sur quelques situations difficiles .

FIG. 2.10 – Exemple de détection pour les trois méthodes testées : 4h - *4histogrammes*, 2d - *2dérivées*, md - *mdiscont* (la détection d'un "cut" est marquée par un rectangle bleu et la non-détection par un rectangle rouge).

Dans la Figure 2.10 les situations A et C correspondent à des "cuts" et les situations B et D correspondent à des changements visuels forts qui sont typiquement la source de fausses détections. Les résultats obtenus sont :

- **dans la situation A**, suite à une forte similarité entre les couleurs, seule la méthode *mdiscont* basée sur l'analyse du mouvement a permis de détecter un "cut",
- **dans la situation B** suite au mouvement de la caméra la méthode *4histogrammes* a détecté un "cut" qui n'en était pas un,
- **dans la situation C** suite au mouvement très rapide de la caméra survenu après le "cut", la méthode *mdiscont* a échoué en détectant une discontinuité du mouvement,
- **dans la situation D** le mouvement très rapide d'un personnage a provoqué une fausse détection des méthodes *2dérivées* et *4histogrammes* mais quelques images plus tard.

Vers la fusion couleur-mouvement

Suite aux tests précèdents la méthode *mdiscont* a obtenu le meilleur taux de *rappel* (94.53%), et donc le nombre le plus élevé de bonnes détections. Cette méthode est efficace en ce qui concerne la détection des discontinuités dans la séquence, mais elle est sensible aux mouvements très rapides de caméra, d'où le taux faible de précision (89.39%). Il est donc intéressant d'utiliser une méthode efficace d'analyse du mouvement qui permettrait de détecter également les "cuts" (voir le rapport [Ionescu 05c]).

Pour améliorer la détection de la méthode *2dérivées* nous avons utilisé les plans de discontinuité définis dans la Section 3.2.2, qui sont obtenus par la détection du mouvement. Ces plans correspondent normalement à des transitions vidéo ou des mouvements très rapides de caméra ou des effets couleurs. La fusion des résultats est effectuée en éliminant tous les "cuts" détectés par la méthode *2dérivées* qui se trouvent à l'intérieur d'un plan de discontinuité.

Les erreurs globales obtenues sur les deux séquences utilisées dans les tests précédents, sont : un taux de *précision* de *95.6%* et un taux de *rappel* de *92.2%* par rapport aux taux respectifs de *94.92%* et de *92.6%* sans l'utilisation du mouvement. On note donc une *légère*

amélioration par rapport aux fausses détections mais également une légère perte par rapport aux bonnes détections. Ces tests ne sont pas totalement satisfaisants et il serait nécessaire d'envisager une coopération plus élaborée de ces deux approches.

Campagne ARGOS

Nous avons participé à la campagne "ARGOS" [ARGOS 06] consistant à évaluer des outils d'analyse de contenus vidéo, et en particulier le découpage en plans. La méthode de détection des "cuts" *2dérivées* a été testée sur le corpus SFRS des films documentaires qui est composé de 21 films d'une durée totale de 10 heures et 24 minutes. En ce qui concerne le réglage des paramètres nous avons utilisé : un pas d'analyse $l = 2$, une réduction des couleurs avec la diffusion d'erreur et un sous-échantillonnage spatial avec $n = 4$.

Globalement nous avons obtenu un taux de *précision* de *91%* et un taux de *rappel* de *90.5%* ce qui nous situe au-dessus de la moyenne générale (*précision moyenne* de *85.9%* et *rappel moyen* de *85.4%*) obtenue sur un nombre de 10 participants et plus de 23 méthodes testées.

Les fausses détections sont liées à la qualité des séquences utilisées. La plupart des films datent de plus de 10 ans et ont été numérisés à partir d'un support magnétique (cassettes VHS [ARGOS 06]). En effet nous avons noté un niveau de bruit important dans les images, un mouvement saccadé, et des changements aléatoires de l'intensité lumineuse. Des situations causant des fausses détections sont présentées dans la Figure 2.11.

FIG. 2.11 – Quelques situations qui ont donné un nombre élevé de fausses détections successives.

Dans les situations mentionnées ci-dessus, la méthode *2dérivées*, qui est basée sur l'analyse de la discontinuité visuelle en représentant les images avec une palette de couleurs réduite, est très sensible aux variations répétitives discontinues de couleurs et donnent des séquences de fausses détections pour ces passages.

2.4.5 Conclusions

Dans ce chapitre nous avons proposé une méthode de détection de "cuts" (la méthode *2dérivée*) adaptée aux particularités des films d'animation. Les dissimilarités visuelles introduites par les "cuts" sont traduites par des différences entre histogrammes couleurs et la méthode de détection de ces "cuts" a été améliorée par l'utilisation de la dérivée seconde par rapport au temps de la différence entre histogrammes couleurs des images successives composant la séquence. De plus, le seuil de détection est calculé d'une manière automatique. Cette méthode donne de meilleurs résultats que les deux autres approches envisagées : la méthode *4histogrammes* (approche classique sur les histogrammes) et la méthode *mdiscont* (approche sur la discontinuité du mouvement).

Nous avons également testé notre méthode sur un corpus de films documentaires de la compagne d'évaluation ARGOS [ARGOS 06]. La méthode que nous avons utilisée pour la

réduction des couleurs s'avère moins efficace que prévu à cause de la variation des couleurs due au bruit, aux mouvements discontinus et à la variation de la luminosité. Nous obtenons alors un nombre élevé de fausses détections mais en même temps un taux élevé de bonnes détections (supérieure à 95%).

La présence des autres transitions vidéo, comme par exemple les "fades" ou les "dissolves" augmente le nombre de fausses détections car ils sont vus comme des discontinuités. Par contre, une détection de ces transitions permet de corriger la détection de "cuts" et, dans un deuxième temps, aide au découpage en plans comme nous le verrons dans la suite.

Le choix des paramètres de la méthode est lié à la réussite de la détection. Un premier paramètre important est le pas d'analyse. Un petit pas d'analyse ($l < 3$) donne une meilleure précision de détection. Un pas d'analyse trop élevé a comme conséquence la perte de certains "cuts".

Le deuxième paramètre important est le choix du seuil de détection des "cuts", puisqu'il permet de décider si un "cut" est présent ou non. Théoriquement chaque séquence nécessite un seuil particulier. Un seuil trop bas entraîne une sur-détection et un seuil trop élevé donne une sous-détection. En général, il est préférable d'avoir une sur-segmentation temporelle (donc avoir trop de "cuts") qu'une sous-détection, parce qu'en analysant les plans, il est toujours possible d'agréger plusieurs plans en un seul, selon un critère de similarité couleur, par exemple.

Le troisième paramètre important est le choix de la réduction des couleurs, car il influe sur la quantification des discontinuités visuelles entre les images. Pour la tâche de détection des "cuts" une réduction des couleurs qui entraîne une qualité visuelle moyenne est suffisante. Mais, dans des situations particulières, comme lors de déplacements discontinus de la caméra, ou des variations de l'intensité lumineuse dans l'image, la réduction des couleurs augmente le nombre de fausses détections (voir la Section 2.4.4 sur les résultats expérimentaux).

2.5 La détection de SCC

Comme nous l'avons déjà mentionné, l'une des particularités des films d'animation est la présence d'effets couleurs qui sont utilisés par l'artiste pour augmenter l'intérêt d'une scène de la séquence. Les changements brefs de couleurs ou SCC ("short color change") [Ionescu 05d] sont une des techniques utilisées. Un SCC est un effet de courte durée, typiquement de moins de 20 images, qui se caractérise par *un changement brutal et bref des couleurs d'une image suivi du retour à l'image initiale ou à une image similaire*.

Un SCC n'est pas une transition vidéo mais plutôt un effet visuel. Dans cette catégorie on peut mentionner les flashs, les explosions, les éclairs, etc. Quelques exemples sont présentés dans la Figure 2.12.

L'intérêt de détecter ce type d'effet est double. D'une part un SCC produit une double discontinuité visuelle dans la séquence : une au début de l'effet et une à la fin. Les méthodes de détection des "cuts" basées sur l'analyse de l'image vont détecter ces deux transitions comme des "cuts". Donc, dans un premier temps, la détection des SCC permet d'améliorer les taux de détection et particulièrement de réduire le nombre de fausses détections (voir la Section 2.4.3). D'autre part la présence de nombreux SCC dans une séquence nous donne une information sémantique. Dans certains films d'animation l'action est reliée à l'utilisation de tels effets. Par exemple, un passage contenant un nombre élevé de SCC est un passage qui va particulièrement retenir l'attention du spectateur. De la même manière, un film qui

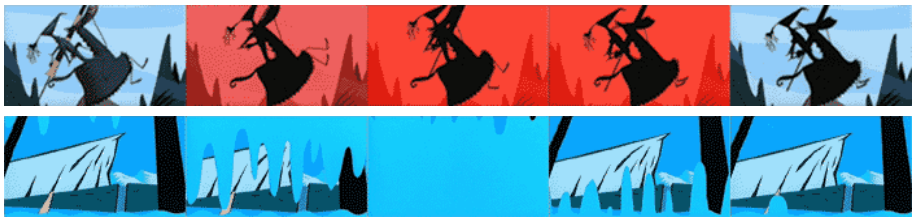


FIG. 2.12 – Quelques changements brefs de couleurs ou SCC (film "François le Vaillant" [Folimage 06b]).

utilise un pourcentage élevé de SCC est un film que nous qualifierons *"d'explosif"*.

Dans les films naturels, une situation similaire correspond aux flashs dans une scène. L'effet d'un flash est une variation soudaine de l'intensité lumineuse et il sera détecté comme un "cut". Les détecteurs existants se sont inspirés du modèle physique de la génération d'un flash d'appareil photo. Pour un état de l'art sur les méthodes de détection de flashs, on pourra se rapporter à [Heng 99] et [Truong 01].

2.5.1 La méthode proposée

La méthode de détection des SCC que nous proposons est inspirée du principe général de la détection d'un flash dans les séquences naturelles (voir [Heng 99]). Le principe est de rechercher des images semblables, temporellement voisines et séparées par des images comportant un changement important de couleurs.

La détection de SCC débute dès qu'un "cut" a été détecté. Les changements de couleurs entre les images sont mis en évidence en utilisant la distance Euclidienne entre les histogrammes couleurs calculés sur l'image entière. Ce type d'effet se manifeste par une variation couleur si importante que l'utilisation d'une réduction couleur par quantification de l'espace RVB, méthode rapide mais moins précise, est suffisante (voir Section 2.4.3). Comme les histogrammes sont calculés sur l'image entière, un sous-échantillonnage spatial avec $n = 4$ (voir dans la Section Chapitre 2.4.3 les pré-traitements) ne produit aucun changement sur la précision de la méthode mais il diminue le temps de traitement. La séquence est aussi sous-échantillonnée temporellement avec un pas constant l qui est un paramètre de la méthode. L'algorithme présenté (voir Algorithme 3) est illustré dans la Figure 2.13.

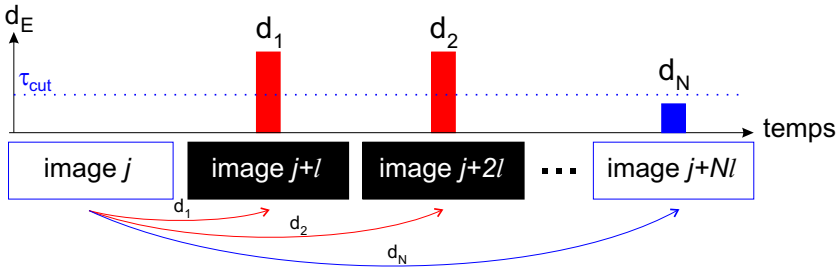


FIG. 2.13 – L'algorithme de détection des SCC.

Pendant l'effet, les différences entre les histogrammes de l'image de début et les images voisines, correspondant aux instants $j + i \cdot l$ avec $i > 2$, sont importantes et donc plus élevées

que le seuil τ_{cut}. A la fin de l'effet, on trouve une image à l'instant $j + N \cdot l$, avec N entier, similaire à l'image de début, $image_k$, et donc pour laquelle la différence $d_E(H_j, H_{j+N \cdot l}) = d_N < \tau_{cut}$.

Algorithm 3 Détection des SCC

si ("cut" détecté à l'instant j) **alors**
  {j est un multiple entier du pas d'analyse l}
  $i \leftarrow 1$ $H_j \leftarrow Histogramme(\widehat{image}_j)$
  $H_{j+i \cdot l} \leftarrow Histogramme(\widehat{image}_{j+i \cdot l})$ {le chapeau indique la version sous-échantillonnée spatialement et avec la réduction des couleurs dans l'espace RVB, de l'image}
  $d_i \leftarrow d_E(H_j, H_{j+i \cdot l})$ {$d_E()$ est la distance Euclidienne entre les 2 histogrammes couleur globaux}
  si ($d_i > \tau_{cut}$) **alors**
    {τ_{cut} est le même seuil que celui qui est utilisé pour la détection des "cuts"}
    faire
      $i \leftarrow i + 1$
      $H_{j+i \cdot l} \leftarrow Histogramme(\widehat{image}_{j+i \cdot l})$
      $d_i \leftarrow d_E(H_j, H_{j+i \cdot l})$
    tant que [($d_i > \tau_{cut}$) ou ($i \cdot l < t_{SCC}$)]
    {t_{SCC} est la taille maximale d'un SCC, valeur fixée a priori à 20 images de manière empirique en observant un grande nombre de SCC}
    si ($d_i < \tau_{cut}$) **alors**
      SCC détecté entre l'instant j et $j + i \cdot l$
    fin si
  fin si
fin si

2.5.2 Résultats expérimentaux

Pour valider cette méthode de détection des SCC nous l'avons testé sur un ensemble de 14 films d'animation de [CICA 06] d'une durée totale de 102 minutes et contenant 120 SCC. Pour chaque séquence les SCC ont été marqués à la main, cela nous servant de vérité terrain.

De plus, pour avoir une idée sur la difficulté de l'analyse de chaque film, nous avons labellisé les 14 séquences utilisées selon leurs particularités :

- une séquence dite *"facile"* comporte peu de mouvement et une délimitation aisée des "cuts",
- une séquence dite *"moyenne"* comporte du mouvement et des transitions vidéo,
- une séquence dite *"difficile"* est une séquence avec de nombreux mouvements de caméra ou d'objets et comportant de nombreux effets de couleurs et de nombreuses transitions.

Les films utilisés sont présentés dans le Tableau 2.6.

Pour la détection des SCC nous avons utilisé un pas d'analyse de $l = 2$, une réduction des couleurs dans l'espace RVB, le même seuil de similarité que celui utilisé pour la détection des "cuts", à savoir $\tau_{cut} = 0.2$ et une fenêtre de recherche de 20 images. Les résultats obtenus sont résumés dans le Tableau 2.7.

index	nom	vignette	nb. images	durée	appréciation
1	A Crushed World		10059	6m42s	moyenne
2	A Viagem		11312	7m32s	difficile
3	François le Vaillant		13416	8m56s	moyenne
4	Paradise		21083	14m03s	difficile
5	Gazoon		4195	2m47s	facile
6	The Buddy System		9566	6m22s	difficile
7	Le Moine et le Poisson		8977	5m59s	moyenne
8	Casa		9128	6m05s	moyenne
9	Circuit Marine		8378	5m35s	difficile
10	Ferrailles		9392	6m15s	moyenne
11	The Hill Farm		24993	16m39s	difficile
12	L'Egoïste		4727	3m09s	moyenne
13	Le Chat d'Appartement		10073	6m42s	moyenne
14	Le Château des Autres		7460	4m58s	difficile

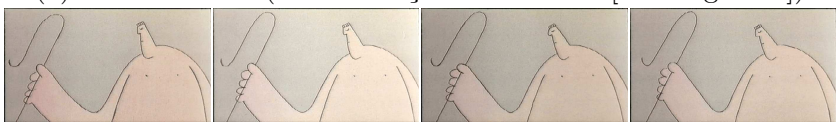
TAB. 2.6 – Les 14 films d'animation utilisés pour la détection des SCC et des "fades".

film	1	2	3	4	5	6	7	8	9	10	11	12	13	14
N_t	1	0	39	7	0	7	0	0	4	2	45	3	12	0
BD	1	0	38	6	0	5	0	0	4	1	40	2	9	0
FD	0	0	2	0	0	0	1	0	0	0	0	0	5	0

TAB. 2.7 – Les résultats de la détection des SCC pour les 14 films testés (N_t est le nombre total de SCC, BD et FD représentent respectivement le nombre de bonnes détections et de fausses détections).

Globalement nous avons obtenu un taux de *précision* de *93%* et un taux de *rappel* de *88.3%*. Les situations pour lesquelles nous obtenons de fausses détections correspondent à des changements de plans dans lesquels les distributions des couleurs dans ces plans sont voisines, ou à une variation forte de l'intensité lumineuse dans les images d'un même plan, ce qui aboutit à des images très differentes après l'utilisation de la réduction des couleurs (voir Figure 2.14.b). Les non-détections sont plutôt liées à la dissimilarité entre l'image de début et de fin de l'effet, images qui sont suffisamment différentes pour que la distance Euclidienne entre leurs histogrammes soit significative (voir la Figure 2.14.a).

(a) non-détection (film "François le Vailant" [Folimage 06b])



(b) fausse détection (film "The Hill Farm" [CICA 06]).

FIG. 2.14 – Exemples d'erreurs de détection des SCC.

2.5.3 Conclusions

Une des particularités des films d'animation est la présence d'effets visuels. C'est pourquoi nous avons proposé une méthode de détection d'un de ces effets particulièrement utilisé, appelé "changement bref de couleurs" ou SCC , qui peut correspondre à une explosion, un éclair, un flash, etc. La méthode proposée a été inspirée du principe de détection de flashs de la caméra dans les films naturels (voir [Heng 99]) qui, adaptée aux films d'animation, a donné de bons résultats.

L'intérêt de la détection des SCC est double :

- **corriger la détection des "cuts"** : premièrement il permet de corriger la détection des "cuts", car un SCC est souvent détecté comme plusieurs "cuts",
- **analyser le contenu** : deuxièmement l'utilisation d'un tel effet donne une signification particulière au contenu de la séquence. Les films d'animation comportant un nombre élevé de SCC attirent particulièrement l'attention du spectateur.

2.6 La détection des "fades"

Un "fade" est une transition graduelle, très souvent utilisée, caractérisée par un effet optique qui permet, en partant d'un fond constant, le plus souvent noir, de faire apparaître progressivement une image. C'est le "fade-in". Le processus inverse, qui consiste à passer progressivement d'une image à un fond constant, est le "fade-out". Souvent les deux sont utilisés l'un après l'autre, dans l'ordre "fade-out" - "fade-in". Dans ce cas les deux constituent une seule transition qui est appelée un "fade group". Un exemple de transition de type "fade" est présenté dans la Figure 2.2.

La séquence d'images constituant un "fade", noté $F(x, y, t)$ de durée T, est définie comme la transformation de l'intensité des pixels d'une séquence $S_1(x, y, t)$ par une fonction monotone $f(t)$ [Lienhart 01b] :

$$F(x, y, t) = f(t) \cdot S_1(x, y, t), \quad 0 \leq t \leq T \tag{2.23}$$

Un "fade-in" utilise comme fonction monotone la fonction définie par $f(0) = 0$ et $f(T) = 1$ et de la même façon un "fade-out" est défini par la fonction $f(0) = 1$ et $f(T) = 0$. La plupart du temps, la fonction $f(t)$ est choisie linéaire et est définie de la façon suivante :

$$f_{fade\ in}(t) = \frac{t}{T} \quad f_{fade\ out}(t) = 1 - \frac{t}{T} \tag{2.24}$$

Dans la suite nous allons présenter les différentes méthodes de détection de "fades" trouvées dans la littérature.

2.6.1 État de l'art

Par rapport à la détection des transitions abruptes ou "cuts", la détection des "fades" a été nettement moins abordée dans la littérature. Les méthodes existantes sont orientées vers deux axes prédominants :

- les approches basées sur l'*intensité des pixels*,
- les approches basées sur *l'analyse des contours*.

Différents états de l'art sur la détection des "fades" ont été proposés dans [Lienhart 01b], [Hanjalic 02] ou [Ren 03].

L'analyse de l'intensité des pixels

Une des premières approches a été proposée dans [Zhang 93] où les transitions graduelles sont détectées en utilisant deux seuils. La méthode est basée sur l'analyse des distances entre histogrammes, méthode utilisée dans un premier temps pour la détection des "cuts". Un "cut" est détecté si la dissimilarité entre deux images successives est supérieure au seuil $\tau_{début}$. Si à un instant k, la dissimilarité entre deux images successives passe au dessus d'un deuxième seuil, τ_{cand} ($\tau_{cand} < \tau_{début}$), tout en restant en dessous du seuil $\tau_{début}$, alors l'image k est une image candidate potentielle de début d'une transition graduelle. Ensuite, cette image est comparée avec les images suivantes, et les écarts sont cumulés. Quand le cumul des écarts dépasse le seuil $\tau_{début}$ et si en même temps les différences entre les images consécutives restent inférieures à τ_{cand}, la détection d'un "fade" est validée.

Pendant un "fade" une des informations caractéristiques est le changement des intensités lumineuses des pixels. D'autres approches étudient l'évolution de l'écart-type de l'intensité lumineuse des pixels calculé dans l'image entière. En supposant l'hypothèse d'ergodicité vérifiée, l'écart-type lors d'un "fade" peut être exprimé par :

$$\sigma(F(x,y,t)) = f(t) \cdot \sigma(S_1(x,y)) \tag{2.25}$$

où $F(.)$, $f(.)$ et $S_1(.)$ ont les mêmes significations que dans la définition du "fade" dans l'équation 2.23 et $\sigma()$ représente l'écart type.

La méthode présentée dans [Lienhart 99a] propose dans un premier temps de localiser dans la séquence les images monochromatiques pour lesquelles la variance de l'intensité lumineuse est proche de zéro. Ces images sont des candidates potentielles pour être des images de début (respectivement de fin) d'un "fade in" (respectivement d'un "fade-out"). Les "fade-in/out" sont détectés en analysant la croissance de l'intensité lumineuse des pixels et de l'écart-type, dans le sens positif et négatif du défilement du temps. La linéarité est détectée par l'évaluation de l'erreur entre ces évolutions et les droites de régressions les modélisant.

Une approche similaire basée sur l'analyse de la variance est proposée dans [Alattar 97]. Les "fades" sont d'abord détectés par l'analyse de tous les extrema négatifs de la dérivée seconde des valeurs de la variance des intensités lumineuses des pixels. Puis, ils sont confirmés si la dérivée première de l'intensité lumineuse moyenne est constante entre deux extrema négatifs.

Une autre approche, proposée dans [Truong 00a], combine les deux méthodes proposées dans [Lienhart 99a] et [Alattar 97]. Dans un premier temps les images monochromatiques sont détectées. Seules les images qui sont proches d'un extremum négatif de la dérivée seconde de la variance de l'intensité lumineuse des pixels sont retenues. Pendant une transition de type "fade" on rencontre les conditions suivantes :

- la dérivée première de la courbe lissée de l'intensité moyenne reste constante et ne change pas de signe,
- la valeur moyenne de la pente de la dérivée première doit être supérieure à un certain seuil,
- la valeur de la variance de l'intensité lumineuse de la première et dernière image du "fade" doit également être supérieure à un certain seuil.

Une des principales sources d'erreurs dans la détection des "fades" est la présence de mouvement dans la séquence qui rend la variation de l'intensité lumineuse dans l'image non linéaire. Dans [W.A.C.Fernando 99] la détection de "fades" est effectuée en utilisant des mesures statistiques sur l'intensité des pixels mais aussi sur le signal de chrominance, pour réduire l'influence du mouvement. L'analyse est effectuée en utilisant l'espace YCbCr (Y-luminance et Cb, Cr-différences chromatiques). La moyenne du signal de chrominance C ($C = \frac{C_b+C_r}{2}$) est moins sensible à la présence du mouvement que la moyenne de la luminance, Y. Ces deux informations sont utilisées pour définir le paramètre $R(k)$ qui est le rapport du changement incrémental de la moyenne du signal de luminance relative au signal de chrominance :

$$R(k) = \begin{cases} \frac{\Delta_k^Y}{\Delta_k^C} & \text{si } k < L_1 \text{ ou } k \geq (L_1 + T) \\ \frac{|C_0 - m_{k+1}^Y + (L_1 - k)\cdot\Delta_k^Y|}{|C_0 - m_{k+1}^C + (L_1 - k)\cdot\Delta_k^C|} & \text{si } L_1 \leq k < (L_1 + T) \end{cases} \tag{2.26}$$

où Δ_k^Y et Δ_k^C sont les changements incrémentaux de la moyenne du signal de luminance Y et du signal de chrominance C, m_{k+1}^Y et m_{k+1}^C sont les valeurs moyennes de Y et C à l'instant $k+1$, L_1 est l'instant de début du "fade", T est la durée du "fade", et C_0 est le niveau du signal vidéo au début du "fade". L'analyse des valeurs de $R(k)$ est utilisée ensuite pour la détection d'un "fade" : pendant un "fade", le rapport $R(k)$ doit rester approximativement constant et donc la dérivée, exprimée par $|R(k) - R(k-1)|$ reste proche de 0.

L'analyse des contours

Un autre axe d'étude est l'analyse basée sur les contours. Les méthodes existantes utilisent le fait que pendant une transition de type "fade-out" les contours des objets disparaissent graduellement. De même, pendant une transition de type "fade-in" les contours apparaissent graduellement. Une mesure de la quantité de changements des contours, mesurée par un rapport noté ECR, est proposée dans [Zabih 95]. Cette mesure est également utilisée lors de la détection des "cuts" (voir la Section 2.4.1). Pendant un "fade in" le nombre de pixels contours qui apparaissent, ECR_{in}, est prédominant par rapport au nombre de pixels contours qui disparaissent, ECR_{out} ($ECR_{in} > ECR_{out}$). Pour un "fade-out" on rencontre la situation inverse ($ECR_{in} < ECR_{out}$).

Si les "cuts" se retrouvent comme des valeurs de maxima dans la suite temporelle des valeurs ECR, les "fades" et les autres transitions graduelles sont caractérisés par des intervalles de valeurs élevées de ECR. Cette méthode a été utilisée dans [Zabih 95] et [Zabih 99] pour

la segmentation en plans. Deux approches similaires ont été proposées dans [Lupatini 98] et [Yu 97].

Les approches basées sur l'analyse des contours sont moins efficaces que les approches basées sur l'analyse de l'intensité lumineuse des pixels [Lienhart 01b], car elles sont très sensibles à la présence du mouvement et à des effets de couleurs qui provoquent des changements de contours importants.

Autres approches

Dans cette catégorie on peut d'abord mentionner des approches voisines comme celles qui utilisent différentes mesures statistiques de l'intensité des pixels, mais qui font l'analyse dans le domaine compressé des coefficients DCT (par exemple dans [Bimbo 99]). Un certain nombre d'autres approches ont tenté d'exploiter de nouvelles sources d'information pour contourner les points faibles des méthodes classiques. On peut citer :

- l'utilisation de l'information de mouvement pour réduire son influence et donc le nombre de fausses détections [Porter 01],
- le "rythme visuel" des histogrammes, utilisé pour réduire l'influence du bruit présent dans les images [Guimaraes 03],
- l'exploitation des paramètres calculés dans le domaine fréquentiel comme les coefficients issus de la FFT de l'image [Miene 01]
- des approches statistiques génériques sur la détection des transitions [Heng 01].

Généralement, chaque méthode proposée a un certain nombre d'avantages et de d'inconvénients, liés à certaines situations permettant d'obtenir de meilleurs résultats par rapport aux approches classiques, mais échouant dans d'autres situations. Par exemple l'approche proposée dans [Guimaraes 03] détecte les "fades" avec une précision de localisation d'une image, alors que dans les approches classiques cette localisation est faite de manière grossière. Mais cette méthode a échoué lors de la détection des "fades" qui présentent un faible écart de luminosité entre l'image de début et de fin.

2.6.2 La méthode proposée

La méthode de détection de "fades" proposée est inspirée des travaux présentés dans [W.A.C.Fernando 99] et utilise l'hypothèse que lors d'un "fade in" l'intensité lumineuse augmente progressivement jusqu'à la fin de la transition où elle devient approximativement constante. Pour être moins sensible à la présence du bruit et du mouvement qui produisent de petites variations de l'intensité lumineuse des pixels, [W.A.C.Fernando 99] a proposé d'analyser en même temps l'information de chrominance, car elle est moins sensible aux changements de l'intensité lumineuse. La détection des "fades" est alors effectuée en analysant l'évolution temporelle d'un certain nombre de paramètres statistiques liés à la luminance et à la chrominance.

Cependant cette méthode présente quelques inconvénients pour notre application. En particulier elle se base sur des hypothèses de constance de mesures statistiques mêlant la luminosité et la chrominance qui ne sont pas toujours vérifiées dans le cas de nos séquences (voir films d'animation). En repartant du principe de l'analyse conjointe de l'évolution de la luminance et de la chrominance, nous avons donc proposé une nouvelle mise en oeuvre de cette approche.

Pré-traitements

Avant la détection, un certain nombre de pré-traitements sont effectués. Pour avoir une information continue sur la variation des paramètres utilisés la séquence n'est pas sous échantillonnée temporellement, donc toutes les images de la séquence sont analysées. Par contre, comme pour la détection des autres transitions (voir les sections précédentes) les images sont sous-échantillonnées spatialement avec $n = 4$ (pour chaque bloc de $n \times n = 4 \times 4$ pixels, sans recouvrement, nous retenons seulement un seul pixel). La réduction de la taille de l'image n'agit pratiquement pas sur la détection car les mesures utilisées sont des statistiques calculées sur l'ensemble de l'image, comme la moyenne ou la variance.

Les paramètres proposés sont calculés dans l'espace couleur YCbCr, espace très utilisé en télévision. Cela nous permet de séparer l'information de luminance (la composante Y), de l'information de chrominance (les deux différences chromatiques : Cb et Cr). Selon l'utilisation que l'on veut en faire (HDTV - télévision haute définition ou SDTV - télévision standard) et la quantification des composantes RVB, il existe plusieurs types de transformations YCbCr [Bruns 00].

Dans notre situation, nous avons retenu la transformation suivante, adaptée à la représentation des images sur ordinateur :

$$\begin{cases} Y_{601} = 0.257R + 0.504V + 0.098B + 16 \\ Cb = -0.148R - 0.291V + 0.439B + 128 \\ Cr = 0.439R - 0.368V - 0.071B + 128 \end{cases} \tag{2.27}$$

Un exemple de décomposition d'une image RVB dans l'espace YCbCr est présenté par la Figure 2.15. On peut remarquer que la composante Y est liée aux variations de l'intensité lumineuse dans l'image.

FIG. 2.15 – Exemple de représentation dans l'espace YCbCr. Les valeurs ont été normalisées entre 0 et 255 pour la visualisation (image du film "Le Roman de Mon Âme" [CICA 06]).

Nous avons essayé d'autres espaces couleur qui permettent d'isoler l'information de luminance de l'information de chrominance, comme par les espaces TSL et Lab, où L est la composante de l'intensité lumineuse. Nous avons constaté que la composante Y de l'espace YCbCr est moins sensible aux petites variations de luminance que la composante L.

L'algorithme de détection

La détection d'un "fade" est effectuée par l'analyse de l'évolution temporelle des paramètres suivants (voir [Ionescu 05d]) :

- la valeur moyenne de la composante de l'intensité lumineuse, $\overline{Y}_k$, calculée dans l'image à l'instant k,
- la variance de Y, $\sigma^2(Y_k)$, calculée dans l'image à l'instant k,
- la valeur absolue de la différence entre les valeurs moyennes des composantes Cb et Cr, $|\overline{Cb}_k - \overline{Cr}_k|$, calculée dans l'image à l'instant k.

La variance $\sigma^2(Y_k)$ est utilisée pour détecter le début d'un "fade-in" et également la fin d'un "fade-out", car pour une image constante la variance de Y est nulle.

En étudiant le comportement de deux autres paramètres, $\overline{Y}_k$ et $|\overline{Cb}_k - \overline{Cr}_k|$, pour différentes situations de "fades" dans un certain nombre de films d'animation, nous avons trouvé que leurs évolutions sont typiquement croissantes pour un "fade-in" et décroissantes pour un "fade-out". Le paramètre $|\overline{Cb}_k - \overline{Cr}_k|$ est moins sensible à la présence du mouvement que la moyenne $\overline{Y}_k$ (voir la Figure 2.16). Par contre, en absence de mouvement, le paramètre $\overline{Y}_k$ est souvent plus discriminant que le paramètre $|\overline{Cb}_k - \overline{Cr}_k|$. C'est donc l'analyse conjointe de ces deux paramètres qui contrôle la détection.

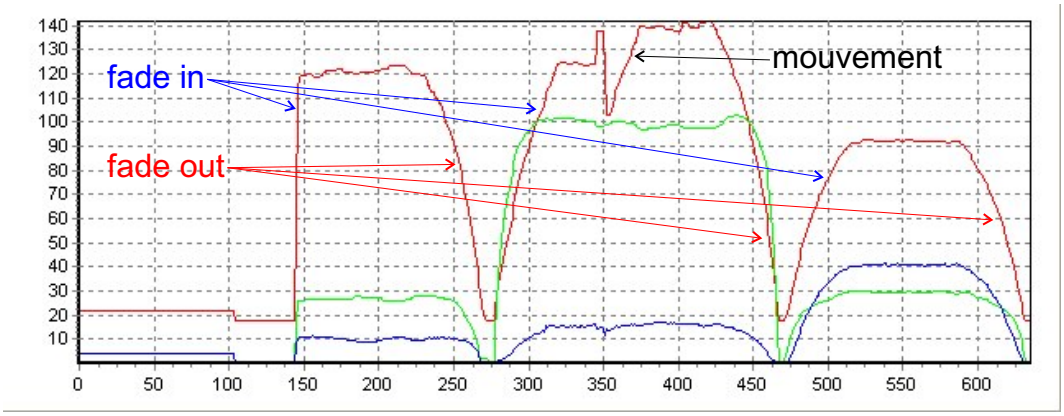


FIG. 2.16 – L'évolution temporelle des paramètres proposés : $\sigma^2(Y)$ normalisé - ligne bleu, $\overline{Y}$ - ligne rouge, $|\overline{Cb} - \overline{Cr}|$ - ligne verte, l'axe oX est l'axe temporel (extrait du film "Coeur de Secours" [CICA 06]).

Dans ce qui suit nous allons nous focaliser sur la détection d'un "fade-in". Une transition "fade-out" correspond à l'inverse d'un "fade-in" et pour la détection il suffit de changer le sens de parcours de la séquence, et donc d'utiliser les mêmes règles que pour la détection d'un "fade-in" mais appliquées dans le sens inverse de défilement du temps (voir la Figure 2.16).

Dans un premier temps nous calculons pour toute la séquence les valeurs des trois paramètres proposés : $\sigma^2(Y_k)$, $\overline{Y}_k$ et $|\overline{Cb}_k - \overline{Cr}_k|$, pour $k = 0, ..., N$ où N est le nombre total d'images de la séquence (sans sous-échantillonnage temporel). Pour réduire l'influence du bruit et lisser les transitions nous avons appliqué un filtrage médian avec une fenêtre de taille 5 sur les évolutions de $\overline{Y}_k$ et $|\overline{Cb}_k - \overline{Cr}_k|$. Le choix du filtre médian est motivé par le fait que ce type de filtrage ne détruit pas les transitions graduelles dans l'évolution temporelle des valeurs des paramètres analysés, comme le ferait un filtre moyenneur.

La détection d'un début de "fade-in" est réalisée si on réunit les conditions suivantes :

- $\sigma^2(Y_k) < \tau_{fade}$, qui traduit que l'image à l'instant k est une image constante. La valeur

du seuil a été fixée empiriquement à $\tau_{fade} = 5$, après l'analyse d'un grand nombre d'images.
- $\sigma^2(Y_{k+1}) > \tau_{fade}$, l'image à l'instant $k+1$ n'est pas constante, condition qui permet de positionner le début du "fade in" juste avant le moment où l'image de fin du "fade-in" commence à réapparaître.
- $D_{k+2,k+1} > 0.2 \cdot D_{k+1,k}$ où $D_{k+2,k+1} = \overline{Y}_{k+2} - \overline{Y}_{k+1}$ et $D_{k+1,k} = \overline{Y}_{k+1} - \overline{Y}_{k}$. Cette dernière condition évite la confusion avec un "cut" qui se traduit par une valeur de $D_{k+2,k+1}$ faible (les images aux instants $k+2$ et $k+1$ sont similaires et appartiennent au même plan).

Si les conditions énumérées ci-dessus sont satisfaites pour l'image à l'instant k, on poursuit la validation du "fade-in" en commençant avec l'image voisine $k+1$. Puis, pour chaque image analysée aux instants $k+i$, avec $i > 1$, on vérifie que les valeurs de $\overline{Y}_{k+i}$ et $C_{k+i} = |\overline{Cb}_{k+i} - \overline{Cr}_{k+i}|$ sont bien croissantes en utilisant la condition suivante :

$$\overline{Y}_{k+i} > 0.98 \cdot \overline{Y}_{k+i-1} \quad \textbf{ou} \quad C_{k+i} > 0.98 \cdot C_{k+i-1} \tag{2.28}$$

Pour la comparaison des valeurs nous avons utilisé, à cause du bruit, une tolérance de 2% choisie empiriquement.

Si pour une certaine image à l'instant $k+i$ la condition n'est pas satisfaite où si i est supérieur à la taille maximale d'une transition de type "fade", c'est à dire $i > t_{max}$, avec $t_{max} = 20$, la détection s'arrête. Enfin, avant de valider la présence d'un "fade-in" nous vérifions que la taille de la transition détectée, c'est à dire i, est supérieure à la taille minimale d'un "fade-in", donc $i > t_{min}$, avec $t_{min} = 3$. Cette dernière condition permet d'éviter la situation où un "cut" est suivi par quelques images présentant des variations d'intensité moyenne semblables à ce que l'on a dans un "fade".

2.6.3 Résultats expérimentaux

Pour valider la méthode proposée nous l'avons testée sur les 14 films d'animation du [CICA 06], utilisés aussi pour la détection des SCC (voir Tableau 2.6), d'une durée totale de 102 minutes et contenant un nombre de 37 "fade-in" et 56 "fade out". Les transitions "fades" ont été détectées dans un premier temps manuellement pour nous servir de référence pour l'évaluation des résultats obtenus. Pour la détection nous avons utilisé un seuil de variance $\tau_{fade} = 5$ et des tailles minimale et maximale d'un "fade" $t_{min} = 3$ et $t_{max} = 20$. Les résultats obtenus sont présentés dans le Tableau 2.8.

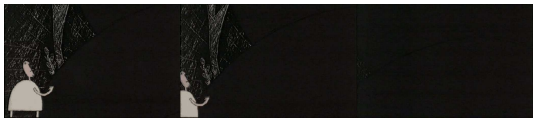
film	1	2	3	4	5	6	7	8	9	10	11	12	13	14
N_{fi}	7	4	2	7	0	1	5	1	0	0	5	0	3	2
BD_{fi}	7	4	2	7	0	1	5	0	0	0	5	0	3	2
FD_{fi}	0	1	1	1	0	0	0	1	0	1	3	0	0	0
N_{fo}	6	4	4	7	0	1	5	3	0	7	6	5	5	3
BD_{fo}	6	4	4	7	0	0	5	3	0	6	6	5	5	2
FD_{fo}	1	1	0	0	0	0	0	0	0	0	3	0	0	0

TAB. 2.8 – Les résultats de la détection de "fades" pour les 14 films testés (N est le nombre de transitions, BD et FD sont les nombres de bonnes détections et de fausses détections (notation : fi="fade-in" et fo="fade-out").

Globalement nous avons obtenu pour la détection des "fade-in" une *précision* de *81.9%* et un *rappel* de *97.2%*, et pour la détection des "fade-out" une *précision* de *91.4%* et un *rappel* de *94.6%*.

Les fausses détections obtenues sont liées à la difficulté particulière des films utilisés. Comme exemple de situations pour lesquelles les fausses détections sont survenues nous pouvons mentionner l'apparition et le déplacement progressif d'objets à l'intérieur d'une image constante, ce qui provoque une variation de l'intensité lumineuse, ou des "cuts" survenus entre une image constante et des images comportant un fort mouvement de caméra par rapport aux objets, ce qui entraîne une variation de l'intensité lumineuse. Des exemples de fausses détections sont présentés dans la Figure 2.17.

"A Viagem"



"The Hill Farm"

"Paradise"

FIG. 2.17 – Exemples de fausses détections de "fades" (films de [CICA 06]).

Les non détections sont liées à des "fades" très courts pour lesquels l'intensité lumineuse augmente jusqu'à 90% en deux images et jusqu'à 100% pendant plusieurs images, ce qui provoque une variation importante, difficile à détecter.

2.6.4 Conclusions

Dans cette section nous avons proposé une méthode de détection de "fades" basée sur l'analyse de mesures statistiques de l'intensité lumineuse et du signal de chrominance calculées dans l'espace YCbCr. L'information sur le signal de chrominance a été utilisée pour réduire l'influence du mouvement sur la qualité de la détection, car elle est moins sensible que l'intensité lumineuse Y. Testée sur un certain nombre de films d'animation la méthode proposée a donné de bons résultats.

Les transitions de type "fade-in" et "fade-out" sont souvent liées les unes aux autres. Nous pouvons considérer un "fade-out" comme un "fade-in" qui se produit en parcourant l'échelle du temps en sens inverse. De plus, ces deux effets peuvent être utilisés ensemble : un "fade out" peut être, après un certain nombre d'images noires, suivi par un "fade-in". Ces deux transitions sont souvent utilisées dans les génériques ou dans différents passages pour lesquels le texte peut apparaître et disparaître. Elles sont également utilisées comme lien entre deux scènes ou entre différents moments où l'action est élevée. D'une manière générale, l'utilisation de passages avec un certain nombre d'images noires entre un "fade-out" et un "fade-in" permet d'obtenir un moment de repos dans le déroulement de l'action de la séquence.

Pour la détection des "fades" l'utilisation d'un sous-échantillonnage temporel n'est pas efficace. En utilisant un pas d'analyse plus élevé que $l = 1$ les variations des paramètres

utilisés deviennent discontinues. Comme la taille maximale d'une transition "fade" est d'environ $t_{max} = 20$ images, en augmentant le pas d'analyse les paramètres n'ont pas le temps de varier suffisamment ce qui nuit à une bonne détection.

Une des principales sources d'erreurs dans la détection des "fades" est le mouvement d'objets qui produit une variation lumineuse importante et le mouvement de caméra qui apporte en plus une variation des couleurs. Dans ces situations l'évolution temporelle des paramètres analysés est similaire avec celle d'un "fade".

Concernant le temps de calcul, la détection se déroulent à une cadence d'environ 20 images par seconde. Les tests ont été effectués sur une machine Pentium IV à 3.4GHz et 1Go de RAM sur des images de 187×103 pixels. Ces bonnes performances proviennent d'une complexité de calcul réduite (l'image est balayée une seule fois pour le calcul des paramètres). En optimisant les procédures de calcul une implantation en temps réel est envisageable.

La méthode proposée a besoin des valeurs d'un certain nombre de paramètres (seuil temporel, seuil d'intensité, etc.). Ces valeurs sont choisies empiriquement. Cependant leur choix n'est pas critique, c'est à dire que des variations de ces valeurs ne dégradent pas sensiblement les performances des méthodes. De plus, les valeurs choisies présentent une certaine généricité car elles conviennent pour l'ensemble des séquences testées.

2.7 Agrégation en plans vidéo

Une séquence d'images est définie comme un ensemble de plans qui sont liés les uns aux autres à l'aide de transitions vidéo. La détection des plans, donc de l'unité de base de la séquence, est la première étape d'analyse pour toutes les méthodes d'analyse des séquences d'images.

Dans ce chapitre nous avons présenté plusieurs méthodes de détection de transitions vidéo et d'effets de couleurs comme : les "cuts", les "SCC" et les "fades". Pour la détection des "dissolves" nous avons utilisé la méthode de [Lienhart 01b]. La détection de ces transitions n'est pas suffisante pour déterminer les plans, et une *étape d'agrégation* est nécessaire. En analysant les résultats de la détection pour les transitions citées, nous avons constaté que ce sont les "cuts" qui sont les plus sensibles aux fausses détections car toutes les autres transitions graduelles sont au moins détectées comme étant un "cut". De même, certains changements de couleurs, comme par exemple les SCC, qui ne sont pas des changements de plans, sont détectés comme des "cuts". Donc, pour délimiter correctement les plans nous avons besoin de définir un certain nombre de règles d'agrégation.

Après avoir mis toutes les transitions détectées dans l'ordre chronologique de leur apparition, les plans vidéo sont définis comme les intervalles continus compris entres deux transitions successives, résultats de l'application des règles suivantes (voir [Ionescu 05d]) :

- tous les "cuts" détectés pendant un intervalle de temps correspondant à une transition graduelle sont supprimés,
- un effet visuel de type SCC ne doit pas être la source d'un changement de plan,
- les plans vidéo d'une durée inférieure à T_{plan} (fixé empiriquement à 5 images) sont supprimés puisque ce sont des passages trop courts et difficilement décelables dans la séquence,
- les plans contenus entre deux transitions "fade-out" et "fade-in", ne contenant que des images noires, sont effacés si leur durée est inférieure à T_{plan}. Sinon cette insertion

d'images noires est alors faite volontairement par l'auteur et possède une signification sémantique comme par exemple une période de repos ou de tranquillité.

Les plans sont représentés en gardant deux informations : l'intervalle des cadres et le type des transitions de début et de fin. Par exemple le $plan_n$ est compris entre les images $[100, 340]$ et encadré par les transitions $[\text{"}cut\text{"}, \text{"}fade - out\text{"}]$. Le principe d'agrégation est présenté par la Figure 2.18.

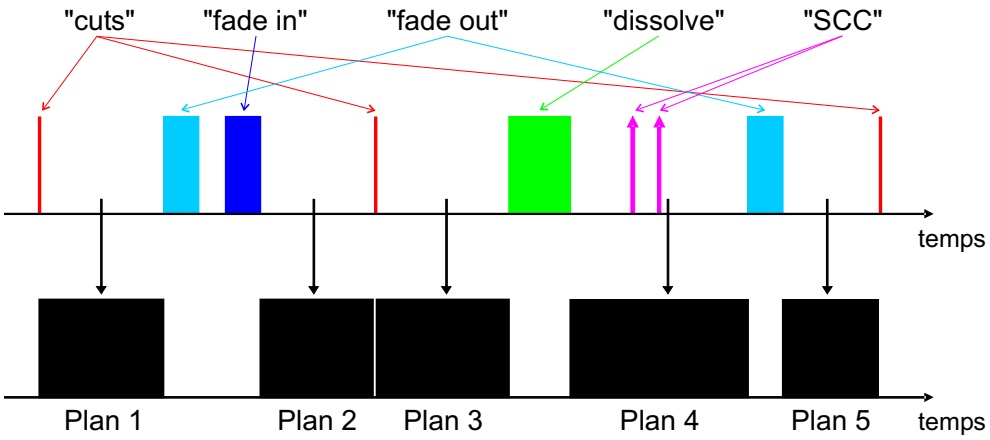


FIG. 2.18 – Agrégation en plans vidéo.

2.8 Annotation visuelle des transitions

Généralement les méthodes d'annotation visuelle sont utilisées dans les systèmes de navigation comme outils permettant à l'utilisateur d'accéder facilement au contenu de la séquence. Les méthodes existantes utilisent différentes techniques de représentation graphique pour visualiser aussi bien l'information temporelle que l'information spatiale sous la forme de certaines images clés de la séquence. Un état de l'art sur les outils de navigation a été présenté dans la Section 1.2.3.

Notre approche est différente (voir [Ionescu 05g]). En utilisant la distribution temporelle des transitions et des plans nous proposons une représentation visuelle de la structure de la séquence sous la forme d'un graphe 2D. Le graphe proposé décrit l'évolution temporelle de la séquence comme un signal continu d'amplitude 1 (arbitraire) qui est interrompu par la présence de certaines transitions (voir la Figure 2.19).

Dans ce graphe chaque transition est représentée par un signal de forme particulière :

- **"cut"** : un *"cut"* correspond au passage par 0 du signal,
- **SCC** : un SCC est un pic du signal,
- **"fade-in"** : un *"fade-in"* est représenté comme un "cut" suivi par une pente croissante du signal,
- **"fade-out"** : un *"fade-out"* est l'inverse d'un "fade-in",
- **"dissolve"** : un *"dissolve"* est représenté comme une pente décroissante suivie par une pente croissante du signal.

La durée de chaque symbole correspond à la durée réelle de la transition. Les formes

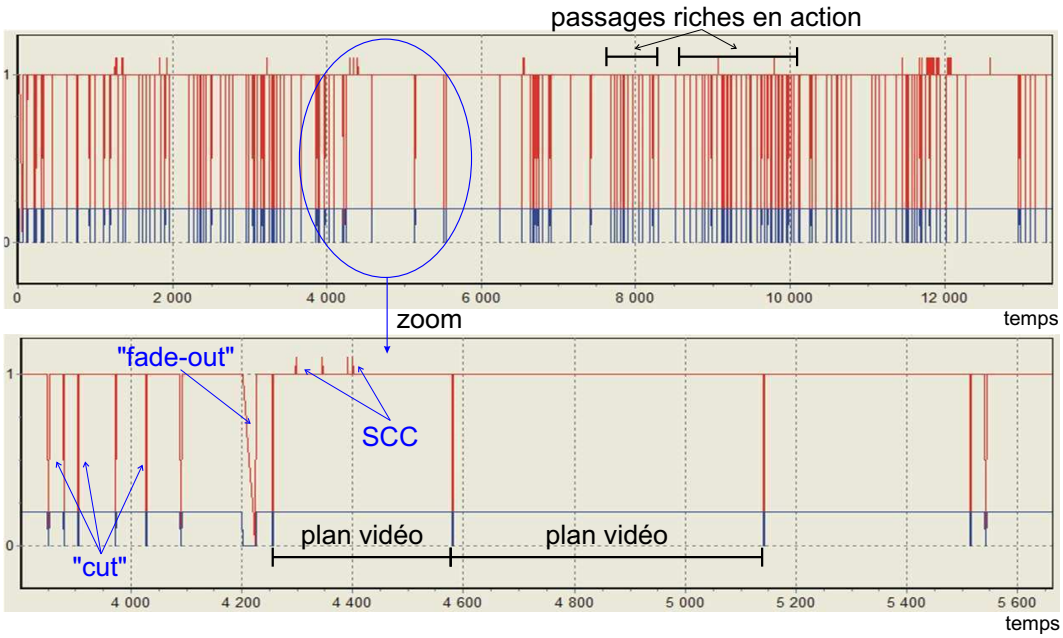


FIG. 2.19 – Annotation visuelle. La distribution des transitions est marquée avec la ligne rouge et les plans vidéo avec la ligne bleue.

particulières des signaux associés à chaque transition sont présentées dans la Figure 2.20.

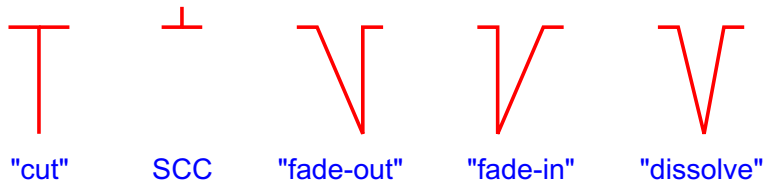


FIG. 2.20 – Les symboles utilisés pour l'annotation visuelle.

L'intérêt de l'annotation proposée est double. Dans un premier temps c'est un *outil d'analyse de la structure* de la séquence permettant aux spécialistes et aussi aux autres utilisateurs d'étudier la façon dont la séquence a été créée, ainsi que la fréquence de l'utilisation de certaines transitions. Le graphe fourni correspond au cheminement du montage de la séquence où les plans vidéo ont été juxtaposés les uns aux autres à l'aide de certaines transitions, formant ainsi le film.

Dans un deuxième temps, la visualisation globale de l'ensemble des transitions permet déjà, d'un seul coup d'œil, de faire une première *analyse sémantique* de la séquence. Par exemple un "dissolve" est souvent utilisé pour introduire un retour en arrière. De la même façon, avoir de nombreux "dissolves" dans une séquence donne une signification particulière. La fréquence des changements des plans est liée au rythme de la séquence : plus il y a de "cuts" par unité de temps, plus il y a de changements visuels et donc plus le rythme est important. Ainsi, en localisant les régions comportant une densité élevée de lignes verticales, et donc de changements de plans, nous pouvons alors déterminer les passages importants du film comportant beaucoup d'action (voir la Figure 2.19). Généralement une scène riche en

action correspond aux passages comportant un nombre élevé de changements visuels.

2.9 Paramètres de bas niveau des plans

En utilisant la distribution des transitions vidéo obtenue nous proposons de calculer un certain nombre de *paramètres de bas niveau* (voir le rapport [Ionescu 05b]). Les paramètres envisagés serviront de point de départ pour la caractérisation sémantique du contenu de la séquence.

2.9.1 L'analyse de la distribution des plans

Dans un premier temps, nous définissons un indicateur de base reliée à la structure temporelle de la séquence. Cet indicateur, noté $\zeta_T(i)$, représente le nombre de changements de plans survenus dans une plage de durée T (exprimée en secondes) démarrant à l'image de l'instant i. La cadence vidéo des séquences étant de 25 images par seconde, cette durée T correspond à $T \cdot 25$ images. L'indicateur $\zeta_T(i)$ est calculé pour chaque image. Il existe donc un fort recouvrement entre les fenêtres de mesure comme le montre la Figure 2.21.

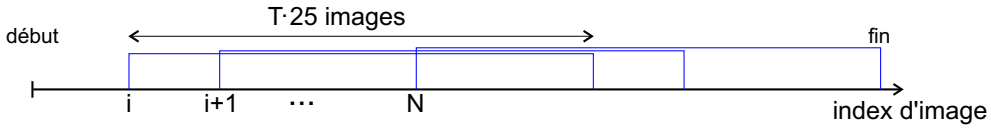


FIG. 2.21 – Les fenêtres de mesure de l'indicateur $\zeta_T(i)$ (N est le nombre total de fenêtres d'analyse).

Ce recouvrement alourdit très légèrement les calculs, mais évite d'avoir à choisir la valeur d'un recouvrement plus faible. On peut noter qu'entre les paramètres introduits, il existe la relation suivante :

$$N = (T_{film} - T) \cdot 25 + 1 \tag{2.29}$$

où N est le nombre total de fenêtres d'analyse et T_{film} est la durée totale de la séquence mesurée en seconde.

A partir de $\zeta_T(i)$ nous définissons deux caractéristiques de la distribution des plans : la *vitesse moyenne de changements* de plans et *la mesure de l'action*.

La vitesse moyenne de changements de plans

Le premier paramètre très caractéristique de la séquence est la valeur moyenne de $\zeta_T(i)$, notée $\bar{v}_T$ et appelée dans la suite *vitesse moyenne de changements de plans*. Il peut être défini de la manière suivante. Les valeurs de $\zeta_T(i)$ sont considérées comme les réalisations particulières d'une variable aléatoire discrète, dont la densité de probabilité est :

$$f_{\zeta_T}(t) = \frac{1}{N} \sum_{i=1}^{N} \delta(\zeta_T(i) - t) \tag{2.30}$$

où N est le nombre total de fenêtres temporelles d'analyse d'une durée de T secondes, $\delta(t) = 1$ pour $t = 0$ et 0 ailleurs, et i est l'indice de la fenêtre courante analysée. $\bar{v}_T$ s'exprime alors

par :

$$\bar{v}_T = E\{\zeta_T\} = \sum_{t=1}^{T \cdot 25} t \cdot f_{\zeta_T}(t) \tag{2.31}$$

Ce paramètre est bien sûr dépendant de la valeur de T. Le choix de cette valeur sera discuté à la fin de cette sous-section 2.9.1.

La mesure de l'action

Dans un second temps en utilisant l'évolution de l'indicateur $\zeta_T(i)$, nous définissons une seconde caractérisation liée à l'action contenue dans la séquence. Cette relation entre la fréquence des changements de plans et l'action est très souvent utilisée dans les techniques de génération automatique de résumés sémantiques de séquences, comme les "bande-annonces". Les passages les plus importants de la séquence sont mis en évidence en utilisant des informations statistiques sur le pourcentage de changements de plans [Hauptmann 98] ou sur la vitesse de changements des contours [Lienhart 97]. Un état de l'art sur les techniques d'extraction de résumés de plus haut niveau est présenté dans [Truong 06] (voir aussi le Chapitre 6 sur la construction des résumés).

L'action correspond à des passages (segments) de la séquence où la fréquence des changements de plans est élevée. Pour cela nous définissons d'abord les *segments d'action*. Ils sont obtenus après 4 étapes :

- **a. seuillage.** Dans un premiere temps, nous construisons un signal binaire fonction du temps, défini par :

$$action(i) = \begin{cases} 1 & \text{si } \zeta_T(i) > \bar{v}_T \\ 0 & \text{sinon} \end{cases} \tag{2.32}$$

  c'est à dire que nous ne gardons que les passages de la séquence où le nombre de changements de plans est élevé (supérieur à la moyenne de la séquence entière, voir le graphe *a* dans la Figure 2.22).

- **b. concaténation.** Les effets SCC, contenant des informations importantes et attractives, sont marqués comme des segments d'action. Les segments d'action voisins, c'est à dire d'une distance inférieure à la taille de la fenêtre d'analyse T seront concaténés en un seul segment. Cette étape nous permet d'effacer les petits trous présents dans les segments d'action obtenus après l'étape de seuillage (voir le graphe *b* dans la Figure 2.22).

- **c. effacement.** Les segments d'action d'une durée inférieure à la taille T de la fenêtre d'analyse sont effacés. Cette étape nous permet d'enlever les petites segments isolés (voir le graphe *c* dans la Figure 2.22).

- **d. nettoyage.** Enfin, tous les segments d'action ne contenant qu'un seul plan vidéo sont effacés. Ces segments sont de faux passages d'action contentant une ou plusieurs transitions graduelles comme par exemple un "fade-out" suivi d'un "fade-in" ce qui entraîne une valeur élevée de ζ_T (voir le graphe *d* dans la Figure 2.22).

Concernant la valeur de la durée de la fenêtre d'analyse, T, plusieurs tests ont été effectués sur un certain nombre de films d'animation de [CICA 06] pour différentes valeurs de T,

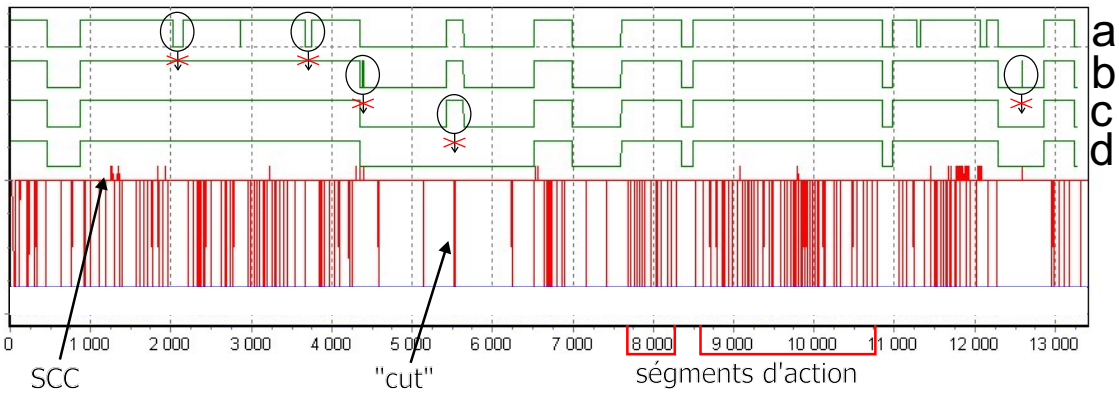


FIG. 2.22 – Exemple de segments d'action pour le film "François le Vaillant" [Folimage 06b]. L'annotation visuelle des transitions est marquée par la ligne rouge, les segments d'action sont marqués par la ligne verte ($T = 5s$). Les lettres de a à d correspondent aux étapes de calcul.

$T \in \{1, ..., 10\}$ secondes. La valeur de T est liée à la granularité des segments d'action obtenus. En utilisant de faibles valeurs de T on obtient une densité élevée de segments d'action de courte durée et donc une sur-segmentation des segments. Inversement, lorsque la valeur de T augmente, la séquence comportera moins de segments d'action. Un bon compromis entre la taille des segments d'action et leur nombre a été obtenu pour $T = 5s$. Des exemples de segments d'action obtenus pour différentes valeurs de T sont présentés dans la l'Annexe B.

En utilisant les segments d'action obtenus, on définit un deuxième paramètre de bas niveau, R_{action}, qui représente le pourcentage de segments d'action par rapport à la séquence entière. Il est défini par :

$$R_{action} = \frac{T_{action}}{T_{film}} \tag{2.33}$$

où T_{action} est la durée totale des segments d'action et T_{film} est la durée totale du film.

2.9.2 L'analyse des transitions

Un autre paramètre proposé tient compte des transitions "fade-in", "fade-out" et "dissolves". Ce paramètre est important du point de vue sémantique parce que les transitions vidéo sont utilisées dans les films avec un but bien précis comme nous l'avons déjà mentionné dans le chapitre précédent. Par exemple, les "fades" sont utilisés pour faire une transition lente entre deux scènes différentes et également pour augmenter le suspens. Les "dissolves" sont souvent utilisés pour modifier le temps de l'action (ils sont alors associés à des retours en arrière) ou simplement comme les "fades" pour introduire une transition lente. On définit alors le rapport des transitions, R_{trans}, comme :

$$R_{trans} = \frac{T_{fades} + T_{dissolves}}{T_{film}} \tag{2.34}$$

où T_{fades} et $T_{dissolves}$ sont les durées totales des transitions "fade-in", "fade-out" et "dissolve", et T_{film} est la durée totale du film.

De la même façon, nous avons défini le paramètre R_{SCC} qui traduit l'importance des effets couleurs de courte durée, particuliers aux films d'animation, nommés SCC (voir la

Section 2.5). Comme ce type d'effet correspond aux changements rapides de couleurs le film aura une signification particulière si ce type d'effet est souvent utilisé. R_{SCC} est défini par :

$$R_{SCC} = \frac{T_{SCC}}{T_{film}} \tag{2.35}$$

où T_{SCC} est la durée totale des SCC sur toute la séquence.

En résumé, les 4 paramètres de bas niveau que nous avons défini à partir des plans vidéo sont :

- **la vitesse moyenne** des changements de plans, $\bar{v}_T$,
- **le rapport d'action**, R_{action},
- **le rapport des transitions**, R_{trans},
- **le rapport des SCC**, R_{SCC}.

Ces paramètres seront utilisés ultérieurement, dans la deuxième partie de la thèse, comme point de départ pour la caractérisation sémantique des séquences d'image.

2.10 Conclusions générales

Le découpage en plans est *l'étape de base* de toutes les méthodes d'analyse de bas niveau, mais aussi d'analyse sémantique, des séquences d'images. Les plans sont liés les uns aux autres par des transitions vidéo.

Dans le contexte particulier des films d'animation nous avons été amenés à proposer de nouvelles méthodes de détection des transitions vidéo les plus souvent utilisées dans les films d'animation : les "cuts", les "fades" et les "dissolves". Pour la détection de "dissolves" nous avons utilisé une méthode existante qui a été adaptée au contexte des films d'animation. De plus, nous avons analysé un effet particulier qui a été appelé "changement bref de couleurs" ou SCC ("short color change") qui a une signification sémantique importante dans l'analyse de la séquence.

La détection des transitions et des effets de couleurs (par exemple les SCC dans les films d'animation ou les flashs dans les vidéos) est également importante car ces effets sont souvent détectées comme des "cuts". Leur détection permet ainsi d'améliorer le découpage en plans.

Les méthodes spécifiques de détection des transitions que nous avons proposées sont les suivantes :

- **détection des "cuts"** : la méthode proposée est basée sur l'analyse des distances entre histogrammes couleurs. Elle utilise un certain nombre d'améliorations et est adaptée au domaine spécifique des films d'animation : les images sont divisées en quadrants pour réduire l'influence de l'apparition/disparition des personnages/objets dans la scène, la dérivée seconde est utilisée pour réduire l'influence du mouvement répétitif de la caméra sur la détection. Un seuillage automatique qui permet la détection des "cuts" a été également mis en place. La méthode proposée a donné de bons résultats comparée à deux autres approches : une approche classique basée sur l'histogramme et une approche basée sur la discontinuité du mouvement. Les deux taux de détection obtenus, la précision et le rappel, sont élevés et supérieurs à 92%. (voir la Section 2.4.3). En ce qui concerne le temps de calcul, selon la réduction des couleurs utilisée, une implantation en temps réel est possible après l'optimisation des procédures de calcul. La

réduction couleur dans l'espace Lab et la diffusion d'erreur sont les seules opérations ayant une complexité de calcul élevée. Dans ce cas, pour une implantation en temps réel, des dispositifs hardware spécifiques sont nécessaires.

– **détection des SCC** : la méthode proposée a été inspirée des méthodes de détection de flashs dans les films naturels. L'intérêt de la détection de cet effet particulier est double : premièrement cela permet de corriger la détection des "cuts" et deuxièmement cela apporte des informations sémantiques sur le contenu de la séquence. Testé sur un certain nombre de films d'animation, la méthode proposée a donné de bons résultats : un taux de précision de 93% et un taux de rappel de 88.3%. En ce qui concerne les performances, la complexité de calcul est réduite. L'algorithme peut être implanté en parallèle à la détection de "cuts" sans alourdir les traitements (voir la Section 2.5).

– **détection des "fades"** : La méthode proposée utilise des mesures statistiques. Elles sont calculées sur l'intensité des pixels mais aussi sur l'information de chrominance qui est moins sensible aux fluctuations de l'intensité lumineuse et à la présence du bruit dans les images. Testée sur un certain nombre de films d'animation, la méthode proposée a obtenu de bons taux de détection. Pour la détection des "fade-in" la précision est de 81.9% et le rappel de 97.2%, et pour la détection des "fade-out" la précision est de 91.4% et le rappel de 94.6% (voir la Section 2.6). La complexité de calcul est réduite car les paramètres proposés ne demandent qu'un seul balayage de l'image ce qui permet une implantation en temps réel.

L'utilisation des transitions dans la séquence n'est pas aléatoire. Chaque transition a un but précis. Par exemple les "cuts" sont les transitions usuelles qui changent le point de vue ou la scène, les "dissolves" sont utilisées généralement pour changer le temps de l'action ou pour revenir en arrière. Dans ce chapitre nous avons proposé un certain nombre de *paramètres de bas niveau* permettant la caractérisation du contenu de la séquence (voir la Section 2.9). Ils serviront comme point de départ pour la caractérisation sémantique des plans, proposé dans la deuxième partie de cette thèse.

Chapitre 3

L'Analyse du mouvement

Résumé : *L'évolution temporelle de l'information visuelle est une des caractéristiques fondamentales des séquences d'images. Dans ce chapitre nous présentons les différentes directions que nous avons étudiées pour extraire l'information de mouvement présente dans les séquences. Nous présentons d'abord un état de l'art sur les techniques d'estimation du mouvement. Puis, nous proposons une méthode de caractérisation du mouvement global de la caméra. Enfin, un certain nombre de paramètres de bas niveau sont calculés, servant de point de départ pour une caractérisation sémantique du mouvement.*

Une des particularités les plus importantes des séquences d'images est la présence de mouvement. Par rapport aux images fixes, une séquence apporte une information supplémentaire, qui se caractérise par une certaine évolution continue du contenu des images dans le temps. On peut alors dire que les séquences d'images sont des *"images en mouvement"*.

Si aujourd'hui, de nombreux moteurs de recherche permettent de trouver en quelques secondes de nombreuses informations stockées sous format texte, aucun mécanisme ne permet de retrouver de la même manière des contenus multimédia. Afin de pallier ce manque, des chercheurs du Moving Picture Experts Group (MPEG) ont développé un nouveau standard ISO connu sous le nom de MPEG-7. Selon leurs créateurs, "la principale ambition de MPEG-7 est de rendre les informations multimédia aussi faciles à trouver sur le Web que le texte l'est aujourd'hui". Ainsi, en ce qui concerne le mouvement, le standard MPEG-7 *intègre des informations obtenues à partir des meilleures méthodes existantes d'analyse du mouvement.* Ce sont ces méthodes que nous allons détailler par la suite.

En ce qui concerne le mouvement, on trouve deux grandes directions d'analyse : d'une part l'analyse du *mouvement global* de la caméra et d'autre part l'analyse locale du *mouvement des objets* [Jeannin 01]. Ces deux directions d'analyse sont présentées dans la Figure 3.1.

Le mouvement global. L'analyse du *mouvement global* est effectuée au niveau des segments vidéo ou des groupes d'images. Une première information analysée est le *mouvement de la caméra*. Dans cette étape, le type particulier de déplacement de caméra est déterminé parmi tout un ensemble de mouvements possibles. Typiquement les informations retenues,

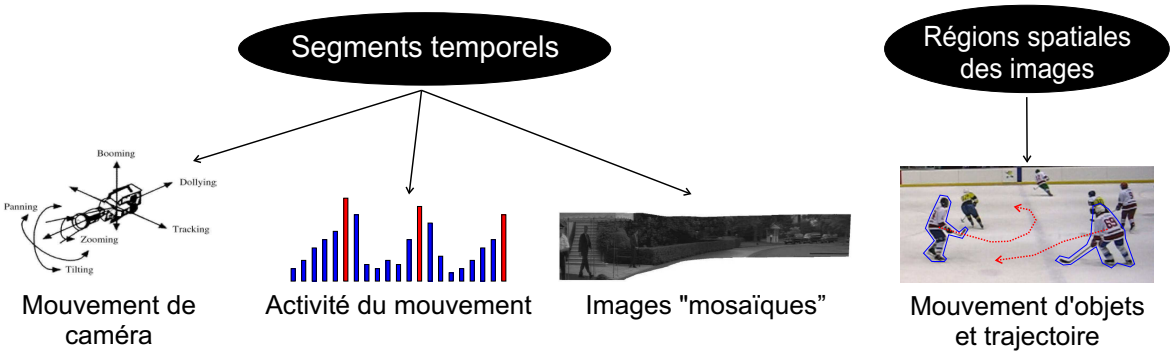


FIG. 3.1 – Les principales directions d'analyse du mouvement dans les séquences d'images : niveau global (segments temporels) et niveau local (régions spatiales) (source MPEG-7).

pour un mouvement spécifique, sont l'amplitude du mouvement, sa position dans la séquence et sa durée. Cette étape d'analyse est très importante pour la compréhension du contenu de la séquence. Dans certaines situations, à l'aide du mouvement de caméra, on peut localiser les passages importants de la séquence comme par exemple le fait de focaliser l'attention des spectateurs (arrêt sur une scène précise, puis zoom sur le visage d'un personnage).

Une seconde information souvent exploitée est *l'activité du mouvement* qui est une mesure de la perception que l'on a du mouvement en regardant la séquence. Une classification du mouvement global est réalisée à travers un certain nombre de paramètres de bas niveau. Cette classification en plusieurs niveaux d'activité concerne l'intensité de l'action, auquel est ajouté un niveau correspondant à l'absence d'action ("pace of action"). Une activité intense est liée aux événements dynamiques comme par exemple les scènes de but dans les matchs de football ou les poursuites de voitures. Au contraire, une activité faible est liée aux scènes comportant un mouvement faible (interviews, journaux télévisés).

Une autre utilisation de l'information sur le mouvement global est la construction *d'images "mosaïques"* (voir [Irani 95] ou [Aner 01]). Une "mosaïque" est la représentation sous la forme d'une seule image statique du contenu global d'une scène constituée d'images en mouvement. Elle est réalisée en regroupant les différentes images de la scène, après un recalage prenant en compte le déplacement global dans la scène (voir Figure 3.1). Les images "mosaïques" sont utilisées comme des résumés compacts des différents segments de la séquence. La complexité des calculs peut être simplifiée grâce aux paramètres de "déformation" standardisés dans le format MPEG-7.

Le mouvement local. La deuxième direction d'analyse est la caractérisation du *mouvement local ou mouvement des objets* qui n'affecte que des régions de pixels de l'image. Pour ce type d'analyse, contrairement à la caractérisation globale du mouvement, les mesures de déplacement sont effectuées au niveau du pixel. Les méthodes existantes utilisent généralement une modélisation paramétrique du mouvement permettant de retrouver dans la séquence les mêmes objets avec des déplacements similaires, malgré d'éventuelles déformations géométriques apportées par des zooms, rotations, etc. Les résultats de cette analyse sont souvent résumés dans la *trajectoire* des objets qui est définie par l'évolution temporelle de certains points d'intérêt de l'objet, comme le centre de gravité ou certains points de contour.

Dans la littérature spécialisée, les méthodes d'analyse de la trajectoire des objets, dans le

contexte de l'indexation du contenu vidéo, ont été beaucoup plus abordées que les méthodes d'analyse du mouvement global (mouvement de caméra). Cela vient du fait que, dans une séquence, la plupart des événements importants sont liés à des mouvements d'objets. Par exemple dans une séquence sportive, il sera plus intéressant de caractériser et de segmenter la trajectoire d'un joueur qui est en train de marquer un but que de caractériser le mouvement global de caméra, qui a suivi ce joueur. Pour un état de l'art sur les techniques d'analyse du mouvement des objets on pourra se rapporter à [Smith 04] [Fablet 02] [Dagtas 00].

En conclusion, toutes les méthodes d'analyse du mouvement s'appuient sur le résultat de *l'estimation du mouvement* qui est obtenu par la mesure du déplacement d'un pixel, ou d'une région de pixels, entre l'image courante et l'image suivante. Dans la suite de ce chapitre nous allons nous focaliser sur l'étude du mouvement de caméra.

3.1 L'estimation du mouvement

Le principe de l'estimation du mouvement repose sur la recherche du déplacement d'un pixel, ou d'un bloc de pixels, entre deux images successives minimisant la variation d'intensité de ce pixel, ou blocs de pixels (DFD = "displaced frame difference"). Ce principe repose sur l'hypothèse que l'intensité du pixel, ou du bloc, n'a pas sensiblement varié d'une image à la suivante. Ceci peut s'exprimer par :

$$DFD(\vec{r}, \vec{d}, \Delta t) = I(\vec{r} + \vec{d}, t + \Delta t) - I(\vec{r}, t) \tag{3.1}$$

où $\vec{r}$ est la position du pixel, ou du bloc, dans l'image courante analysée, $\vec{d}$ est le vecteur de déplacement entre les instants t et $t + \Delta t$ et I est l'image.

Typiquement les méthodes d'estimation du mouvement emploient des algorithmes de minimisation d'une fonction de coût liée au déplacement entre pixels ou blocs de pixels. On trouve plusieurs approches (voir [Marichal 98]) :

- **les méthodes différentielles** : *les méthodes différentielles* basées sur l'équation du flot optique (comme les méthodes itératives et "pel-recursives") donnent comme résultat un champ vectoriel dense. Pour ces méthodes, la présence du bruit dans l'image réduit fortement la précision de l'estimation. De plus le temps de traitement est très élevé.
- **les méthodes paramétriques** : *les méthodes paramétriques* modélisent les déplacements des pixels de l'image en utilisant un certain nombre de paramètres. Le problème de l'estimation du mouvement dans ce cas est transformé en un problème d'estimation des paramètres du modèle, qui peut être un modèle affine, quadratique, etc.
- **les méthodes stochastiques** : *les méthodes stochastiques* utilisent des modèles probabilistes. L'exploration de l'espace des paramètres est guidée par des processus aléatoires (modèle Bayesien, modèle Markovien ou algorithme génétique). Les méthodes stochastiques introduisent une complexité de calcul élevée mais les résultats obtenus sont plus proches de la réalité.
- **les approches par blocs** : dans *les approches par blocs*, l'estimation du mouvement est faite par l'analyse de blocs de pixels ("block-based", voir l'Annexe C). Ce type d'approche a été proposé pour la première fois dans [Jain 91] et se trouve être le meilleur compromis entre la complexité des calculs et la qualité des résultats obtenus. Le choix de la taille des blocs utilisés est important et résulte d'un compromis sensibilité/robustesse. Des blocs de dimension réduite donne une *sensibilité élevée* de la

détection, situation adaptée aux faibles déplacements des objets. Une taille de blocs élevée confère une *bonne robustesse* à la présence de bruit dans les images. Grâce à ces propriétés les méthodes par blocs ont été utilisées dans la plupart des standards de codage vidéo, comme par exemple les standards H.263, MPEG 1, 2, 4 et 7.

Lorsque les séquences d'images sont au format MPEG, une solution pour récupérer les vecteurs de mouvement est de les extraire directement du *flux MPEG* (voir [Pilu 97], [Gilvarry 99]). Les vecteurs de mouvement du flux MPEG ont en effet été calculés au moment du codage de la séquence d'images. On peut supposer que la qualité des images utilisées au moment du codage est supérieure à la qualité des images après décodage car le codage est un codage avec perte d'information. Dans la réalité, l'information du mouvement issue du flux MPEG n'est pas partout cohérente. En effet, il existe certaines régions qui nécessiteraient une étape de correction (voir Figure 3.2). Par exemple les vecteurs de mouvement ne sont pas présents dans les régions de l'image qui ne sont pas texturées [Pilu 97]. Si l'on souhaite faire une analyse sémantique fine du mouvement, les résultats fournis directement par le flux MPEG ne sont pas suffisants. Il est alors nécessaire de décompresser la séquence et d'analyser précisément le mouvement [Pineau 05].

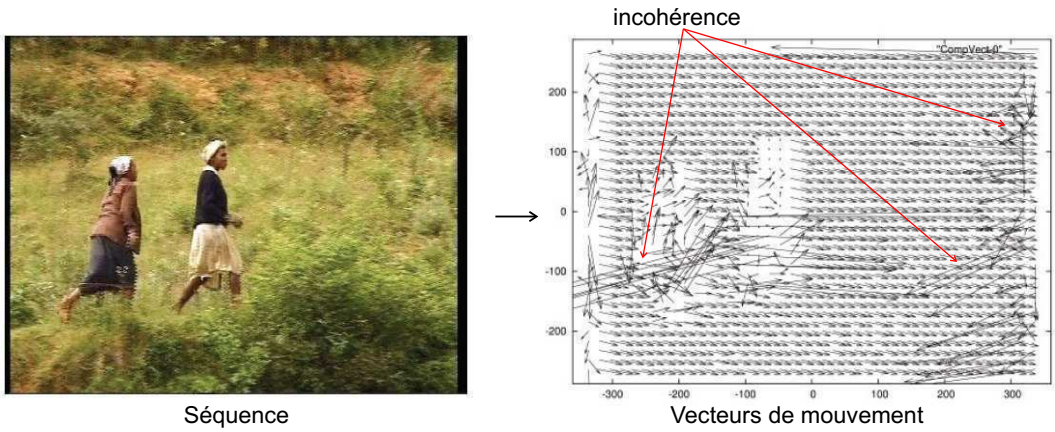


FIG. 3.2 – Exemple des vecteurs de mouvement obtenus à partir du flux MPEG2 (source des données : Projet Analyse et Indexation Vidéo [Pineau 05]).

Notre objectif étant une *analyse sémantique du mouvement* dans les séquences d'images, nous avons donc choisi d'estimer le mouvement en utilisant une méthode par blocs de pixels, car la précision du résultat se règle en jouant sur la taille des blocs et le temps de calcul s'en trouve réduit. Pour la caractérisation du mouvement de caméra, il n'est pas nécessaire d'avoir un champ vectoriel dense. Les méthodes par blocs sont le meilleur compromis entre la précision de l'estimation et la complexité de calcul. Nous avons testé trois de ces méthodes : la recherche complète, la recherche logarithmique et l'algorithme du standard H.263+ (voir l'Annexe C). Nous utiliserons cette estimation du mouvement pour la caractérisation du mouvement de caméra dans les séquences d'images.

3.2 L'analyse du mouvement de caméra

Parmi les applications utilisant l'estimation du mouvement nous nous sommes intéressés à la caractérisation du mouvement global de la caméra dans la scène. Ce type d'analyse permet

d'avoir une caractérisation globale de l'action du film. Par exemple les changements de plans vidéo utilisent souvent des mouvements de translation pour se focaliser sur certains points d'intérêt de la scène, une action importante est fréquemment accompagnée d'un mouvement rapide de caméra, les personnages sont mis au premier plan en utilisant un zoom, etc.

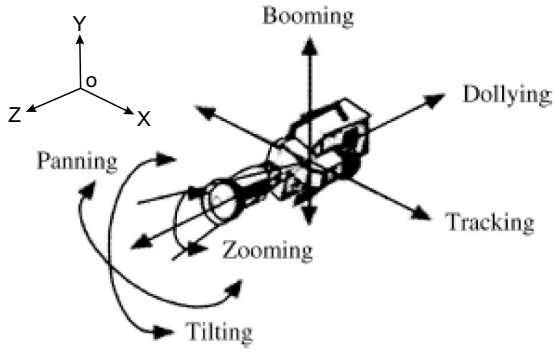


FIG. 3.3 – Les mouvements de caméra.

Globalement le mouvement d'une caméra est un mouvement libre dans l'espace 3D. Mais dans la réalité, à cause des contraintes techniques et physiques de la caméra (le nombre réduit de degrés de liberté), seul un nombre restreint de mouvements sont réalisables. Les mouvements plus généraux sont en effet une approximation d'un mouvement libre dans l'espace infini 3D, et ils sont réalisés comme une suite de mouvements primaires. Les plus importants sont illustrés dans la Figure 3.3.

On retrouve deux types différents de mouvements :

- d'une part il y a *les mouvements de translation* selon les trois axes XYZ, comme le "dollying" ou "zoom out" (agrandissement), le "zoom in" (rétrécissement), le "tracking" (translation vers la droite ou vers la gauche) et le "booming" (translation vers l'haut ou vers le bas).
- d'autre part, il y a *les mouvements de rotation*, comme les rotations horaires et anti-horaires (dans le plan XoY), le "panning" (rotation dans le plan XoZ) et le "tilting" (rotation dans le plan YoZ).

3.2.1 État de l'art

Les méthodes d'analyse du mouvement de caméra peuvent être regroupées en deux catégories principales : les méthodes qui analysent l'information de mouvement directement dans le *domaine compressé* (flux MPEG) et les méthodes qui font l'analyse dans le *domaine spatio-temporel* des images décodées du flux vidéo. On trouve différents états de l'art sur l'analyse du mouvement de caméra dans [Kramer 05], [Ngo 00], [Tardini 05] ou [Duan 06]. Nous présentons, dans la suite, les points les plus marquants de ces approches.

L'analyse du mouvement dans le domaine compressé

Une approche probabiliste pour la détection du mouvement de caméra de type "zoom in/out" est proposée dans [Jin 02]. Elle utilise l'algorithme EM (Esperance-Maximization)

pour l'estimation de la probabilité d'occurrence d'un mouvement de type "zoom" par rapport aux autres mouvements. L'information de mouvement est récupérée directement à partir du flux MPEG-1 ou MPEG-2. L'avantage de l'approche probabiliste est d'être moins sensible à la présence de bruit affectant les vecteurs de mouvement, bruit causé par les erreurs de quantification ou la présence d'artefacts liés au codage.

Une approche générale pour la détection des 6 mouvements de base de caméra est proposée dans [Kim 04]. La méthode proposée utilise une interprétation qualitative d'un certain nombre de paramètres des modèles de mouvement, paramètres qui sont estimés directement à partir du flux MPEG-2. Les vecteurs de mouvement récupérés du flux MPEG-2 sont comparés avec le modèle affine du mouvement :

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} \tag{3.2}$$

où (u, v) est le vecteur de mouvement du bloc de pixels centré sur les coordonnées (x, y).

D'une manière générale, le mouvement global pour une certaine image peut être exprimé comme un vecteur de paramètres $\phi = (a_1, a_2, ..., a_6)$. Le vecteur ϕ est estimé à partir du champ vectoriel de mouvement en utilisant la méthode des moindres carrés. [Kim 04] propose de représenter l'information de mouvement par des vecteurs de paramètres qui sont exprimés en fonction des mouvements de caméra. Chaque image est paramétrée par un vecteur $\phi_c = (pan,\ tilt,\ div,\ rot,\ hyp)$ où les 5 paramètres utilisés sont définis en fonction du modèle affine du mouvement :

$$pan = a_1, \quad tilt = a_4 \tag{3.3}$$

$$div = \frac{1}{2}(a_2 + a_6), \quad rot = \frac{1}{2}(a_5 - a_3) \tag{3.4}$$

$$hyp = \frac{1}{4}(|a_2 - a_6| + |a_3 + a_5|) \tag{3.5}$$

où *pan* représente le mouvement de translation horizontale, *div* représente le mouvement de type "zoom" et *hyp* représente le "flou hyperbolique" qui correspond typiquement aux situations de mouvements prédominants d'objets.

Les différents mouvements de caméra ("tracking", "tilting", "rotation" et "zoom") sont déterminés par le seuillage de ces vecteurs de paramètres. Une approche similaire qui utilise le paramétrage du modèle affine du mouvement du flux MPEG est proposée dans [Kramer 05]. L'algorithme proposé a des performances 3 à 4 fois plus rapides que le temps réel (= 25 images par seconde).

Dans la méthode proposée par [Lee 02], la classification des mouvements de caméra est réalisée en comparant les mouvements mesurés à ceux issus de certains modèles prédéfinis. Dans un premier temps, le champ des vecteurs de mouvement, pour une image, est calculé à partir du flux MPEG. Les vecteurs de mouvement obtenus pour les différents blocs de l'image répondent forcément à l'une des deux alternatives suivantes : soit ils appartiennent à des régions du fond, soit ils appartiennent à des régions contenant des objets. Les mouvements de caméra sont alors déterminés par comparaison aux modèles prédéfinis (voir Figure 3.4). La similarité entre vecteurs est mesurée par la distance entre les histogrammes des phases des vecteurs mouvement de chaque bloc.

L'avantage principal des méthodes d'analyse du mouvement de la caméra utilisant le domaine compressé est le temps de calcul. Elles permettent des performances supérieures au temps réel. Cependant, la précision des vecteurs de mouvement utilisés est directement

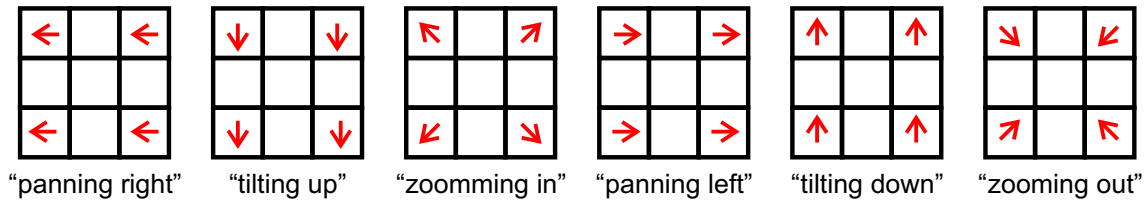


FIG. 3.4 – Les modèles de mouvement de caméra utilisés dans [Lee 02].

proportionnelle au niveau de compression du flux vidéo. Et souvent les vecteurs ainsi obtenus ne modélisent pas proprement le mouvement réel. Pour améliorer l'analyse une solution consiste alors à décompresser les données et à ré-estimer plus correctement le mouvement.

L'analyse du mouvement dans le domaine spatio-temporel de l'image

Les méthodes d'analyse du mouvement de caméra qui utilisent l'information spatio-temporelle de la séquence sont très variées. Une grande diversité de techniques ont été proposées, comme l'utilisation de modèles prédéfinis, l'analyse des volumes spatio-temporels, l'utilisation des réseaux neuronaux, l'utilisation d'ondelettes, etc.

Par exemple [Akutsu 92] propose de faire la reconnaissance des différents mouvements de caméra, en comparant les vecteurs de mouvement obtenus avec des modèles prédéfinis, comparaison effectuée dans l'espace de la transformée de Hough. [Xiong 98] propose une méthode d'analyse du flot optique en analysant les projections des vecteurs de mouvement sur les axes oX et oY. Dans [Bouthemy 99] l'information du mouvement est estimée en s'appuyant sur un modèle global 2D affine du mouvement. La détection des différents types de mouvement de caméra est effectuée par le seuillage d'une mesure de similarité sur les paramètres du mouvement.

Une approche inédite est proposée dans [Ngo 00]. La caractérisation du mouvement de caméra et des objets est effectuée par l'analyse des volumes spatio-temporels des images. Une séquence peut être considérée comme un volume où les 2 premières dimensions sont les dimensions spatiales (x, y) et la troisième dimension est le temps t. En considérant un autre point de vue, on peut représenter ce volume comme un ensemble de couches temporelles 2D, sur les dimensions (x, t) (couche horizontale) et (y, t) (couche verticale). Dans l'espace ainsi formé, les différents types de mouvements sont représentés par des modèles orientés (voir Figure 3.5). La caractérisation de ces modèles est effectuée par le calcul des histogrammes des tenseurs (les tenseurs étant des dérivées partielles selon le temps et les axes xy) qui servent ensuite à la caractérisation du mouvement de caméra ou des objets.

FIG. 3.5 – Les différents types de mouvements de caméra en utilisant une représentation par des couches [Ngo 00].

Par rapport aux méthodes utilisant le domaine compressé, l'analyse du mouvement dans le domaine spatio-temporel de l'image est moins rapide mais plus précise. Le domaine spatio-temporel (voir la suite d'images) a l'avantage d'être la source d'autres types d'informations, plus riches et moins affectées par le bruit que celles fournies par les coefficients MPEG.

3.2.2 Méthode proposée

La méthode d'analyse du mouvement de caméra que nous proposons utilise l'information du mouvement mesurée dans le domaine décompressé. Les vecteurs de mouvement sont calculés en utilisant une approche d'estimation par blocs, qui se trouve d'être le meilleur compromis entre la précision de l'estimation et la complexité de calcul.

La classification des différents mouvements de caméra est basée sur la comparaison d'un certain nombre de paramètres de mouvement à des *modèles prédéfinis* (voir [Ionescu 07b] ou le rapport [Ionescu 05c]). Ce type d'approche a l'avantage d'être également un bon compromis performance/complexité, ce qui en fait une des approches la plus utilisée. Pour la comparaison, nous utilisons des règles de décision basées sur un certain nombre de seuils. La méthode est inspirée des travaux proposés dans [Lee 01] où les auteurs utilisent des modèles prédéfinis de mouvements de caméra et une classification par réseaux de neurones.

Par rapport à cette méthode (voir [Lee 01]), notre approche apporte un certain nombre de modifications ou d'améliorations liées au contexte que nous avons abordé. D'abord, elle classifie un nombre plus élevé de types de mouvements de caméra et gère les situations de discontinuité du mouvement et le mouvement d'objets. Ensuite, les modèles prédéfinis utilisés sont plus élaborés. Pour leur construction nous avons en effet pris en compte des situations plus complexes, telles que les erreurs d'estimation des vecteurs de mouvement ou la similarité de certains modèles correspondant à des mouvement de caméra différents. Enfin, l'étape de classification a été simplifiée. Les réseaux de neurones ont été remplacés par des règles de décision pour réduire le temps de calcul, ce qui permet d'envisager une implantation en temps réel. Le diagramme de la méthode proposée est présenté dans la Figure 3.6.

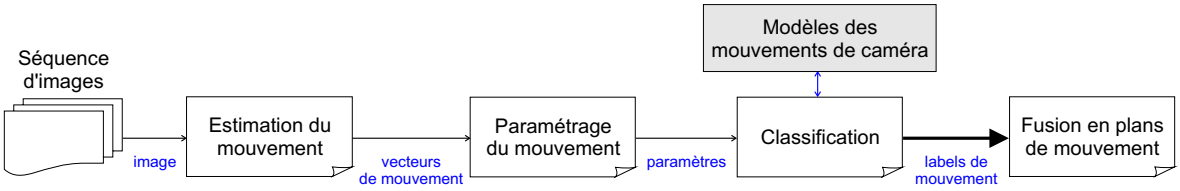


FIG. 3.6 – Méthode proposée pour la classification des mouvements de caméra.

Les étapes d'analyse les suivantes :

- **l'estimation du mouvement** : dans un premier temps pour chaque image de la séquence nous estimons les vecteurs de mouvement en utilisant une méthode par bloc.
- **le paramétrage du mouvement** : à partir du champ vectoriel obtenu à l'étape précédente et en s'appuyant sur des modèles prédéfinis de mouvement, un certain nombre de paramètres sont calculés,
- **les règles de décision** : les paramètres ainsi obtenus sont comparés aux paramètres d'un certain nombre de modèles prédéfinis déterminés empiriquement pour chaque type de mouvement de caméra. Pour la comparaison, nous avons utilisé des règles de

décision basées sur des seuils.
- **la fusion des résultats - les plans de mouvement** : en analysant l'ensemble des résultats obtenus sur chaque image, il arrive que l'on obtienne des incohérences : par exemple la présence d'une image correspondant à une rotation de la caméra à l'intérieur d'un ensemble d'images où l'on a détecté une translation. Ces "points faux" peuvent être éliminés par une étape de fusion qui se base sur une hypothèse de continuité des mouvements de caméra et permet de générer des plans de mouvement.

Les classes de mouvements à détecter

Parmi les différents mouvements possibles de caméra (voir Figure 3.3), l'algorithme proposé permettra de détecter les situations suivantes :

- **"pas de mouvement"**,
- **"discontinuité du mouvement"** : ce sont des situations pour lesquelles il n'y a pas continuité du mouvement entre l'image courante et l'image suivante. Ces situations correspondent aux changements de plans vidéo, aux transitions de mouvement ou aux mouvements de caméra très rapides,
- **"mouvement d'objets"** : déplacement de régions de pixels ou mouvement très faible,
- *"zoom in/out"* : agrandissement/rétrécissement de l'image,
- **"déplacement vers la gauche/droite"** : mouvement de translation dans le plan horizontal,
- **"déplacement vers le haut ou vers le bas"** : mouvement de translation dans le plan vertical,
- **"rotation dans le sens horaire/anti-horaire"** : rotation dans le plan de l'image.

En ce qui concerne les mouvements de rotation de type "panning" et "tilting", présentés dans la Figure 3.3, leurs effets dans la scène sont similaires aux mouvements de translations sur les axes horizontal ou vertical. Nous ferons donc une approximation de ces deux types de mouvement par des mouvements de translation.

L'estimation du mouvement

Pour l'estimation du mouvement nous avons choisi d'utiliser une méthode basée sur l'analyse des blocs de pixels (voir l'Annexe C). Ce type d'approche moins précis en ce qui concerne la densité du champ résultant qu'une méthode basée sur le flot optique, est plus efficace et bien adapté à nos besoins.

Chaque bloc de 16×16 pixels de l'image courante est caractérisé par un vecteur de mouvement (voir Figure C.1), $\vec{d} = (\Delta x, \Delta y)$ où , Δx et , Δy sont les déplacements sur les axes oX et oY du centre du bloc. Le vecteur $\vec{d}$ contient une information sur l'amplitude du mouvement, donnée par le module $|\vec{d}|$, et une information sur l'orientation du mouvement exprimée par l'angle qu'il fait avec l'axe oX, $\theta = arctangente(\frac{\Delta y}{\Delta x})$.

L'estimation du mouvement est réalisée par minimisation, selon le déplacement $\vec{d}$, d'une fonction de coût, $F_c()$, définie comme l'erreur absolue moyenne (MAE) entre les intensités du bloc à l'instant t et les intensités du bloc voisin à l'instant $t + l$:

$$F_c(\vec{d}) = MAE(I(\vec{r}, t), I(\vec{r} + \vec{d}, t + l))$$

où $\vec{r}$ est la position du bloc dans l'image à l'instant t, $\vec{r} + \vec{d}$ la position du bloc dans l'image

à l'instant $t + l$, l étant le pas temporel d'analyse.

Trois situations peuvent alors se produire :

- **pas de mouvement** : la valeur minimale de $F_c()$ est obtenue pour le bloc de la fenêtre de recherche S dans l'image suivante (instant $t + l$) à la même position que le bloc courant analysé dans l'image à l'instant t. Dans cette situation le vecteur de déplacement est nul, il n'y a donc *pas de mouvement*.
- **le mouvement est discontinu** : la valeur minimale de $F_c()$ est très élevée et supérieure à un certain seuil, $\tau_{discont}$. Dans cette situation il n'y a pas de continuité du mouvement car le bloc analysé ne se retrouve pas dans la fenêtre de recherche S, donc le *mouvement est discontinu*. Le bloc est alors marqué avec un label particulier.
- **mouvement du bloc** : la valeur minimale de $F_c()$ est inférieure au seuil de discontinuité, $\tau_{discont}$ et correspond à un bloc différent du bloc courant analysé. Dans cette situation le déplacement obtenu correspond au *mouvement du bloc*.

Le seuil de discontinuité, $\tau_{discont}$ a été choisi empiriquement, $\tau_{discont} = 20000$, après avoir analysé différentes séquences contenant des passages de changements de plan et des passages continus. Un seuil trop bas diminue la sensibilité de l'estimation en augmentant le nombre de blocs de discontinuité dans l'image et un seuil trop élevé conduit à des vecteurs de mouvement incohérents.

Le paramétrage du mouvement

Dès que l'estimation du mouvement est finie, chaque image du film est caractérisée par une matrice de vecteurs de mouvement. A chaque élément de cette matrice est associé l'amplitude $A(i, j)$ et l'orientation $O(i, j)$ du mouvement, avec i, j les indices des blocs dans l'image, $i = 0, ..., N_{oX}$ et $j = 0, ..., N_{oY}$, où N_{oX} et N_{oY} sont les nombres de blocs sur les axes oX et oY.

Dans un premier temps nous proposons deux paramètres calculés sur la matrice d'amplitudes $A(i, j)$, permettant de caractériser globalement le mouvement et la discontinuité du mouvement dans l'image analysée. Ces deux paramètres sont définis par :

$$R_{mouv} = \frac{N_{mouv}}{N_{total}}, \quad R_{discont} = \frac{N_{discont}}{N_{total}} \tag{3.6}$$

où N_{mouv} est le nombre de blocs en mouvement (situation 3 de l'analyse de la fonction de coût $F_c()$), $N_{discont}$ est le nombre de blocs de pixels correspondant à une discontinuité temporelle du mouvement (*situation 2* de l'analyse de la fonction de coût), et N_{total} est le nombre total de blocs dans l'image.

Une autre caractérisation proposée est *la matrice des vecteurs moyens* du mouvement. L'image courante analysée est découpée d'une manière uniforme en 3×3 macro-blocs, chaque macro-bloc étant composé d'un certain nombre de blocs de taille 16×16 pixels. L'objectif est de résumer de manière compacte et plus facile à analyser les résultats obtenus sur les bloc 16×16. A chacun de ces macro-blocs sont associées une amplitude et une direction moyenne.

L'amplitude moyenne des vecteurs de mouvement, notée $\bar{A}$, est calculée comme la moyenne arithmétique des amplitudes des blocs composant le macro-bloc. Pour la direction moyenne, notée $\bar{O}$, le calcul doit prendre en compte la circularité de l'information angulaire. En utilisant des statistiques circulaires [Fisher 93], on calcule l'orientation moyenne des vecteurs de

mouvement selon la formulation suivante :

$$\bar{O} = atan(\frac{b}{a}) \tag{3.7}$$

où a et b sont définis par :

$$a = \sum_i cos(\theta_i), \quad b = \sum_i sin(\theta_i) \tag{3.8}$$

où les valeurs θ_i avec $i = 1, ..., n$ représentent les directions des n blocs 16×16 composant le macro-bloc.

Un exemple est donné dans la Figure 3.7. Les orientations moyennes ainsi calculées seront utilisées pour caractériser les différents mouvements de caméra qui se traduiront par des configurations particulières de ces orientations.

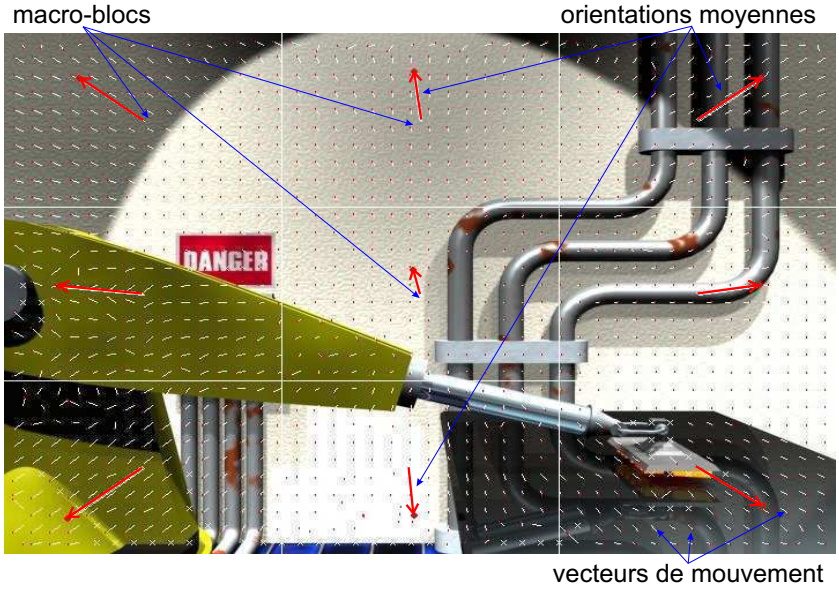


FIG. 3.7 – Exemple d'estimation du mouvement "zoom in" : les vecteurs de mouvement sont représentés par des petites flèches blanches proportionnelles à l'amplitude du déplacement et les orientations moyennes normalisées par la taille de la région sont représentées par de grandes flèches rouges.

Les règles de décision

En analysant les paramètres de bas niveau proposés, R_{mouv}, $R_{discont}$, $\bar{O}(k,l)$ et $\bar{A}(k,l)$ ($k,l \in \{0,1,2\}$ sont les indices des macro-blocs), nous avons défini des règles de classification pour retrouver les mouvements de la caméra envisagés en début de Section 3.2.2. Ces règles s'inspirent des travaux proposés dans [Lee 01].

Chaque situation de mouvement est caractérisée par un ensemble de règles disjointes (liées par l'opérateur "sinon"), ce qui donne :

a. "discontinuité du mouvement" : si $R_{discont} > 0.5$. Dans cette situation le nombre de blocs qui ont une discontinuité du mouvement est plus élevé que la moitié du nombre total

de blocs.

b. "pas de mouvement" : si $R_{mouv} \leq 0.01$. Dans cette situation le nombre des blocs de pixels comportant un mouvement est proche de 0.

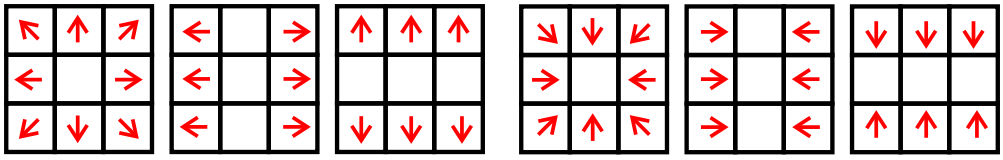
c. "mouvement d'objets" : si $R_{mouv} > 0.01$ et $R_{mouv} \leq 0.5$. Le mouvement d'objets a lieu si le nombre des blocs qui comportent un mouvement est suffisamment élevé mais également inférieur à la moitié du nombre total de blocs.

d. "mouvement de caméra" : si $R_{mouv} > 0.5$. Dans cette situation, la plupart des blocs de pixels comportent un déplacement, donc on se trouve dans la situation d'un mouvement global de caméra.

e. "autre situation de mouvement" : si aucune des situations ci-dessus n'est réalisée.

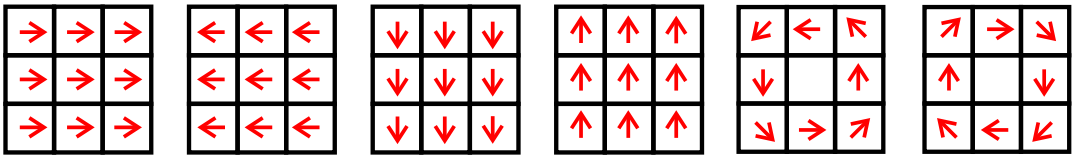
Les valeurs des seuils utilisés pour la classification du mouvement présentée ci-dessus ont été déterminées empiriquement en analysant plusieurs séquences d'images contenant des passages représentatifs de chaque situation particulière de mouvement.

Dans le cas d'un mouvement de caméra, la détermination du type de mouvement se fait par la comparaison des distributions des vecteurs moyens de mouvement (définies par $\bar{O}(k,l)$ et $\bar{A}(k,l)$) aux situations de référence présentées en Figure 3.8 et 3.9.



(a) "zoom in" (b) "zoom out"

FIG. 3.8 – Orientation des vecteurs moyens pour : (a) "zoom in" (3 situations possibles) et (b) "zoom out" (3 situations possibles).



(a) (b) (c) (d) (e) (f)

FIG. 3.9 – Orientation des vecteurs moyens pour : (a) "déplacement vers la gauche", (b) "déplacement vers la droite", (c) "déplacement vers le haut", (d) "déplacement vers le bas", (e) "rotation horaire", (f) "rotation anti-horaire".

Pour vérifier l'appartenance à l'une des situations présentées ci-dessus, nous analysons la matrice $\bar{O}(k,l)$ où $k,l \in \{0,1,2\}$. Seuls les vecteurs qui ont une amplitude supérieure à un pixel sont pris en compte, c'est-à-dire $\bar{A}(k,l) > 0$. Les angles servant à définir l'orientation du mouvement sont comparés à ceux qui sont définis ci-dessus, en utilisant un intervalle

de tolérance. Cet intervalle est égal à $d = \pm 45$ degrés pour les mouvements de translation (horizontale et verticale) et les situations 2 et 3 du mouvement "zoom in/out" (voir Figure 3.8) et $d = \pm 22$ degrés pour les mouvements de rotation et les premières situations du mouvement "zoom in/out" (voir Figure 3.8). Si la direction estimée se trouve à l'intérieur de la plage angulaire de tolérance pour une certaine direction de déplacement, elle sera approximée par cette direction. Le principe est présenté par la Figure 3.10.a.

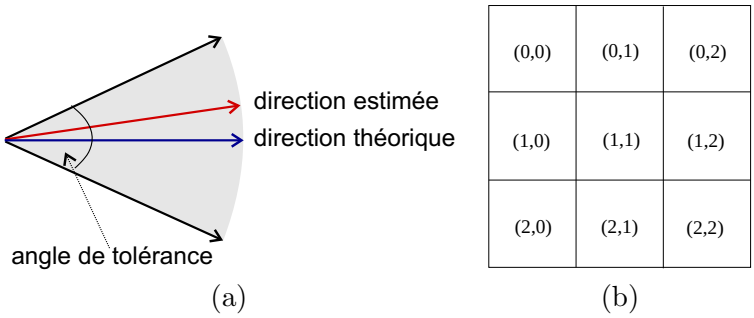


FIG. 3.10 – (a) La tolérance des valeurs des angles (le vecteur rouge est la direction estimée, le vecteur bleu est la direction théorique idéale, l'angle noir est la région de tolérance), (b) Les index utilisés pour les vecteurs moyens de mouvement.

En utilisant les notations de la Figure 3.10.b nous avons défini, en adoptant une stratégie de décision majoritaire, les règles de classification suivantes (dans ces règles les flèches indiquent l'orientation et l'opérateur $Card\{\}$ indique le nombre d'éléments d'un ensemble) :

d.1. "zoom in" : plusieurs situations sont possibles :

– *Situation 1* (voir la première image de la Figure 3.8.a) dans cette situation l'orientation du macro-bloc central n'est pas pris en compte :

$$\begin{aligned} Card\{\bar{O}(0,0) = \nwarrow; \bar{O}(0,1) = \uparrow; \bar{O}(0,2) = \nearrow; \bar{O}(1,0) = \leftarrow; \\ \bar{O}(1,2) = \rightarrow; \bar{O}(2,0) = \swarrow; \bar{O}(2,1) = \downarrow; \bar{O}(2,2) = \searrow \} \geq 7 \end{aligned} \tag{3.9}$$

– *Situation 2* (voir la deuxième image de la Figure 3.8.a) dans cette situation les orientations des macro-blocs situés sur la colonne centrale ne sont pas prises en compte

$$\begin{aligned} Card\{\bar{O}(i,0) = \leftarrow; \bar{O}(i,2) = \rightarrow / i = 0,..,2\} \geq 4 \\ \textbf{et } Card\{\bar{O}(i,1) = \leftarrow / i = 0,..,2\} < 2 \\ \textbf{et } Card\{\bar{O}(i,1) = \rightarrow / i = 0,..,2\} < 2 \end{aligned} \tag{3.10}$$

– *Situation 3* (voir la troisième image de la Figure 3.8.a) dans cette situation les orientations des macro-blocs situés sur la ligne centrale ne sont pas prises en compte

$$\begin{aligned} Card\{\bar{O}(0,j) = \uparrow, \bar{O}(2,j) = \downarrow / j = 0,..,2\} \geq 4 \\ \textbf{et } Card\{\bar{O}(1,j) = \uparrow / j = 0,..,2\} < 2 \\ \textbf{et } Card\{\bar{O}(1,j) = \downarrow / j = 0,..,2\} < 2 \end{aligned} \tag{3.11}$$

Dans la deuxième et troisième situation de "zoom in" l'orientation des vecteurs moyens est similaire à l'orientation des vecteurs moyens correspondant à un mouvement de translation, ce qui justifie l'ajout des deux dernières règles.

d.2. "zoom out" : similaire au "zoom in", les conditions sont :
- *Situation 1* (voir la première image de la Figure 3.8.b)

$$Card\{\bar{O}(0,0) = \searrow; \bar{O}(0,1) = \downarrow; \bar{O}(0,2) = \swarrow; \bar{O}(1,0) = \rightarrow; \\ \bar{O}(1,2) = \leftarrow, \bar{O}(2,0) = \nearrow, \bar{O}(2,1) = \uparrow, \bar{O}(2,2) = \nwarrow \} \geq 7 \quad (3.12)$$

- *Situation 2* (voir la deuxième image de la Figure 3.8.b)

$$Card\{\bar{O}(i,0) = \rightarrow; \bar{O}(i,2) = \leftarrow / i = 0,..,2\} \geq 4 \\ \textbf{et } Card\{\bar{O}(i,1) = \rightarrow / i = 0,..,2\} < 2 \\ \textbf{et } Card\{\bar{O}(i,1) = \leftarrow / i = 0,..,2\} < 2 \quad (3.13)$$

- *Situation 3* (voir la troisième image de la Figure 3.8.b)

$$Card\{\bar{O}(0,j) = \downarrow; \bar{O}(2,j) = \uparrow / j = 0,..,2\} \geq 4 \\ \textbf{et } Card\{\bar{O}(1,j) = \downarrow / j = 0,..,2\} < 2 \\ \textbf{et } Card\{\bar{O}(1,j) = \uparrow / j = 0,..,2\} < 2 \quad (3.14)$$

d.3. "déplacement vers la gauche" : (voir Figure 3.9.a)

$$Card\{\bar{O}(i,j) = \rightarrow / i,j = 0..2\} \geq 6 \quad (3.15)$$

d.4. "déplacement vers la droite" : (voir Figure 3.9.b)

$$Card\{\bar{O}(i,j) = \leftarrow / i,j = 0..2\} \geq 6 \quad (3.16)$$

d.5. "déplacement vers le haut" : (voir Figure 3.9.c)

$$Card\{\bar{O}(i,j) = \downarrow / i,j = 0..2\} \geq 6 \quad (3.17)$$

d.6. "déplacement vers le bas" : (voir Figure 3.9.d)

$$Card\{\bar{O}(i,j) = \uparrow / i,j = 0..2\} \geq 6 \quad (3.18)$$

d.7. "rotation dans le sens horaire" : (voir Figure 3.9.e)

$$Card\{\bar{O}(0,0) = \swarrow; \bar{O}(0,1) = \leftarrow; \bar{O}(0,2) = \nwarrow; \bar{O}(1,0) = \downarrow; \\ \bar{O}(1,2) = \uparrow; \bar{O}(2,0) = \searrow; \bar{O}(2,1) = \rightarrow; \bar{O}(2,2) = \nearrow \} \geq 6 \quad (3.19)$$

d.8. "rotation dans le sens anti-horaire" : (voir Figure 3.9.f)

$$Card\{\bar{O}(0,0) = \nearrow; \bar{O}(0,1) = \rightarrow; \bar{O}(0,2) = \searrow; \bar{O}(1,0) = \uparrow; \\ \bar{O}(1,2) = \downarrow; \bar{O}(2,0) = \nwarrow; \bar{O}(2,1) = \leftarrow; \bar{O}(2,2) = \swarrow \} \geq 6 \quad (3.20)$$

d.9. "autre mouvement de caméra" : si aucune des situations énumérées ci-dessus n'est valide.

La fusion des résultats - les plans de mouvement

Dès que la classification du mouvement sur la séquence entière est achevée, chaque image est caractérisée par un label approprié au type de mouvement. Par exemple : $image_n$="pas de mouvement", $image_{n+1}$="discontinuité du mouvement", etc. Les *plans de mouvement* sont obtenus en regroupant les mouvements similaires par plan. Ainsi, si pour les images aux instants $0, ..., n$ nous avons trouvé l'absence de mouvement, donc le label "pas de mouvement", le plan de mouvement associé sera défini par $plan_{[0-n]}$ ="pas de mouvement".

Dans de nombreuses situations, les plans de mouvement sont artificiellement fractionnés par de fausses détections de courte durée : à l'intérieur d'un plan on trouvera d'autres types de mouvement. Pour corriger ces résultats nous proposons une étape de fusion dont le principe est exposé ci-dessous.

Pour le plan de mouvement courant analysé, $plan^i_{[a,b]}$, d'indice i, situé entre les instants a et b et de label de mouvement M, on cherche, en suivant la chronologie du film, si le même type de mouvement est présent dans des plans voisins dans un intervalle de temps limité, noté T_r (la valeur de T_r a été fixée empiriquement à 5 images, car les fausses détections ont typiquement une durée inférieure à 5 images). Si le même label de mouvement, M, est rencontré dans le $plan^j_{[c,d]}$ $(j > i)$ à l'intérieur de l'intervalle T_r, tous les plans compris entre le plan d'indice i et le plan d'indice j sont concaténés en un seul plan de label de mouvement M qui devient le plan courant.

L'algorithme est ainsi appliqué à toute la séquence. Dans le cas particulier du mouvement de type "discontinuité du mouvement" aucune fusion n'est réalisée. Un exemple de fusion est présenté dans le Tableau 3.1.

plans de mouvement	*plans de mouvement après la fusion*
$plan^1_{[0,20]}$="pas de mouvement"	$plan^1_{[0,100]}$="pas de mouvement"
$plan^2_{[20,22]}$="mouvement d'objets"	$plan^2_{[100,101]}$="discontinuité du mouv."
$plan^3_{[22,100]}$="pas de mouvement"	$plan^3_{[101,200]}$="zoom in"
$plan^4_{[100,101]}$="discontinuité du mouv."	
$plan^5_{[101,130]}$="zoom in"	
$plan^6_{[130,133]}$="déplacement vers la gauche"	
$plan^7_{[133,134]}$="mouvement d'objets"	
$plan^8_{[134,200]}$="zoom in"	
...	

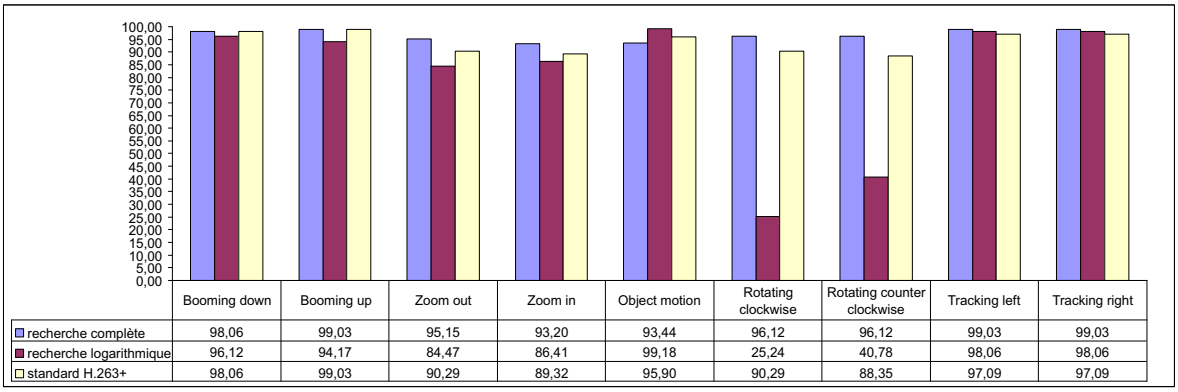
TAB. 3.1 – Exemple de fusion des plans de mouvement.

3.2.3 Résultats expérimentaux

Pour valider la méthode proposée concernant la classification des différents types de mouvements de caméra, nous avons utilisé un certain nombre de séquences de synthèse générées à partir d'un logiciel de graphique 3D ("3D Studio Max"). 26 séquences synthétiques ont été testées (24 pour le mouvement de caméra et 2 pour le mouvement d'objet). Pour un même mouvement, plusieurs séquences ont été générées en changeant les caractéristiques de ce mouvement (variation de la vitesse de translation, variation sur la nature de l'objet en mouvement, etc.).

Pour l'estimation du mouvement nous avons utilisé trois algorithmes de recherche : *la recherche complète*, *la recherche logarithmique* et *la recherche en trois étapes*. Un comparatif des ces trois méthodes est présenté dans l'Annexe C. Les images utilisées sont de taille 720×486 pixels et pour les tests nous avons utilisé un seuil de discontinuité noté $\tau_{discont} = 20000$.

Les résultats ont été évalués en utilisant comme mesure le taux de bonne détection qui est défini comme le rapport entre le nombre d'images où le mouvement a été bien classifié et le nombre total d'images utilisées. Les résultats obtenus sont présentés dans la Figure 3.11.



	Booming down	Booming up	Zoom out	Zoom in	Object motion	Rotating clockwise	Rotating counter clockwise	Tracking left	Tracking right
recherche complète	98,06	99,03	95,15	93,20	93,44	96,12	96,12	99,03	99,03
recherche logarithmique	96,12	94,17	84,47	86,41	99,18	25,24	40,78	98,06	98,06
standard H.263+	98,06	99,03	90,29	89,32	95,90	90,29	88,35	97,09	97,09

FIG. 3.11 – Les résultats de la classification du mouvement de caméra pour trois méthodes de recherche (sans correction des plans de mouvement).

Nous avons obtenu des taux de détection supérieurs à 93% pour la plupart des situations de mouvement.

	b.d.	b.u.	z.o.	z.i.	o.m.	r.cw.	r.ccw.	t.l.	t.r.	n.m.	oth.
	3seq.	3seq.	3seq.	3seq.	2seq.	3seq.	3seq.	3seq.	3seq.		
b.d.	101	0	0	0	2	0	0	0	0	0	0
b.u.	0	102	0	0	1	0	0	0	0	0	0
z.o.	0	0	98	0	5	0	0	0	0	0	0
z.i.	0	0	0	96	6	0	0	0	0	1	0
o.m.	0	0	0	0	114	0	0	0	0	8	0
r.cw.	0	0	0	0	4	99	0	0	0	0	0
r.ccw.	0	0	0	0	4	0	99	0	0	0	0
t.l.	0	0	0	0	1	0	0	102	0	0	0
t.r.	0	0	0	0	1	0	0	0	102	0	0

TAB. 3.2 – La matrice de confusion pour la classification du mouvement (*recherche complète* sans la correction des résultats par les plans de mouvement).

Les fausses détections sont généralement liées à la non-linéarité du mouvement : il y a des passages avec un mouvement trop faible qui sont détectés comme des passages sans mouvement ou des situations pour lesquelles les vecteurs de mouvement sont nuls à cause de la similarité des textures de l'image. Les matrices de confusion sont présentées dans les Tableaux 3.2, 3.3 et 3.4). Les notations suivantes ont été adoptées : *b.d./b.u.*=déplacement vers le bas/haut ("booming down/up"), *z.o./z.i.*=zoom out/in, *o.m.*=mouvement d'objets, *r.cw/r.ccw.*= ro-

tation horaire/anti-horaire ("rotation clockwise/counterclockwise"), *t.l./t.r.*=déplacement vers la gauche/droite ("tracking left/right"), *n.m.*=pas de mouvement, *oth.*=autre mouvement.

	b.d.	b.u.	z.o.	z.i.	o.m.	r.cw.	r.ccw.	t.l.	t.r.	n.m.	oth.
	3seq.	3seq.	3seq.	3seq.	2seq.	3seq.	3seq.	3seq.	3seq.		
b.d.	99	0	0	0	4	0	0	0	0	0	0
b.u.	0	97	0	0	2	0	0	0	2	0	2
z.o.	0	0	87	0	11	0	0	0	0	1	4
z.i.	0	0	0	89	11	0	0	0	0	1	2
o.m.	0	0	0	0	121	0	0	0	0	1	0
r.cw.	0	0	10	0	6	26	0	0	0	0	61
r.ccw.	0	0	12	0	9	0	42	0	0	0	40
t.l.	0	0	0	0	2	0	0	101	0	0	0
t.r.	0	0	0	0	2	0	0	0	101	0	0

TAB. 3.3 – La matrice de confusion pour la classification du mouvement (*recherche logarithmique* sans la correction des résultats par les plans de mouvement).

	b.d.	b.u.	z.o.	z.i.	o.m.	r.cw.	r.ccw.	t.l.	t.r.	n.m.	oth.
	3seq.	3seq.	3seq.	3seq.	2seq.	3seq.	3seq.	3seq.	3seq.		
b.d.	100	0	0	0	2	0	0	0	0	1	0
b.u.	0	102	0	0	1	0	0	0	0	0	0
z.o.	0	0	93	0	9	0	0	0	0	1	0
z.i.	0	0	0	92	10	0	0	0	0	1	0
o.m.	0	0	0	0	117	0	0	0	0	5	0
r.cw.	0	0	0	0	5	93	0	0	0	1	4
r.ccw.	0	0	0	0	5	0	92	0	0	1	5
t.l.	0	0	0	0	2	0	0	100	0	1	0
t.r.	0	0	0	0	2	0	0	0	100	1	0

TAB. 3.4 – La matrice de confusion pour la classification du mouvement (l'algorithme de recherche du *standard H.263+* sans la correction des résultats par les plans de mouvement).

Parmi les trois méthodes de recherche utilisées, celle donnant les meilleurs résultats est *la recherche complète*, mais au détriment de la vitesse de calcul, car c'est la méthode la plus lente. L'algorithme de recherche du *standard H.263+* donne des résultats inférieurs à la *la recherche complète*, mais il possède une complexité de calcul plus réduite, ce qui le rend rapide. Cependant, la méthode la plus rapide est la *recherche logarithmique* mais la précision de détection n'est pas toujours bonne (voir l'Annexe C). En particulier, dans les cas de rotations, la *recherche logarithmique* a échoué en obtenant des taux de bonnes détections très faibles (voir Figure 3.11).

3.2.4 Application : la détection des "cuts"

La détection du mouvement peut être également utilisée pour la détection *des changements de plans*. Un changement brusque, ou "cut", sera représenté par un plan de mouvement de durée égale à une image et de type "discontinuité du mouvement". De la même façon,

une transition vidéo sera également représentée par un mouvement discontinu, mais sur une durée supérieure à une image.

Cette technique de détection des "cuts" à partir de l'estimation du mouvement de caméra (appelée *mdiscont* dans la suite) a été testée sur deux longs métrages : "A Bug's Life" (84min46s) contenant 1597 "cuts" et "Toy Story" (73min18s) contenant 1569 "cuts", soit un temps total de 158min8s et un total de 3166 "cuts". Pour bien choisir le seuil de discontinuité, $\tau_{discont}$ (utilisé pour l'estimation du mouvement, voir la Section 3.2.2) nous avons testé la détection des "cuts" sur la première séquence en utilisant plusieurs valeurs pour $\tau_{discont}$. Les résultats sont présentés dans le Tableau 3.5. Les erreurs de détection sont exprimées en utilisant les critères de précision et de rappel (voir la Section 2.3).

$\tau_{discont}$	N_t	BD	FD	Précision	Rappel
10000	1597	1492	134	**91.76%**	**93.43%**
15000	1597	1331	98	93.14%	83.34%
20000	1597	1074	69	93.96%	67.25%

TAB. 3.5 – Les résultats de détection des "cuts" pour différentes valeurs du seuil $\tau_{discont}$ (séquence "A Bug's Life", N_t est le nombre total de "cuts", BD et FD sont les nombres de bonnes détections et de fausses détections).

En augmentant la valeur de $\tau_{discont}$ la détection est moins sensible aux discontinuités du mouvement. Après avoir analysé les résultats nous avons choisi d'utiliser la valeur $\tau_{discont} = 10000$ pour laquelle nous avons obtenu le meilleur compromis précision-rappel. En utilisant cette valeur nous avons testé la détection des "cuts" pour les deux films long métrage. Les résultats sont présentés dans le Tableau 3.6.

film	N_t	BD	FD	Précision	Rappel
"A Bug's Life"	1597	1492	134	91.76%	93.43%
"Toy story"	1569	1501	221	87.17%	95.66%

TAB. 3.6 – Résultats de la détection des "cuts" en utilisant le mouvement ($\tau_{discont} = 10000$, N_t est le nombre total de "cuts", BD et FD sont les nombres de bonnes détections et de fausses détections).

Un test comparatif avec deux autres méthodes de détection des "cuts" est présenté dans la Section 2.4.4, Tableau 2.5. La méthode *mdiscont* a obtenu une moins bonne précision et donc un nombre plus élevé de fausses détections que les deux autres méthodes à cause des mouvements très rapides de caméra (mouvements qui sont souvent présents dans les films d'animation). D'un autre côté, elle a obtenu le meilleur rappel et donc le nombre le plus élevé de bonnes détections.

L'utilisation des plans de discontinuité de mouvement pour la détection des transitions vidéo graduelles (comme par exemple les "fades" ou les "dissolves") ne s'est pas révélée efficace. Les plans de discontinuité ne coïncident avec les transitions graduelles que dans une faible proportion. La plupart des transitions détectées en utilisant les plans de discontinuité du mouvement correspondent en fait à des mouvements très rapides de caméra ou à des effets de couleurs.

3.2.5 Conclusions

Dans cette section nous avons proposé une méthode d'analyse du mouvement de caméra. Elle est basée sur la définition d'un certain nombre de modèles de mouvement correspondant aux différents types de mouvements de caméra analysés. La classification est effectuée en utilisant un jeu de règles basé sur le choix d'un certain nombre de seuils. Nous avons également utilisé cette approche pour la détection des "cuts" où nous avons obtenu de bons résultats. La méthode d'analyse proposée se limite seulement à l'analyse d'un certain nombre de mouvements de base de caméra : les mouvements de translation sur les axes oX et oY, de rotation et les "zoom in/out". Dans la réalité, les mouvements de caméra sont plus complexes ce qui rend difficile l'évaluation des résultats.

Souvent, plusieurs types de mouvements sont utilisés conjointement : par exemple un déplacement à droite mélangé avec un agrandissement ("zoom in"), mouvement qui peut être considéré aussi bien comme un "zoom" que comme une translation. La pertinence de l'évaluation est bien sûr liée à la construction de la vérité terrain. Pour les mouvements 3D de caméra, même si on les décompose en mouvements classiques, la vérité terrain est très lourde à construire et peut être inexacte. De plus certains mouvements de caméra ont des apparences similaires (par exemple le "panning" et le déplacement vers le haut, voir Figure 3.3). C'est pour cette raison que nous avons testé notre méthode sur des séquences artificielles qui ont été générées pour chaque type de mouvement recherché.

A partir des informations de mouvement, nous avons envisagé de détecter les transitions vidéo. Si les résultats sont bons pour la détection des "cuts", les transitions lentes sont mal détectées. Une solution pour améliorer ce point serait d'augmenter le pas temporel d'analyse.

3.3 Paramètres de bas niveau du mouvement

L'information sur les différentes situations de mouvement présentes dans la séquence peut servir pour l'analyse sémantique du contenu. Comme pour les transitions vidéo, les mouvements de caméra ne sont pas utilisés de manière aléatoire, mais avec un but précis. Par exemple les *mouvements de translation* changent le point de vue de la scène, un *mouvement d'agrandissement* ("zoom in") est plutôt utilisé pour se focaliser sur une région de l'image ou sur un personnage, *l'absence de mouvement* définit un moment de calme dans l'action du film, etc.

En utilisant les plans de mouvement obtenus avec la méthode proposée nous avons défini un certain nombre de paramètres de bas niveau pour caractériser le mouvement global de la séquence. Les deux premiers paramètres proposés estiment le pourcentage de passages de la séquence sans mouvement, R_{pm}, et le pourcentage de passages dans lesquels on rencontre des mouvements d'objets, R_{mo}. Ils sont définis par :

$$R_{pm} = \frac{N_{pm}}{N_{img}}, \quad R_{mo} = \frac{N_{mo}}{N_{img}} \tag{3.21}$$

où N_{pm} est le nombre total d'images sans mouvement (donc ayant le label "pas de mouvement"), N_{mo} est le nombre total d'images comportant un mouvement d'objets (label "mouvement d'objets") et N_{img} est le nombre total d'images de la séquence.

Le pourcentage de passages du film avec un mouvement de translation est calculé par :

$$R_{m.trans} = \frac{N_{depl.d} + N_{depl.g} + N_{depl.h} + N_{depl.b}}{N_{img}} \tag{3.22}$$

où $N_{depl.x}$ est le nombre total d'images qui comportent un mouvement de type déplacement vers x, avec $x \in \{d, g, h, b\}$ où $d = droite$, $g = gauche$, $h = haut$ et $b = bas$.

Les pourcentages de mouvements de rotation, R_{rot}, et de mouvement "zoom", R_{zoom}, sont définis par :

$$R_{rot} = \frac{N_{rot.h} + N_{rot.ah}}{N_{img}}, \quad R_{zoom} = \frac{N_{zi} + N_{zo}}{N_{img}} \tag{3.23}$$

où $N_{rot.x}$ est le nombre total d'images qui comportent un mouvement de rotation dans le sens x, avec $x \in \{h, ah\}$ où $h = horaire$ et $ah = anti - horaire$, et N_{zi}, N_{zo} sont respectivement les nombres d'images qui comportent un mouvement de type "zoom in" (agrandissement) et "zoom out" (rétrécissement).

A partir des paramètres ainsi calculés, nous avons envisagé une description sémantique du mouvement contenu dans la séquence (voir la Section 7.4).

3.4 Conclusions générales

Dans ce chapitre nous avons présenté les différentes méthodes d'analyse du mouvement dans les séquences d'images. Nous avons également proposé une méthode de classification d'un certain nombre de mouvements de base de caméra. Généralement, toutes les méthodes existantes ont besoin pour l'analyse de *l'estimation du champ vectoriel du mouvement*.

Selon la précision recherchée, les méthodes d'estimation du mouvement se divisent en deux catégories. Il y a d'abord les méthodes qui calculent un champ vectoriel dense, où un vecteur de déplacement est calculé en chaque pixel. Ce type de méthodes est typiquement basé sur l'estimation du flot optique et la complexité de calcul est généralement élevée. D'autre part il y a les méthodes par blocs qui estiment le déplacement sur des blocs de pixels de l'image. Ces méthodes s'avèrent être le meilleur compromis entre la complexité de calcul et la qualité du champ de mouvement obtenu. C'est pour cela qu'elles sont employées dans de codage vidéo (par exemple dans les standards MPEG).

Nous pouvons dire que la qualité de l'analyse du mouvement dans une séquence d'images est directement proportionnelle à la qualité de l'estimation de mouvement utilisée. De plus, la méthode d'estimation doit également être choisie en fonction de l'application envisagée. Par exemple, pour la segmentation d'un objet en mouvement, il sera préférable d'utiliser un champ vectoriel dense obtenu par l'estimation du flot optique plutôt que d'utiliser des informations sur les déplacements des blocs de pixels. En revanche, dans une application temps réel on préférera une méthode plus rapide que l'estimation du flot optique.

La méthode de classification de mouvement de caméra proposée a donné de bons résultats sur l'ensemble des séquences testées. L'efficacité de la méthode est liée à l'utilisation de modèles de mouvement définis a priori. L'estimation du mouvement est réalisée à l'aide d'une méthode basée sur les blocs de pixels. Sa précision d'estimation répond à nos besoins : avoir un champ de mouvement moins dense mais une vitesse de calcul élevée. Pour la tâche de classification nous avons utilisé des règles de décision simples. Par rapport aux autres méthodes de classification du mouvement de caméra, notre méthode prend en compte deux situations particulières : la discontinuité et le mouvement d'objets. En utilisant un algorithme d'estimation du mouvement par blocs optimisé (comme par exemple la recherche logarithmique, la recherche en trois étapes, etc.), une implantation en temps réel de l'ensemble de la méthode est possible. Enfin, en même temps que la classification, l'analyse du mouvement permet une détection des changements de plans.

Chapitre 4

L'Analyse des couleurs

Résumé : *Les couleurs jouent un rôle très important dans la transmission d'informations visuelles. Dans ce chapitre nous présentons différentes techniques d'analyse des couleurs à l'intérieur d'une image et à travers des séquences d'images. Nous proposons ensuite une méthode de caractérisation de la distribution globale des couleurs dans une séquence d'images, technique appliquée aux films d'animation. La méthode proposée s'appuie sur le fait que chaque film d'animation a le plus souvent sa propre distribution des couleurs, contrairement aux films conventionnels qui généralement utilisent la plupart des couleurs existantes.*

Les couleurs jouent un rôle très important dans la transmission d'informations visuelles. L'œil humain est ainsi plus sensible aux changements de teinte des couleurs qu'à la présence de mouvement.

Les films d'animation constituent un type particulier d'expression artistique. Et dans ce mode d'expression, l'une des caractéristiques est le fait que chaque film a sa propre distribution des couleurs (voir Chapitre 1.5, Figure 1.7), contrairement aux films conventionnels qui généralement utilisent la plupart des couleurs existant dans la nature. En animation, avant de créer son oeuvre, l'artiste, choisit les couleurs qu'il va utiliser en concordance avec son projet artistique. Trouver les couleurs prédominantes utilisées dans la séquence est donc un élément essentiel à la caractérisation du point de vue artistique de la séquence : procédés couleurs utilisés, combinaisons des couleurs, impressions transmises, etc.

Dans ce chapitre nous proposons une description statistique globale des couleurs prédominantes présentes dans une séquence d'images. La méthode proposée est adaptée au cas particulier des films d'animation. A partir d'un résumé de la séquence, la distribution des couleurs est caractérisée par un *histogramme global pondéré des couleurs*, histogramme prenant en compte l'importance de chaque plan vidéo. Dans cet histogramme chaque couleur qui a une fréquence d'apparition suffisamment importante dans la séquence est caractérisée par son pourcentage d'apparition.

L'histogramme proposé servira de point de départ pour le calcul d'un certain nombre de descripteurs symboliques des couleurs dans l'analyse sémantique des séquences d'images (voir le Chapitre 7).

4.1 État de l'art

Dans le domaine de l'indexation des séquences d'images il n'y a pas beaucoup de travaux consacrés à la caractérisation de la distribution des couleurs d'une séquence. La couleur a été plutôt exploitée dans les systèmes d'indexation d'images statiques. Un état de l'art est présenté dans [Bimbo 99] et [Smeulders 00]. L'information couleur, en conjonction avec d'autres informations telles que le mouvement, la structure de la séquence, peut être utilisée pour définir des descripteurs du contenu et servir dans un but d'indexation (voir Chapitre 1.2).

4.1.1 Les espaces des couleurs

Dans une image numérique, la couleur d'un pixel est représentée comme un point dans un espace, typiquement 3D, qui constitue l'espace des couleurs. La littérature du domaine propose une grande variété d'espaces couleurs qui ont été créés pour des besoins d'analyse. Par exemple, certaines propriétés des couleurs ne sont pas pas directement accessibles dans l'espace initial des images (souvent l'espace RVB). Ces différents espaces couleurs ont été abondamment étudiés depuis une vingtaine d'années [Trémeau 04], et nous nous limiterons ici à quelques notions essentielles dont nous aurons besoin par la suite :

- **les espaces physiques** : ils sont liés aux conditions technologiques d'acquisition et de restitution de la couleur. Dans cette catégorie on citera essentiellement l'espace *RVB* (R-rouge, V-vert, B-bleu), qui est le plus souvent l'espace initial des images numériques. A partir de l'espace RVB, quelques autres espaces dérivés ont été proposés, comme l'espace XYZ, où X, Y, Z sont des primaires imaginaires. L'intérêt de cet espace est que les triplets décrivant les couleurs ont toujours des valeurs positives.
- **les espaces perceptuels** : ils sont basés sur la perception des couleurs. Dans cet catégorie on peut évoquer les modèles des couleurs opposées, comme par exemple l'espace *CIELab*, où L est la luminance, a est une composante chromatique sur l'axe vert/magenta et b est une seconde composante chromatique sur l'axe bleu/jaune. Dans cet espace dit "perceptuellement uniforme", les distances entre composantes sont proches des écarts perçus par l'œil humain. On peut également mentionner la famille des espaces couleurs utilisant la Teinte (T), la Saturation (S) et la Luminance (L) qui expriment les composantes dans des termes proches de la manière dont l'être humain exprime sa perception de la couleur.

4.1.2 La caractérisation des couleurs dans l'image

Il y a plusieurs manières de caractériser l'information couleur d'une image. Une première méthode consiste à utiliser *des informations de bas niveau*, comme par exemple des mesures statistiques sur la distribution des couleurs dans l'image. Le concept de couleur étant lié à la perception, cette méthode est souvent insuffisante pour décrire les différentes techniques d'utilisation des couleurs dans l'image.

Un niveau sémantique supérieur est acquis si *les noms de couleurs* sont utilisés. Associer des noms aux couleurs permet alors de se créer une représentation des couleurs utilisées. Conjointement aux analyses statistiques, la caractérisation des couleurs demande aussi une analyse des couleurs utilisées, et en particulier l'analyse des différentes *relations existant entre les couleurs* : adjacence, complémentarité, etc. Notre approche va s'appuyer sur ce qui se fait

souvent dans le domaine artistique où les couleurs sont choisies et mélangées en partant d'une roue des couleurs. Cette roue est un ensemble arbitraire de couleurs élémentaires disposées de manière à donner une perception progressive.

Dans la suite nous allons détailler les différents concepts énumérés ci-dessus.

Les histogrammes couleurs

Dans une image, la méthode la plus fréquemment utilisée pour caractériser la distribution des couleurs est le calcul de *l'histogramme couleur*, pour l'image entière ou pour certaines régions d'intérêt.

Le calcul de l'histogramme est un moyen de représentation efficace du contenu couleur d'une image. L'histogramme en tant que mesure statistique est en particulier invariant à un certain nombre de transformations géométriques de l'image (rotation, changement de résolution, etc.). Par contre, il est dépendant des variations de l'intensité lumineuse dans l'image et pose également un certain nombre de problèmes techniques liés aux nombres de couleurs comme nous le verrons dans la Section 2.4.2. Pour améliorer son invariance une technique souvent utilisée est le calcul de l'histogramme dans un espace couleur permettant la séparation de l'information lumineuse de l'information chromatique (comme par exemple les espaces TSL ou YCbCr).

Les noms des couleurs

Une autre façon de caractériser l'information de couleur est d'utiliser les *noms des couleurs. En associant des noms aux couleurs on arrive à se fabriquer une image visuelle de la couleur dont on parle.* Les noms des couleurs sont choisis dans un dictionnaire de noms qui ont été associés à chacune des couleurs disponibles au travers d'un système de dénomination des couleurs ("naming system"). Le problème d'association de noms aux couleurs a été bien analysé dans la littérature [Kay 03] [Benavente 04] [Lay 04].

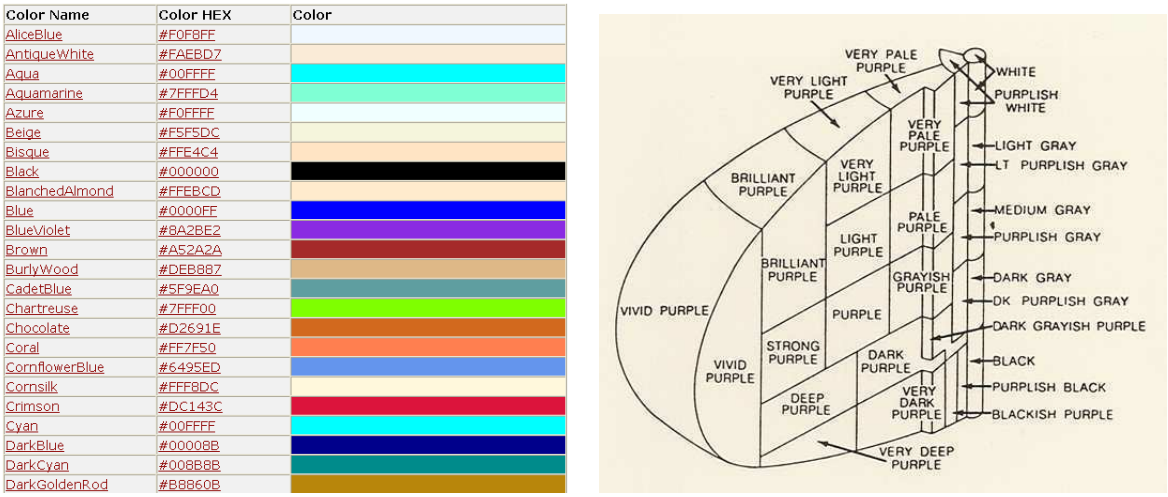
Les travaux proposés dans [Berlin 91] définissent des noms associés à des couleurs de base en imposant que ces noms possèdent les propriétés suivantes :

- ils ne doivent pas constituer une restriction à une certaine classe d'objets (exemple : la couleur olive n'est pas une couleur de base valide),
- leur sens ne doit pas se déduire de la compréhension d'objets (exemple : la couleur d'une feuille n'est pas une couleur de base valide),
- leur sens ne doit pas être inclus dans le nom d'une autre couleur,
- ils ont une constance vis-à-vis de la perception.

Ainsi, en utilisant comme point de départ l'espace de Munsell, [Berlin 91] définit 11 couleurs de base qui présentent un caractère suffisamment général et qui, de plus, se retrouvent dans 20 langues étrangères différentes. Ce sont : *Blanc*, *Noir*, *Rouge*, *Vert*, *Jaune*, *Bleu*, *Marron*, *Rose*, *Pourpre*, *Orange* et *Gris*. Ces couleurs de base constituent un point de départ pour définir des descriptions textuelles pour les autres couleurs. Si les noms des couleurs de base se trouvent être les mêmes dans différentes langues, les frontières entre ces différentes couleurs sont au contraire variables.

Les systèmes d'association de noms aux couleurs emploient différentes techniques pour atteindre une certaine universalité. Par exemple ils utilisent des fonctions floues pour modéliser l'appartenance de telle couleur à telle catégorie, ils associent les noms des couleurs à certains

intervalles de longueur d'onde, lien avec la représentation physique de la couleur, ils utilisent des tables de noms de couleurs ("lookup tables"), définies a priori et qui associent le nom de la couleur avec sa représentation dans un certain espace des couleurs (voir la Figure 4.1.a). Une autre approche est le partitionnement d'un certain espace des couleurs en fonction de la similarité entre les couleurs, etc. (voir Figure 4.1.b). Cependant, on peut constater que les méthodes existantes de dénomination des couleurs ne sont jamais totalement automatiques [Benavente 04].



Color Name	Color HEX	Color
AliceBlue	#F0F8FF	
AntiqueWhite	#FAEBD7	
Aqua	#00FFFF	
Aquamarine	#7FFFD4	
Azure	#F0FFFF	
Beige	#F5F5DC	
Bisque	#FFE4C4	
Black	#000000	
BlanchedAlmond	#FFEBCD	
Blue	#0000FF	
BlueViolet	#8A2BE2	
Brown	#A52A2A	
BurlyWood	#DEB887	
CadetBlue	#5F9EA0	
Chartreuse	#7FFF00	
Chocolate	#D2691E	
Coral	#FF7F50	
CornflowerBlue	#6495ED	
Cornsilk	#FFF8DC	
Crimson	#DC143C	
Cyan	#00FFFF	
DarkBlue	#00008B	
DarkCyan	#008B8B	
DarkGoldenRod	#B8860B	

(a) (b)

FIG. 4.1 – (a) Extrait du tableau des couleurs du langage HTML, (b) Le système des couleurs standardisé ISCC-NBS (définition des pourpres).

Sur ces principes, un certain nombre de dictionnaires des noms de couleurs ont ainsi été proposés. On peut mentionner :

- le système des couleurs ISCC-NBS [Kelly 76] ("Inter Society Color Council" - "National Bureau of Standards") qui est basé sur la sphère des couleurs de Munsell (voir la Figure 4.1.b),
- les dictionnaires de "X11 Window System Distribution", "Netscape Color Names", "HTML-4 Color Names", "Two4U's Big Color Database", "Resene Paint Colours" ou "CNS Color-Naming System" (voir [Dictionaries 06]).

Les sensations introduites

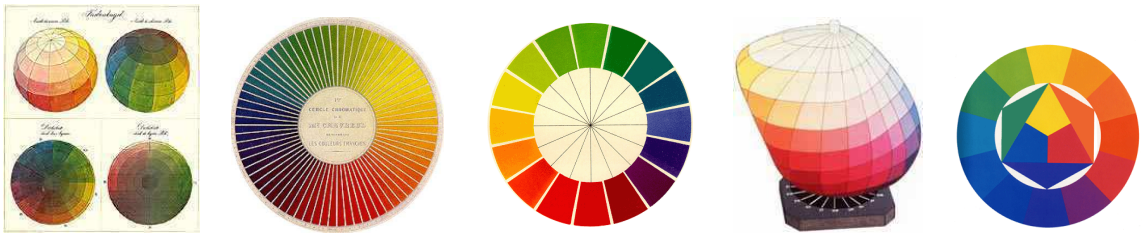
Une autre façon de caractériser la couleur dans une image est de s'intéresser à *la sensation transmise*. Itten [Itten 61], au travers de l'expérience acquise comme peintre dans le mouvement Bauhaus, définit en 1960 un ensemble de *règles formelles* (langage couleur) pour qualifier les effets visuels des différentes combinaisons des couleurs au niveau perceptuel. Ses travaux prennent leur source dans le domaine de l'art pictural.

Pour choisir les couleurs qu'ils vont utiliser dans leurs œuvres et pour analyser la perception produite, les artistes utilisent ce qu'on appelle une *roue des couleurs*. Une introduction sur le concept de roue des couleurs dans l'art est présenté dans [Birren 69].

Les roues des couleurs sont essentiellement des espaces de couleurs construits d'une

manière particulière, dans lesquels les relations entre les couleurs sont inspirées de la théorie des contrastes et de l'harmonie des couleurs. Une roue des couleurs est construite en utilisant un nombre arbitraire de couleurs élémentaires qui sont organisées d'une manière perceptuelle progressive [Lay 04]. De manière similaire, une *sphère des couleurs* est une représentation 3D de même nature.

Dans l'histoire des arts, plusieurs représentations des couleurs sous la forme de roues ou sphères des couleurs ont été proposées : Runge (1810), Chevreul (1864), Hering (1880), Munsell (1910), Itten (1960), etc (voir la Figure 4.2).



(a) (b) (c) (d) (e)

FIG. 4.2 – (a) Sphère des couleurs de Runge (1810), (b) Roue des couleurs de Chevreul (1864), (c) Couleurs opposées de Hering (1880), (d) Solide des couleurs de Munsell (1910), (e) Roue des couleurs d'Itten (1960).

En utilisant la théorie proposée par Itten et la représentation perceptuelle des couleurs à travers une roue des couleurs on peut caractériser la perception visuelle des couleurs en ce qui concerne le *contraste* et *l'accord* des couleurs.

L'étude de la disposition et des relations des couleurs sur une roue des couleurs est essentielle dans le domaine artistique (voir Josef Albers, Faber Birren, Johannes Itten, etc.). Itten [Itten 61] définit la perception des couleurs par les *7 contrastes* de couleurs (voir la Figure 4.3) :

- **le contraste des teintes** : ce contraste est obtenu par la juxtaposition de différentes teintes. Plus la distance entre les teintes utilisées est élevée, plus fort est le contraste (la distance entre couleurs est définie sur la roue des couleurs). Un exemple est proposé dans la figure 4.3.a.

- **le contraste clair-foncé** : il est lié à la perception de l'intensité lumineuse, évaluée sur les niveaux de gris ou sur les teintes. Le contraste est réalisé par la juxtaposition de couleurs claires et de couleurs foncées (voir la Figure 4.3.b).

- **le contraste chaud-froid** : correspond à la chaleur des couleurs. En art, on considère que les couleurs ont une certaine *température ou chaleur*. Le Jaune, le Jaune-Orange, l'Orange, le Rouge-Orange, le Rouge et le Rouge-Violet sont considérés comme des teintes *chaudes*. Au contraire le Jaune-Vert, le Vert, le Bleu-Vert, le Bleu, le Bleu-Violet et le Violet sont considérés comme des teintes *froides*. Le contraste est réalisé par la juxtaposition de couleurs chaudes et froides (voir la Figure 4.3.c).

- **le contraste de complémentarité** : il est lié aux relations de complémentarité entre les couleurs. En pratique sur une roue des couleurs (par exemple la roue d'Itten) les

paires de couleurs complémentaires sont disposées sur un diamètre de la roue. Le contraste de complémentarité est donc réalisé en utilisant des couleurs opposées afin de donner une symétrie dans la perception des teintes (voir la Figure 4.3.d).

– **le contraste de simultanéité** : il s'inspire du phénomène des couleurs opposées. Pour une couleur donnée notre œil exige simultanément la couleur complémentaire et la produit lui-même si elle ne lui est pas donnée. Puisque la couleur complémentaire engendrée simultanément n'existe pas réellement, mais qu'elle n'est engendrée que dans l'œil, elle éveille en nous une impression d'irritation et de vibration vivante dont la force change constamment. Cela peut arriver entre un gris et une couleur principale, entre deux couleurs pures et aussi entre gris. Des illusions optiques intéressantes peuvent être obtenues avec ce contraste. Un exemple est illustré dans la Figure 4.3.e où l'œil accepte mal la juxtaposition des couleurs rouges et vertes qui sont contrastées sans être complémentaires.

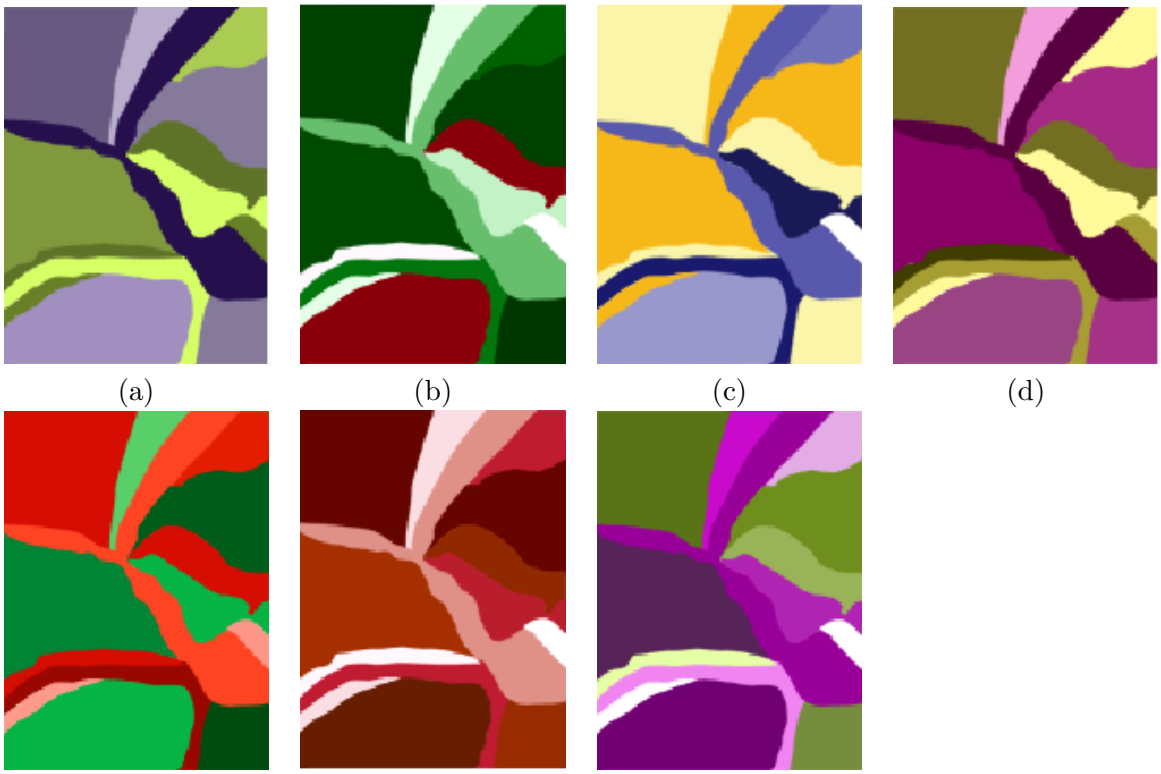


FIG. 4.3 – Les 7 contrastes d'Itten (a) Contraste des teintes, (b) Contraste clair-foncé, (c) Contraste chaud-froid, (d) Contraste de complémentarité, (e) Contraste de simultanéité, (f) Contraste de saturation, (g) Contraste d'extension (source "http ://www.worqx.com/color/itten.htm").

– **le contraste de saturation** : il est réalisé par la juxtaposition de teintes pures saturées avec des teintes diluées ayant une saturation faible. Ce contraste est relatif car une couleur peut apparaître plus saturée si elle est mise à côte d'une couleur pale et l'inverse. Un exemple est illustré dans la Figure 4.3.f.

– **le contraste d'extension** : il est lié à la proportion quantitative d'utilisation des couleurs. L'apparence d'une couleur est influencée par la dimension qualitative de sa luminance mais aussi par sa dimension quantitative d'utilisation (la surface spatiale occupée par la couleur dans l'image). Ce contraste est réalisé en associant à des couleurs des régions physiques dans l'image qui ont une surface proportionnelle au poids perceptuel visuel de la couleur (voir la Figure 4.3.g).

Dans le domaine de l'art, il existe un certain nombre de *principes d'utilisation des couleurs* qui ont des effets particuliers. Ces principes sont souvent décrits comme les moyens d'obtenir l'harmonie des couleurs dans l'image [Birren 69]. Ils expriment les relations entre les différents types de contrastes dans l'image. Ces principes d'harmonie des couleurs sont les suivants :

– **le principe monochromatique** : il est lié à l'harmonie d'une seule teinte. Il est réalisé en utilisant le principe du contraste clair-foncé, on joue alors sur l'intensité de la teinte, ou le principe du contraste de saturation, en faisant varier la saturation de la teinte pour éviter la monotonie.

– **le principe d'adjacence** : il est lié à l'harmonie des teintes similaires. Typiquement il est réalisé par le mélange d'au plus trois couleurs adjacentes sur une roue des couleurs contenant 12 couleurs élémentaires (par exemple la roue d'Itten). Une des couleurs sera utilisée dans une proportion plus élevée que les autres.

– **le principe de complémentarité** : il est lié à l'harmonie de l'équilibre des couleurs complémentaires. La façon le plus simple de le réaliser est l'utilisation de deux couleurs opposées. Une alternative est d'utiliser deux couleurs adjacentes et leur paires correspondantes opposées (donc une structure en X sur la roue des couleurs d'Itten) ce qu'on appelle *double complémentarité*. La *complémentarité divisée* est obtenue en associant une couleur avec deux couleurs voisines de teinte opposée, pour former une structure en Y sur la roue des couleurs d'Itten.

Parmi les systèmes d'indexation d'images, ils en existe quelques uns qui analysent la perception des couleurs d'un point de vue artistique. On peut citer QBIC de "State Hermitage Museum" [QBIC] et PICASSO [Corridoni 99]. Une approche intéressante est présentée dans [Lay 04] pour un système d'indexation des images d'art (SoloArt). Les différents concepts artistiques d'utilisation des couleurs y sont décrits en analysant les contrastes et les harmonies des couleurs.

4.1.3 Caractérisation des couleurs dans les séquences d'images

Bien que moins fréquemment, la couleur est également exploitée pour la caractérisation des séquences d'images. On peut ainsi mentionner les travaux proposés dans [Colombo 99] et [Detyniecki 03].

Dans [Colombo 99] la couleur est utilisée en conjonction avec d'autres paramètres de la séquence pour décrire d'une manière sémantique les séquences de publicité. Deux niveaux différents de perception sont analysés : le niveau expressif et le niveau émotionnel. Au niveau *expressif* les publicités sont divisées en quatre types : *pratique, animé, utopique et critique*. Pour classifier les séquences dans ces catégories les paramètres utilisés sont les suivants : la présence de "cuts" et de "dissolves", la présence ou l'absence récurrente de certaines couleurs, la présence ou l'absence d'effets de montage de la séquence, la présence de lignes verticales ou

horizontales et la présence de couleurs saturées ou non saturées. La caractérisation au *niveau émotionnel* est liée à l'action, au suspens, à la tranquillité, à la relaxation, au bonheur et à l'excitation de la séquence. La couleur est utilisée pour représenter certains niveaux émotionnels. Par exemple, le degré d'action d'une séquence de publicité peut être accentué par l'apport de Rouge et de Pourpre. De même, la tranquillité peut être marquée par la présence de Bleu, d'Orange, de Vert ou de Blanc et atténuée par la présence de Noir et de Pourpre.

Chaque caractérisation sémantique proposée est représentée par un jeu de paramètres de bas niveau qui sont associés aux séquences en utilisant un modèle de représentation floue. Par exemple le type *pratique* est caractérisé par l'absence de couleurs saturées ($\phi_{saturated} = 0$, où ϕ est le degré d'appartenance flou au symbole qui modélise la présence des couleurs saturées dans la séquence), la présence de lignes horizontales ou verticales ($\phi_{hor/vert} = 1$) et la présence de "dissolves" ($\phi_{dissolves} = 1$).

Dans [Detyniecki 03] les arbres de décision flous sont utilisés pour retrouver de la connaissance ("data mining") dans les séquences d'informations. Les arbres de décision flous permettent l'extraction automatique de règles pour la classification des différentes parties thématiques dans les séquences d'informations. La couleur est la seule information utilisée pour décrire ces séquences. Dans un premier temps un certain nombre d'images clés sont extraites pour chaque plan de la séquence. Les couleurs des images sélectionnées sont projetées sur une palette de couleurs réduite (contenant 64 ou 256 couleurs obtenues par quantification uniforme de l'espace RVB). Ensuite, avec cette palette, un histogramme est calculé pour chaque image clé. Ces histogrammes sont alors utilisés pour construire l'arbre de décisions flou. Dans ces séquences d'information, seule la couleur a été discriminante pour classifier deux types d'événements : la présence des incrustations ("inlays") et la présence du journaliste.

4.2 Méthode proposée

La méthode proposée pour caractériser les couleurs dans les films d'animation (voir [Ionescu 05h] ou le rapport [Ionescu 05a]) utilise une approche semblable à celle proposée dans [Detyniecki 03]. Elle est motivée par le fait que dans les films d'animation la couleur est une information discriminante : chaque film a sa propre palette de couleurs (voir Chapitre 1.5, Figure 1.7). La distribution globale des couleurs dans la séquence est représentée par le calcul d'un *histogramme couleur global pondéré*. Le diagramme de la méthode est illustrée dans la Figure 4.4.

Les étapes d'analyse sont les suivantes :

- **le découpage en plans** : dans un premier temps la séquence est divisée en plans vidéo par la détection des transitions vidéo,
- **le calcul du résumé** : un résumé de la séquence est calculé de manière automatique pour réduire la redondance temporelle. Dans ce résumé chaque plan vidéo de la séquence est représenté par un pourcentage de ses images,
- **la réduction des couleurs** : pour toutes les images retenues dans le résumé, les couleurs sont réduites en utilisant une palette particulière de couleurs associée à un mécanisme de diffusion d'erreur,
- **le calcul des histogrammes moyens** : pour chacun des plans, en n'utilisant que le résumé, et après réduction des couleurs, on calcule un histogramme couleur moyen,
- **le calcul de l'histogramme global pondéré**. l'histogramme global pondéré de

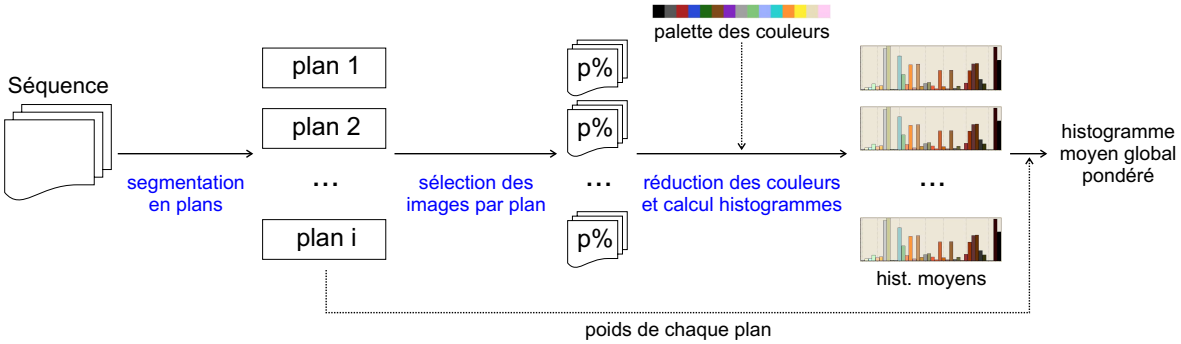


FIG. 4.4 – Le calcul de l'histogramme global pondéré.

la séquence est ensuite calculé comme la somme pondérée de tous les histogrammes moyens de chaque plan vidéo.

4.2.1 Le découpage en plans

Dans un premier temps, la séquence est *découpée en plans* par détection des transitions vidéo. En utilisant les méthodes proposées dans le Chapitre 2 on détecte les transitions suivantes : les "cuts", les "fades", les "dissolves" et également un effet de couleurs particulier spécifique aux films d'animation : les "changements brefs de couleurs", notés SCC dans la suite. Les plans vidéo sont déterminés par l'agrégation des transitions obtenues selon la méthode présentée dans la Section 2.7. Cette étape est nécessaire pour enlever les informations peu pertinentes du point de vue de la couleur comme par exemple les images de transitions, les images noires, les plans trop courts qui sont peu visibles, etc.

Pour chaque plan on calcule alors un indicateur ω_i, où i est l'indice du plan ($i = 1, ..., N_{plans}$, avec N_{plans} le nombre total de plans), qui est une mesure de l'importance du plan i dans la séquence entière. ω_i représente le poids du plan i et il est défini par :

$$\omega_i = \frac{N^i_{img}}{N_{film}} \tag{4.1}$$

où N^i_{img} est le nombre d'images du plan i et N_{film} est le nombre d'images de la séquence entière. Cet indicateur sera utilisé pour le calcul de la distribution globale des couleurs dans la séquence.

4.2.2 La construction du résumé

Pour réduire le contenu de la séquence et donc la redondance temporelle, un *résumé* de la séquence est construit d'une manière automatique. Chaque plan vidéo est résumé par un pourcentage ($p\%$) des images qui le composent. Comme il y a une très forte probabilité pour que l'action importante d'un plan se déroule en son milieu, les $p\%$ images du résumé d'un plan sont retenues sous la forme d'une sous-séquence centrée sur le milieu du plan.

En représentant chaque plan par un pourcentage de ses images, plus de poids est donné aux plans longs qui apportent plus d'informations sur la distribution des couleurs de la

séquence. Le résumé proposé, A_{seq}, est défini comme :

$$A_{seq} = \bigcup_{i=1}^{N_{plans}} seq^i_{p\%} \tag{4.2}$$

où N_{plans} est le nombre total de plans de la séquence et $seq^i_{p\%}$ est la sous-séquence des images retenues du plan i, contenant $p\%$ de ses images. L'influence du paramètre $p\%$ sur les résultats de la méthode proposée sera analysé plus tard dans la Section 4.2.4.

Si l'on prend en compte l'élimination des images peu représentatives des transitions dans l'étape d'agrégation en plans, le résumé proposé réduit le volume des données à traiter dans un rapport supérieur à $100/p$. Ce résumé, préservant l'essentiel de l'information qui nous est nécessaire, sera la base de départ des traitements ultérieurs.

4.2.3 La réduction des couleurs

Dans les images numériques, la couleur est typiquement représentée en utilisant 24 bits (plus de 16 millions de couleurs possibles), 8 bits pour chaque composante Rouge, Verte et Bleue. Ce nombre de couleurs est bien trop élevé et une réduction préalable des couleurs est indispensable pour traiter les images du résumé A_{seq}.

Le méthode de caractérisation de la distribution des couleurs dans les séquences utilise une approche statistique basée sur le calcul d'un histogramme global. Cette approche nous permet de réduire la résolution des images utilisées sans influencer les résultats du fait de la quasi-invariance de l'histogramme à un faible sous-échantillonnage spatial. Aussi, avant d'appliquer la réduction des couleurs, pour diminuer la complexité des calculs, les images du résumé A_{seq} sont d'abord sous échantillonnées spatialement d'un facteur 4 selon les lignes et les colonnes, ou plus si la taille initiale de l'image est importante. La réduction des couleurs est ensuite appliquée aux images sous échantillonnées du résumé A_{seq}.

Le principe de la réduction des couleurs a été présenté dans le Chapitre 2.4.2. Les méthodes existantes comportent deux étapes : le choix de la *palette des couleurs* et *l'association des couleurs* de la palette aux pixels de l'image ("pixel mapping"). En ce qui concerne la palette des couleurs les approches existantes se divisent en deux approches principales : celles qui utilisent une *palette fixe* définie a priori et valable pour toutes les images analysées, et celles qui utilisent une *palette adaptative* propre à chaque image.

Les contraintes de qualité de la réduction des couleurs

Comme nous l'avons déjà dit, les approches qui utilisent *une palette fixe* sont efficaces du point de vue du temps de calcul, mais cela se fait souvent au détriment de la qualité visuelle généralement moyenne. La qualité est bien sûr dépendante en premier lieu de la taille de la palette, mais aussi de la diversité des couleurs utilisées. Les approches qui utilisent *une palette adaptative* déterminent une palette optimale pour chaque image, ce qui engendre un temps de calcul élevé. De plus, la palette étant différente pour chaque image, la comparaison des images à travers leur distribution de couleurs est plus difficile.

Les films d'animation ont bien souvent la particularité d'utiliser une *palette limitée de couleurs* qui est spécifique à chaque film. Dans notre situation l'utilisation d'une palette fixe globale, a priori définie, ne provoquera donc pas trop de pertes dans la qualité visuelle de l'image après réduction des couleurs. Cette perte est minimale sous réserve d'un choix

judicieux de la palette et des performances de l'algorithme d'association des couleurs de l'image à celles de la palette. L'utilisation d'une palette fixe répond aussi à nos besoins en ce qui concerne la comparaison entre les distributions de couleurs des différentes images, comparaison difficile si on dispose pour chaque image d'une palette particulière.

D'autre part, au contraire de la détection des "cuts" où une qualité visuelle réduite suffisait à la mesure des différences entre images (voir la Section 2.4.2), la caractérisation de la distribution des couleurs de la séquence demande une *représentation fidèle du contenu* visuelle des couleurs.

Enfin, la méthode proposée pour la caractérisation de la distribution globale des couleurs de la séquence sera utilisée ensuite pour extraire des informations sémantiques sur les techniques d'utilisation des couleurs. La palette couleur utilisée doit donc *faciliter l'analyse* de la perception des couleurs. La façon la plus simple est d'utiliser les noms des couleurs. Ainsi, le choix d'une palette ayant des couleurs pour lesquelles on dispose d'une description nominative sera primordiale.

En conclusion, pour la caractérisation de la distribution des couleurs dans la séquence nous avons besoin d'une réduction des couleurs basée sur une palette fixe et répondant aux contraintes de qualité suivantes :

- **efficacité** : l'utilisation d'une palette de couleurs efficace, contenant un nombre de couleurs "moyen" mais en même temps présentant une diversité suffisamment élevée pour limiter les pertes visuelles dans l'image,
- **dénomination des couleurs** : la disponibilité des noms des couleurs,
- **qualité visuelle** : l'utilisation d'une méthode d'association des couleurs qui nous fournisse une qualité visuelle élevée de l'image.

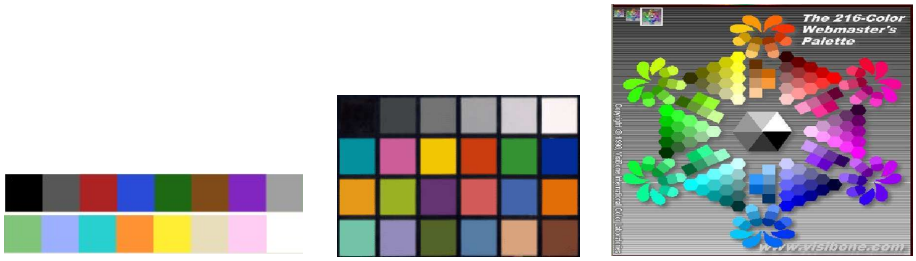
Association des couleurs : la diffusion d'erreur

En ce qui concerne la méthode d'association des couleurs de l'image aux couleurs de la palette, nous avons choisi d'utiliser la diffusion d'erreur, présentée dans le Chapitre 2.4.2. Cette méthode nous avait permis de préserver une qualité visuelle élevée de l'image (voir la Figure 2.4).

La choix de la palette des couleurs

Pour le choix de la palette des couleurs nous avons testé plusieurs palettes pour lesquelles on dispose d'un dictionnaire de noms ou d'informations sur le choix des couleurs proposées : une palette de 16 couleurs (voir la Figure 4.5.a), la roue des couleurs de Chevreul (palette perceptuelle de 72 couleurs, voir Figure 4.2.b), la roue des couleurs de Hering (inspirée de la théorie des couleurs opposées et contenant 17 couleurs, voir la Figure 4.2.c), la palette de test de Gretag Macbeth (25 couleurs, voir la Figure 4.5.b) et la palette "Webmaster" (utilisée pour choisir les couleurs dans la création des sites web, contenant 216 couleurs, voir la Figure 4.5.c). Des exemples de réduction des couleurs obtenues avec ces palettes et la diffusion d'erreur sont présentés dans la Figure 4.6.

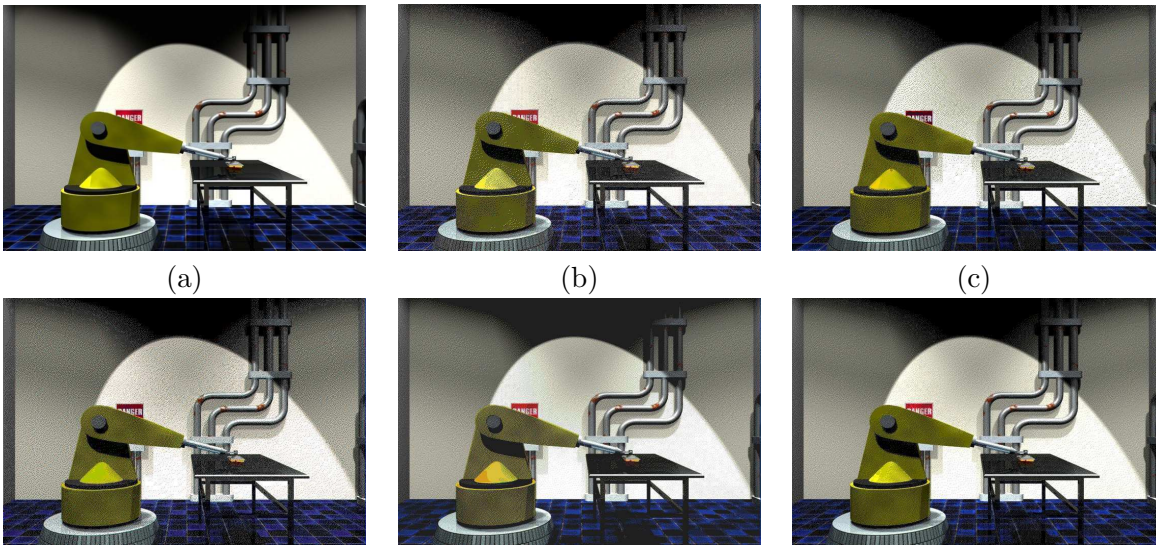
La palette optimale du point de vue du dictionnaire des noms des couleurs et de la diversité des couleurs utilisées est la palette "Webmaster" de 216 couleurs. Le dictionnaire des noms des couleurs fourni avec les autres palettes, quand il existe, est souvent trop détaillé, chaque couleur bénéficiant d'un nom différent. Ceci rend difficile l'analyse, en particulier celle



(a) (b) (c)

FIG. 4.5 – Les différentes palettes de couleurs utilisées : (a) palette 16 couleurs, (b) Gretag Macbeth 25 couleurs, (c) "Webmaster" 216 couleurs [Visibone 06].

des teintes similaires. De plus, les résultats obtenus (voir la Figure 4.6) montrent de manière claire que la meilleure qualité visuelle a été obtenue avec la palette "Webmaster", ce qui n'est pas réellement surprenant compte tenu du nombre de couleurs qu'elle contient. Les autres palettes n'ont pas un nombre et une diversité suffisante de couleurs.



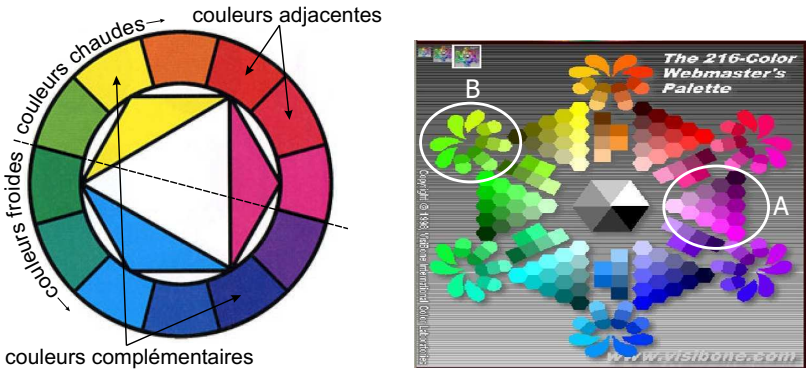
(a) (b) (c)

(d) (e) (f)

FIG. 4.6 – Exemples de réduction couleurs en utilisant les palettes testées : (a) Image initiale, (b) Palette 16 couleurs, (c) Chevreul 72 couleurs, (d) Hering 17 couleurs, (e) Gretag Macbeth 25 couleurs, (f) "Webmaster" 216 couleurs [Visibone 06].

La palette "Webmaster" proposée [Visibone 06] est composée de 12 couleurs élémentaires qui sont (sens de parcours horaire en commençant en haut de la roue, Figure 4.7) : "Orange", "Red", "Pink", "Magenta", "Violet", "Blue", "Azure", "Cyan", "Teal", "Green", "Spring" et "Yellow" auxquelles s'ajoutent 4 niveaux de Gris, le Noir et le Blanc.

Dans cette palette on peut retrouver trois zones différentes : des variations de la saturation et de l'intensité d'une même couleur élémentaire (par exemple la zone A dans la Figure 4.7.b), des mélanges entre différentes couleurs élémentaires (par exemple la zone B dans la Figure 4.7.b) et des niveaux de gris (zone de milieu dans la Figure 4.7.b). Pour chacune des 216



(a) (b)

FIG. 4.7 – (a) La correspondance avec la roue d'Itten, (b) La palette "Webmaster" 216 couleurs [Visibone 06] (zone A : variation d'une couleur élémentaire, exemple "Violet" ; zone B : des mélanges entre des variations de couleurs élémentaires).

couleurs proposées, un nom est associé en fonction de la couleur élémentaire de provenance (la teinte), du degré de saturation et de luminosité de la couleur. Quelques exemples de noms sont présentés dans le Tableau 4.1.

Couleur	Composantes (R, V, B)	Représentation hexa	Nom
	(255, 255, 51)	FFFF33	*"Light Hard Yellow"*
	(204, 0, 102)	CC0066	*"Dark Hard Pink"*
	(204, 204, 204)	CCCCCC	*"Pale Gray"*

TAB. 4.1 – Exemples de noms des couleurs de la palette "Webmaster".

Un autre avantage de la palette "Webmaster" est la correspondance avec la roue des couleurs d'Itten [Itten 61] (voir la Figure 4.7) qui est une représentation perceptuelle des couleurs. Dans la palette "Webmaster" les 12 couleurs élémentaires sont les mêmes que celles utilisées sur la roue d'Itten. Elles sont organisées de la même façon en formant un cercle de couleurs, permettant d'étudier les différentes relations perceptuelles entre les couleurs (les 7 contrastes et les principes d'harmonie des couleurs).

Les principaux avantages de la palette "Webmaster" peuvent être résumés de la manière suivante :

- **efficacité** : bon compromis entre la diversité des couleurs utilisées et le nombre total de couleurs,
- **appropriée pour les films d'animation** : richesse des couleurs suffisante pour la représentation des couleurs dans les films d'animation,
- **diversité** : diversité importante des couleurs : 12 couleurs élémentaires et 6 niveaux de gris avec le Noir et le Blanc,
- **relation entre couleurs** : correspondance avec la roue des couleurs d'Itten qui est une représentation perceptuelle des couleurs,
- **dénomination des couleurs** : disponibilité d'un dictionnaire des noms des couleurs,
- **algorithme efficace de dénomination** : efficacité du système d'association de noms aux couleurs, chaque nom contenant des informations sur la teinte, la saturation et la

luminosité de la couleur.

4.2.4 Le calcul de l'histogramme global pondéré

Pour déterminer la distribution globale des couleurs de la séquence on propose le calcul *d'un histogramme couleur global pondéré*, $h_{seq}(c)$. Cet histogramme demande le calcul préalable des histogrammes de chacun des plans du résumé A_{seq}.

Le calcul des histogrammes de chaque plan

Chaque plan vidéo est représenté par un *histogramme couleur moyen* qui est une mesure de la distribution globale des couleurs du plan. Cet histogramme est calculé sur la sous-séquence d'images, centrée sur le milieu du plan et contenant $p\%$ des images du plan. L'histogramme moyen d'un plan i, $\bar{h}_i(c)$, a pour expression :

$$\bar{h}_i(c) = \frac{1}{N^i_{img}} \cdot \sum_{j=1}^{N^i_{img}} h_{i,j}(c) \tag{4.3}$$

où N^i_{img} est le nombre d'images retenues du plan i, représentant $p\%$ de ses images, $h_{i,j}(c)$ est l'histogramme couleur de l'image retenue j du plan i et c est l'indice des couleurs dans la palette "Webmaster", $c = 1, ..., 216$. Les histogrammes $h_{i,j}(c)$ sont calculés pour les images sous échantillonnées spatialement et avec les couleurs réduites. Les valeurs ainsi obtenues pour les histogrammes moyens sont normalisées entre 0 et 1 et représentent le pourcentage d'apparition des couleurs à l'intérieur de chaque plan.

La méthode de calcul de l'histogramme couleur global

L'histogramme proposé est déterminé en fonction des histogrammes moyens de chaque plan, $\bar{h}_i(c)$ ($i = 1, ..., N_{plans}$ et N_{plans} le nombre total de plans), et en fonction des poids ω_i de chaque plan par :

$$h_{seq}(c) = \sum_{i=1}^{N_{plans}} \bar{h}_i(c) \cdot \omega_i \tag{4.4}$$

où c est l'indice des couleurs dans la palette "Webmaster" de 216 couleurs. Les poids ω_i, définis dans l'équation 4.1, déterminent l'importance de chaque plan par rapport à la séquence entière. Plus le plan est long, plus sa distribution de couleurs est importante pour la distribution globale de la séquence.

Les valeurs de l'histogramme global pondéré, $h_{seq}(c)$, correspondent au pourcentage d'apparition de chaque couleur c de la palette utilisée dans la séquence. Ce sont des valeurs positives et inférieures à 1 (1 = pourcentage d'apparition de 100%) grâce aux coefficients ω_i.

Le choix du pourcentage d'images utilisées

La qualité de la représentation de la distribution globale des couleurs en utilisant $h_{seq}(c)$ est liée à la valeur $p\%$ du pourcentage des images retenues pour chaque plan. Nous avons effectué une étude pour trouver la valeur optimale du paramètre p.

Il est évident que la représentation la plus exacte de la distribution globale des couleurs dans la séquence sera obtenue en utilisant toutes les images de tous les plans de la séquence, c'est-à-dire pour $p = 100\%$. L'histogramme global pondéré obtenu dans cette situation, $\widetilde{h_{seq}(c)} = h_{seq}(c)|_{p=100\%}$, est utilisé comme *référence* pour la mesure de la qualité des différentes distributions de couleurs obtenues pour des valeurs de $p\%$ inférieures.

Pour trouver la valeur optimale de p, p_{opt}, nous avons calculé plusieurs histogrammes globaux pondérés pour différentes valeurs de p, et nous les avons ensuite comparés à la référence $\widetilde{h_{seq}(c)}$. Comme mesure de similarité nous avons utilisé la distance Euclidienne, $d_E()$ entre l'histogramme global pondéré obtenu pour un pourcentage $p\%$ d'images et l'histogramme correspondant à un pourcentage de 100% (histogramme de référence), définie par :

$$d_E^2(p\%, 100\%) = \sum_{c=1}^{216} (h_{seq}(c)|_{p\%} - \widetilde{h_{seq}(c)})^2 \tag{4.5}$$

où c est l'indice des couleurs de la palette "Webmaster" de 216 couleurs.

Les résultats obtenus pour le film "La Cancion du Microsillon" [Folimage 06b], d'une durée totale de 8min28s, sont résumés dans le Tableau 4.2[1].

p	100%	75%	50%	25%	**15%**	~ 1 image	1%
N_{img}^t	12454	9290	6180	3038	**1780**	97	79
$d_E(p\%, 100\%)$	0	0.074	0.161	0.238	**0.267**	0.309	13.89
T_{calcul}	20h	15h	10h	5h	**2h**	10min	8min

TAB. 4.2 – L'influence de la valeur du paramètre $p\%$ sur l'histogramme global pondéré, $h_{seq}(c)$ (N_{img}^t est le nombre total d'images retenues pour différentes valeurs de p et T_{calcul} est le temps de calcul).

Le meilleur compromis entre le temps de calcul et la qualité de la distribution globale des couleurs est obtenu en prenant $p_{opt} \in [15\%, 20\%]$ (temps de traitement moyen de $4.5s$ par image dans les conditions mentionnées dans[1]). La distance d_E augmente avec la réduction du nombre d'images utilisées. La valeur maximale de d_E est obtenue quand certains plans ne sont pas représentés dans le calcul de l'histogramme global, situation qui correspond à $p = 1\%$, où aucune image n'est retenue pour les plans d'une durée inférieure à 4 secondes.

4.3 Résultats expérimentaux : quelques exemples

La caractérisation des couleurs d'un film est fortement dépendante de la qualité du film utilisé. Dans notre cas, les films numérisés proviennent essentiellement de supports magnétiques (cassettes VHS, SVHS ou Beta), parfois datant de plus de vingt ans, et les couleurs obtenues sont souvent très différentes des couleurs originales. Typiquement, la saturation diminue et l'intensité présente des fluctuations importantes (par exemple les films "The Young Lady and The Cellist" (1964) ou "Tamer of Wild Horses" (1964) présentés dans l'Annexe F). Un moyen de combattre cet effet est d'effectuer une normalisation des couleurs, étape que nous n'avons pas réalisée car elle débordait du cadre de ce travail.

[1]les temps de calculs présentés sont donnés à titre indicatif pour permettre des comparaisons relatives car l'algorithme n'a pas été optimisé. Les tests ont été effectués en utilisant des images sous échantillonnées d'une résolution 190×104 pixels, sur une machine Pentium IV à 3GHz avec 1GB de mémoire RAM.

Nous allons présenter les résultats obtenus pour 4 films d'animation représentatifs produits par [Folimage 06b] : "Casa" (6min5s), "Le Moine et le Poisson" (6min), "Circuit Marine" (5min35s) et "François le Vaillant" (8min56s). Quelques images pour chacun de ces films sont données dans la Figure 4.8. Les histogrammes globaux pondérés obtenus sont proposés en Figure 4.9 (seules les couleurs ayant un pourcentage d'apparition supérieur à 1% sont représentées).

Pour les 4 films testés, les couleurs ayant les fréquences d'apparition les plus importantes (supérieures à 4%) sont les suivantes :

- **film "Casa"** : "Pale Gray" - 8.9% ; "Black" - 6.6% ; "Light Weak Yellow" - 9.5% ; "Light Weak Cyan" - 7.4% ; "Light Dull Orange" - 5.6% ; "Obscure Dull Orange" - 5.7% ; "Dark Dull Orange" - 4.3% ; "Dark Weak Red" - 5.7% ; "Obscure Weak Red" - 9.2% ; "Dark Orange-Red" - 4.2% ; "Medium Orange-Red" - 4.3%. Les couleurs élémentaires prédominantes sont *Orange*, *Rouge*, *Jaune*, *Cyan*, *Gris* et *Noir*.
- **film "Le Moine et le Poisson"** : "Black" - 19.6% ; "Pale Dull Yellow" - 5.2% ; "Light Weak Yellow" - 13.7% ; "Light Dull Yellow" - 11.5% ; "Dark Dull Yellow" - 4.9%, "Light Weak Green" - 4%. Les couleurs élémentaires prédominantes sont *Noir*, *Jaune* et *Vert*.
- **film "Circuit Marine"** : "White" - 4.2% ; "Dark Gray" - 11.2% ; "Pale Gray" - 5.3% ; "Light Gray" - 5.6% ; "Pale Weak Blue" - 4.9% ; "Medium Weak Blue" - 4.8% ; "Pale Dull Azure" - 6.5% ; "Light Weak Red" - 5.3% ; "Medium Weak Red" - 7.8% ; "Dark Weak Red" - 4.2%. Les couleurs élémentaires prédominantes sont *Blanc*, *Gris*, *Bleu* et *Rouge*.
- **film "François le Vaillant"** : "Black" - 13.7% ; "Obscure Gray" - 6.1% ; "Pale Dull Azure" - 4.6% ; "Light Cyan-Azure" - 4.2% ; "Dark Dull Azure" - 4.1% ; "Dark Hard Azure" - 19% ; "Obscure Dull Azure" - 5.4%. Les couleurs élémentaires prédominantes sont *Noir*, *Gris* et *Bleu-azur*.

FIG. 4.8 – Quelques images des 4 films d'animation testés.

En analysant les résultats nous pouvons remarquer que l'histogramme global pondéré, $h_{seq}(c)$, que nous venons de présenter, offre des informations précises sur la distribution des couleurs et sur le pourcentage d'apparition de chaque couleur représentative de la séquence.

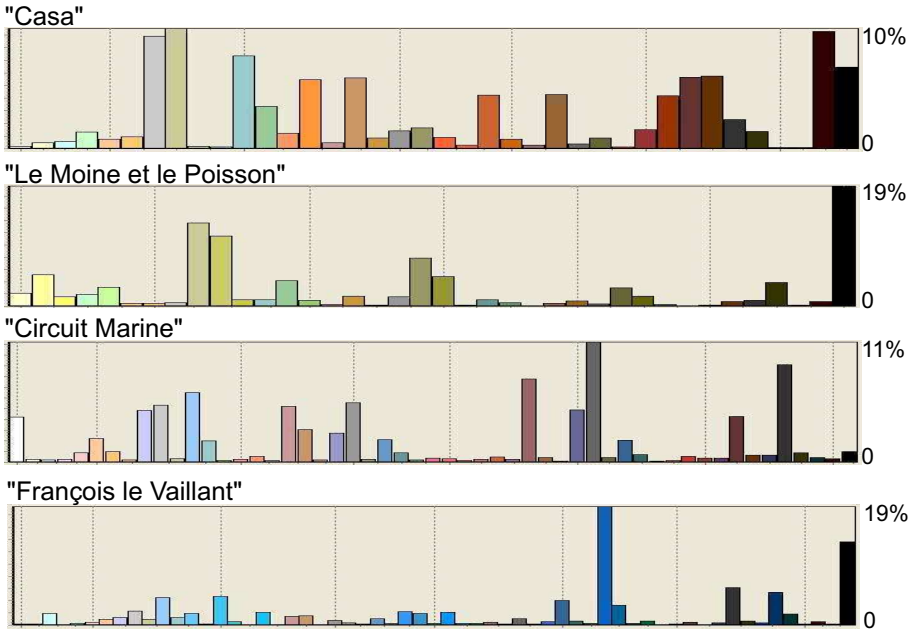


FIG. 4.9 – Les histogrammes globaux pondérés obtenus pour les 4 films d'animation testés (l'axe oY correspond au pourcentage d'apparition des couleurs, l'axe oX à l'index des couleurs et $p = 20\%$).

L'histogramme global pondéré en conjonction avec le dictionnaire des noms des couleurs associé à la palette "Webmaster", va être le point de départ d'une caractérisation des différentes techniques d'utilisation de la couleur dans les films d'animation.

4.4 Conclusions générales

Dans ce chapitre nous avons étudié les différentes façons d'analyser et de caractériser les couleurs dans les séquences d'images.

Dans le cas particulier des films d'animation nous avons proposé une méthode de caractérisation de la distribution globale des couleurs dans le film. La méthode proposée, comme nous l'avons déjà dit, est motivée par le fait que les films d'animation utilisent souvent une *palette de couleurs réduite* et propre à chaque film, une signature couleur en quelque sorte. Cette caractérisation s'appuie sur le calcul d'un *histogramme couleur global pondéré* de la séquence. L'inconvénient de l'utilisation des histogrammes est la perte d'information sur la distribution spatiale des couleurs. Mais notre objectif étant une caractérisation globale des couleurs de la séquence, cette perte d'information spatiale n'est pas pénalisante. Avant de calculer cet histogramme, les couleurs de chaque image analysée sont projetées sur une palette particulière choisie a priori (palette "Webmaster" [Visibone 06]), pour laquelle chaque couleur est décrite de manière textuelle, ce qui va permettre, dans la deuxième partie de la thèse, une analyse de la perception de cette signature couleur (les 7 contrastes d'Itten, l'harmonie des couleurs, etc.).

Nous avons pu tester notre approche sur un certain nombre de films d'animation. L'histogramme global pondéré se trouve être un outil très efficace pour la caractérisation de la

distribution globale des couleurs de la séquence. Même calculé sur un nombre très réduit d'images de la séquence, par exemple avec une image par plan (et donc une compression de la séquence supérieure à 1/120), nous avons obtenu une représentation fidèle de la distribution des couleurs de la séquence. De plus, si nous disposons du découpage en plans de la séquence (nécessaire pour la sélection des sous-séquences), la complexité de calcul de l'histogramme est très réduite et proportionnelle à la complexité du calcul de la réduction des couleurs multipliée par le nombre d'images utilisées. Une implantation en temps réel est alors envisageable, par exemple en parallèle avec le découpage en plans, sans utilisation de dispositif matériel spécifique.

Troisième partie

Vers la description sémantique

Chapitre 5

La détection des scènes

Résumé : *Si les plans sont les unités syntaxiques d'une séquence alors les scènes peuvent être considérées comme les unités sémantiques de la séquence. Dans ce chapitre nous allons d'abord présenter les différentes techniques de détection et de classification de scènes dans les séquences d'images. Nous proposerons ensuite une méthode de détection des scènes basée sur l'analyse des similarités entre les distributions couleurs des plans vidéo voisins. Enfin, nous verrons comment la segmentation en scènes peut être exploitée pour des tâches complémentaires telles que la correction du découpage en plans, la détection des "shot-reverse-shot" (technique particulière de prise de vue) et la représentation hiérarchique du contenu de la séquence.*

La segmentation temporelle d'un film fournit typiquement entre 600 et 1500 plans. Si l'on résume chaque plan par une seule image, on aboutit à une représentation réduite de la séquence comprenant de 600 à 1500 images. Cette représentation contient encore trop d'information pour analyser et visualiser le contenu de la séquence, en particulier pour certains types de séquences comme les films, les documentaires, les informations télévisées, etc. Il est donc nécessaire de disposer d'une représentation structurelle de plus haut niveau que les plans vidéo. Une manière de procéder consiste à définir un découpage en *scènes*.

La détection des éléments structurels de plus haut niveau n'est pas uniquement réservée à l'évaluation du contenu de la séquence, mais elle servira également à accéder au contenu de la séquence d'une *manière sémantique*. Généralement la durée des scènes et leur fréquence d'apparition sont des éléments importants pour l'étude du rythme et du style du réalisateur. Par exemple une scène contenant une suite de plans courts reflète un niveau d'action élevé, une scène contenant des plans dont la durée est de plus en plus courte reflète une augmentation de la tension dans la séquence, etc.

Dans la littérature spécialisée, les scènes sont souvent appelées *les unités de la narration*, ou *les segments de la narration* ou encore *les paragraphes vidéo*. En utilisant la terminologie utilisée dans le domaine de la production de films, une scène est *typiquement composée d'un nombre réduit de plans qui sont unifiés par le lieu de l'action ou par les événements* [Beaver 94]. Dans le langage scientifique nous pouvons dire qu'une scène est une séquence de

plans vidéo qui sont reliés par des caractéristiques sémantiques communes. Comme pour le théâtre classique, on peut donc considérer que le contenu d'une scène doit respecter la règle des trois unités : l'unité de *temps*, de *lieu* et *d'action* [J.M.Corridoni 95]. Un exemple est illustré dans la Figure 5.1.

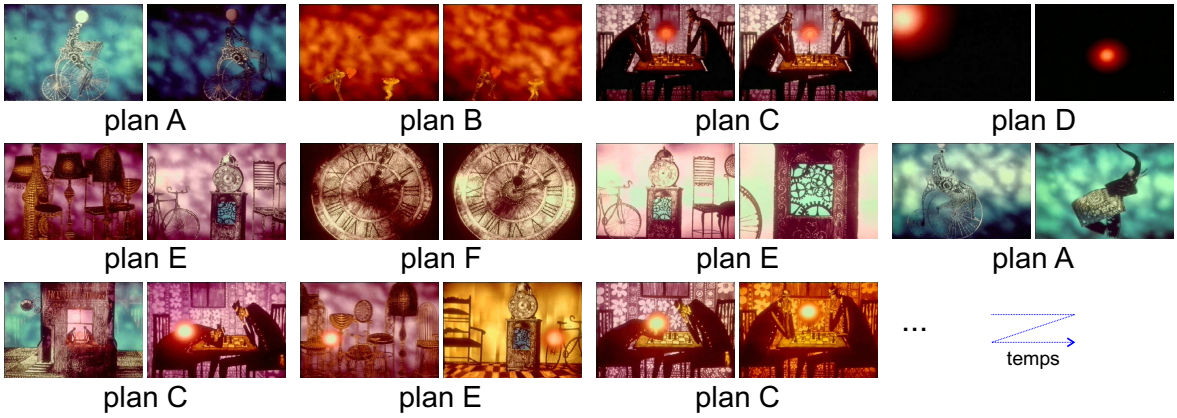


FIG. 5.1 – Exemples de scènes : la scène X est la réunion des plans de type X, où $X \in \{A, B, C, D, E, F\}$ (extrait du film d'animation "Cœur de Secours" [CICA 06]).

A partir des définitions énoncées ci-dessus il devient évident que si les plans vidéo sont les *unités syntaxiques* de la séquence, les scènes constituent *les unités sémantiques*. Leur détection est liée à la compréhension sémantique du contenu de la séquence. Le concept théorique qui présente le film comme une combinaison harmonieuse d'éléments syntaxiques (les plans) en utilisant un méta-langage est connu sous le nom de *paradigme de montage* [J.M.Corridoni 95]. Dans ce cas la scène constitue la brique de base pour la compréhension du contenu sémantique de la séquence.

5.1 État de l'art

Les techniques de traitement d'images et de vision par ordinateur ne sont pas encore suffisamment performantes pour permettre la compréhension sémantique complète du contenu des séquences d'images. Pour réaliser une vraie analyse de scènes, il est nécessaire de comprendre totalement le sujet et l'action de la séquence. Les techniques existantes de recherche des scènes tentent de contourner ce problème en utilisant des paramètres de bas niveau, comme la similarité des couleurs, de la texture, du mouvement, etc., tout en incluant une connaissance a priori sur les techniques de réalisation des films.

Parmi les méthodes existantes de détection des scènes on retrouve principalement deux directions d'analyse :

- les méthodes qui font *le découpage en scènes* de la séquence ou qui détectent *les changements de scènes*,
- les méthodes qui font la *classification des scènes* en fonction de leur contenu sémantique, par exemple en scènes de dialogue, en scènes de violence, en scènes de chasse, etc.

Les classes utilisées sont spécifiques aux différents domaines d'application : documentaires, films, journaux télévisés, etc. Différents états de l'art sur l'agrégation de la séquence

en éléments de plus haut niveau et particulièrement en scènes ont été proposés, comme par exemple [Bimbo 99][Kang 01][Snoek 05b]. Dans la suite nous allons présenter les principales particularités des différentes techniques existantes.

5.1.1 La classification des scènes

Le but de la classification du contenu des scènes est de regrouper différents segments de la séquence en classes sémantiques a priori définies. La classification des scènes peut être réalisée selon plusieurs niveaux sémantiques [Wang 00] :

- **le plus bas niveau** : à ce niveau les segments vidéo peuvent être regroupés en classes élémentaires comme : scènes d'extérieur/scènes d'intérieur, scènes d'action/scènes sans action, etc.
- **niveau intermédiaire** : dans ce niveau on retrouve les scènes contenant des événements de base, comme par exemple des scènes de dialogue entre deux personnes, une scène de concert, une scène se déroulant sur une plage, etc.
- **le plus haut niveau** : pour les scènes de cette catégorie, la compréhension de l'histoire du film est obligatoire. Ce sont des scènes contenant des événements complexes, comme par exemple l'ouragan survenu en Floride en 2004, la célébration du nouvel an au pied de la Tour Eiffel, etc.

Différentes méthodes ont été proposées pour la classification des différents genres de scènes.

Dans [Alatan 01], l'analyse des passages audio associés à la détection et à la localisation des visages ont été utilisés dans des modèles de Markov cachés pour la classification des scènes de dialogue dans les films. Les scènes de violence sont détectées dans [Nam 98] en utilisant des informations mélangeant l'audio et l'image telles que l'activité spatio-temporelle dans la séquence, la détection du sang et des flammes, les changements d'énergie du signal audio.

L'approche proposée dans [Saraceno 98] utilise les informations audio et image de la séquence pour classifier les scènes dans les catégories suivantes : scènes de dialogue, scènes de narration, scènes d'action et scènes de générique. Les séquences sont découpées de manière indépendante, en plans vidéo et plans audio. Le découpage en plans audio est réalisé en analysant l'énergie du signal audio. Les plans audio ainsi obtenus sont classés en passages de silence, de parole, de musique et de bruit. Le découpage en plans vidéo est réalisé en utilisant l'information sur la couleur. Ensuite, une méthode de quantification vectorielle est appliquée pour retrouver des modèles de blocs spécifiques dans les plans vidéo ainsi obtenus. Les plans sont ensuite regroupés selon leurs caractéristiques visuelles et auditives en groupes homogènes (contenant le même label) à partir d'un certain nombre de modèles prédéfinis. Pour la détection des scènes recherchées (dialogue, narration, etc.) les modèles associés sont exprimés sous forme de règles sur les informations dont on dispose. Par exemple, pendant une scène de dialogue le signal audio contient plutôt des passages de parole et les labels des plans vidéo suivent un modèle du type $ABABAB$. Pendant une scène d'action le signal audio ne contient pas de parole et l'information visuelle suit une évolution progressive, donc un modèle du type $ABCDE$.

Une approche utilisant des critères différents pour la classification de scènes est proposée dans [Lienhart 99b]. Les scènes détectées sont les scènes contenant des caractéristiques auditives similaires, les scènes filmées dans des endroits similaires et les scènes de dialogue. La méthode proposée contient quatre étapes : dans un premier temps la séquence est découpée

en plans à partir de la détection des "cuts", des "fades" et des "dissolves". Dans un deuxième temps, un certain nombre d'informations sont extraites : caractéristiques associées au son, à la couleur, aux orientations des contours et aux visages. L'étape suivante consiste à calculer des distances entre chaque paire de plans vidéo en utilisant indépendamment chaque type d'information extraite. Les distances obtenues pour chaque modalité sont organisées sous forme de tableaux de distances. Les plans vidéo sont alors regroupés en fonction de leur distance par type de scène. Par exemple pour détecter les scènes contenant des passages audio similaires, appelés des "séquences audio", la distance entre deux plans est définie comme la distance minimale entre les vecteurs des caractéristiques audio de ces deux plans. Ainsi, les plans voisins ayant des distances inférieures à un certain seuil (proche de la distance minimale) sont regroupés dans la même "séquence audio". Les auteurs prétendent qu'on obtient des résultats meilleurs si l'on fait la classification des scènes en utilisant d'abord les modalités de la séquence de manière indépendante et qu'ensuite on regroupe les scènes obtenues dans une étape liée à l'application.

D'une manière générale, les méthodes de classification de scènes ont besoin de définir a priori les catégories de scènes à détecter. Il n'y a pas d'approche générique permettant de retrouver n'importe quel type de scène sans un minimum de connaissance sur le contenu de la séquence. De plus, les algorithmes de détection sont typiquement spécifiques à chaque type de scène à détecter et ils ne sont pas applicables à la détection d'autres catégories de scènes. Par exemple, l'algorithme de détection des scènes de dialogue ne peut par être utilisé pour les scènes de chasse.

Par contre, on peut envisager une approche plus générale, ne s'intéressant plus à la classification du contenu mais plutôt à la structure temporelle de la séquence : c'est le découpage en scènes.

5.1.2 Le découpage en scènes

Le découpage en scènes ne cherche pas à classifier le contenu des scènes dans la séquence, mais à représenter sa structure temporelle en utilisant des éléments de plus haut niveau sémantique que les plans vidéo. Typiquement, les méthodes existantes sont indépendantes de l'application ou du genre de la séquence. Un ensemble de règles globales pour identifier les scènes dans une séquence d'images a été défini dans [Aigrain 95]. Les règles proposées sont suffisamment invariantes au type de séquence utilisée et elles prennent en compte :

- la façon dont **les transitions vidéo** sont insérées entre les "cuts".
- **la distance entre les plans similaires** de la séquence : un changement de scène est détecté si un plan similaire à celui analysé est retrouvé à une distance de 2 ou 3 plans. La similarité entre les plans est analysée en utilisant des distances entre des images de luminance représentatives de chaque plan.
- **la similarité entre les plans voisins** : la continuité du contenu des plans est détectée en utilisant une mesure de similarité, notée S. Cette mesure est calculée en fonction de la valeur moyenne $m(i)$ et de l'écart type $\sigma(i)$ de la saturation et de la teinte des pixels des images clés de chaque plan. La mesure s'exprime par : $S_{i,j} = |m(i) - m(j)| + |\sigma(i) - \sigma(j)|$, où i, j sont les indices des plans.
- **le rythme du montage** : un changement de scène est détecté seulement si le plan courant analysé est trois fois plus long qu'une durée L ou quatre fois plus court que L. La durée L est estimée en utilisant un modèle autorégressif : $L_n = a \cdot L_{n-1} + b \cdot L_{n-2}$, où L_n est la valeur de L à l'itération n et les paramètres a, b sont estimés dans une

fenêtre glissante de 10 plans.
- la présence de **musique après un passage de silence**.
- **le même type de mouvement de caméra** : les chaînes de trois plans ou plus contenant le même mouvement de caméra sont considérées comme appartenant à la même scène.

Différentes méthodes ont été envisagées : [Huang 98] propose une segmentation hiérarchique de la séquence. Les changements de plan et les changements de scène sont détectés à différents niveaux hiérarchiques. Les changements de scène sont associés aux changements simultanés des couleurs, du mouvement et des caractéristiques sonores. [H. Sundaram 00] propose de modéliser les relations de cohérence entre les plans par l'utilisation d'un modèle de mémoire causale de type FIFO ("First-In-First-Out"). L'approche proposée dans [Aner 01] résume le contenu de chaque plan en utilisant des images "mosaïques" (voir le Chapitre 3). Rechercher la similarité entre les différents plans revient à rechercher la similarité entre les images "mosaïques" associées. Les différents plans sont alors regroupés en scènes en analysant les distances entre leurs images "mosaïques".

Une approche intéressante est l'approche hybride proposée dans [Kang 01] qui utilise conjointement plusieurs techniques de détection. La segmentation hiérarchique en scènes est réalisée en trois étapes illustrées dans la Figure 5.2.

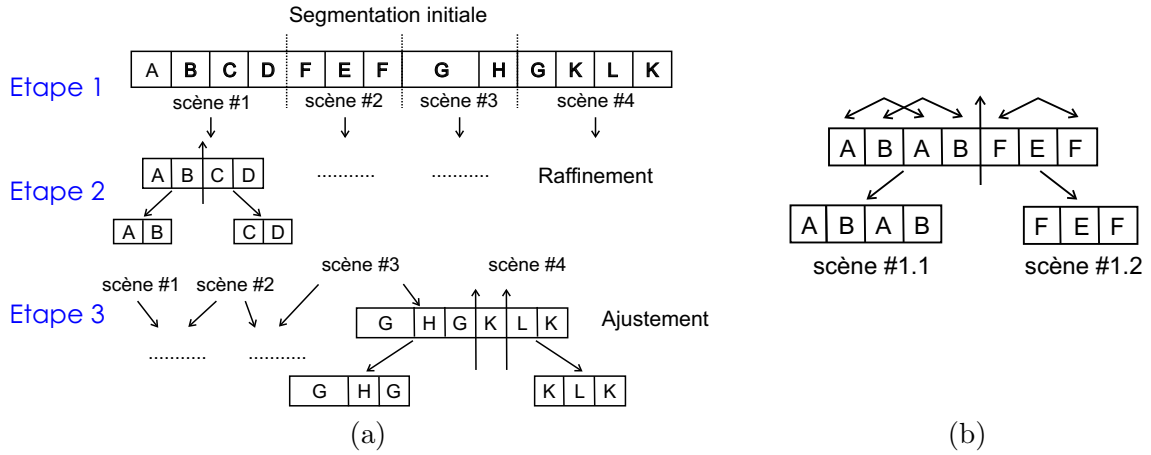


FIG. 5.2 – Une approche hiérarchique de segmentation en scènes [Kang 01] : (a) Les étapes d'analyse, (b) Exemple de raffinement de la détection.

La première étape consiste en une *segmentation initiale* de la séquence en scènes. L'approche utilisée est une approche continue de détection. La cohérence entre les différents plans est calculée en utilisant le modèle de mémoire causale de type FIFO proposé dans [H. Sundaram 00]. La valeur de la cohérence est une valeur continue qui est calculée en fonction de la valeur du rappel des plans. Le rappel entre deux plans A et B, noté $Rappel(A, B)$ à l'instant S_a, où S_a est le point dans la mémoire tampon de départ de l'analyse sur lequel on se focalise ("attention-span"), est défini par :

$$Rappel(A, B) = (1 - dissim(A, B)) \cdot L_A \cdot L_B \cdot (1 - \frac{\Delta n}{N_m}) \cdot (1 - \frac{\Delta t}{T_m}) \qquad (5.1)$$

où $dissim(A, B)$ est une mesure de dissimilarité entre les plans A et B, L_A et L_B sont les tailles des plans A et B pondérées par des coefficients faisant intervenir T_m la taille de la

mémoire tampon, N_m le nombre total de plans contenus dans la mémoire, Δn le nombre de plans contenus entre les plans A et B et Δt l'écart temporel entre les plans A et B. La valeur de la cohérence $Co(S_a)$, calculée à l'instant S_a, est définie par :

$$Co(S_a) = \frac{Rappel(A,B)}{R_{max}(S_a)} \tag{5.2}$$

où $R_{max}(S_a)$ est la valeur maximale du rappel des plans obtenue dans le cas particulier où $dissim(A,B) = 0$.

Les changements de scène sont détectés en analysant les valeurs de la cohérence dans une fenêtre de décision. Un changement de scène est détecté quand le minimum local de la fonction de cohérence se trouve au milieu de la fenêtre de décision considérée.

La deuxième étape d'analyse utilise *une approche discrète* pour raffiner le découpage initial en scènes. Le but de cette étape est de corriger les non détections. Les plans contenus dans les scènes obtenues lors de la première étape sont regroupés en utilisant un certain nombre d'images clés. La classification est effectuée avec un classifieur k-means pour lequel le nombre optimal de classes a été déterminé au préalable par une méthode d'analyse de la véracité des classes. Après la classification on dispose d'un label par type de plan, et les scènes sont alors transformées en une séquence de labels. Les plans ayant le même label sont associés en construisant un lien entre eux. Si à l'intérieur d'une scène on trouve un point pour lequel il n'y a aucune association entre plans, la scène est divisée en deux en ce point (voir Figure 5.2.b).

La dernière étape d'analyse est l'étape d'ajustement de la segmentation. Cette étape permet de corriger les fausses détections. Les plans de deux scènes adjacentes sont regroupés en utilisant un classifieur k-means de la même façon que dans l'étape précédente. Pour déterminer les faux changements de scène, les liens entre les plans d'une scène avec des plans de la deuxième scène sont analysés.

Globalement, les approches existantes de découpage en scènes de la séquence transforment la détection de scènes en un problème de similarité entre plans (à partir de règles) ou plus généralement en un problème de classification des plans. Dans ce contexte une scène est vue comme un ensemble de plans voisins ayant des propriétés similaires (couleur, mouvement, etc.). Différentes mesures de distance entre le contenu des plans sont utilisées pour la détection.

5.2 Méthode proposée

La méthode de détection de scènes que nous proposons (voir [Ionescu 05e]) s'inspire de l'approche de classification de scènes présentée dans [Lienhart 99b], mais nous l'adaptons ici au découpage en scènes. Les films d'animation ayant chacun une distribution spécifique de couleurs, la stratégie que nous utilisons s'appuie sur des tableaux de distances entre les plans vidéo, distances évaluées à partir des similitudes des distributions couleurs de chaque plan. Le diagramme de la méthode est illustré par la Figure 5.3.

Les étapes d'analyse sont les suivantes :

- **résumé de chaque plan** : le contenu de chaque plan vidéo est résumé par un certain pourcentage d'images qui sont extraites de manière uniforme à l'intérieur du plan.
- **prétraitements et calcul des histogrammes** : en utilisant les résumés obtenus dans l'étape précédente, la distribution globale des couleurs de chaque plan est représentée

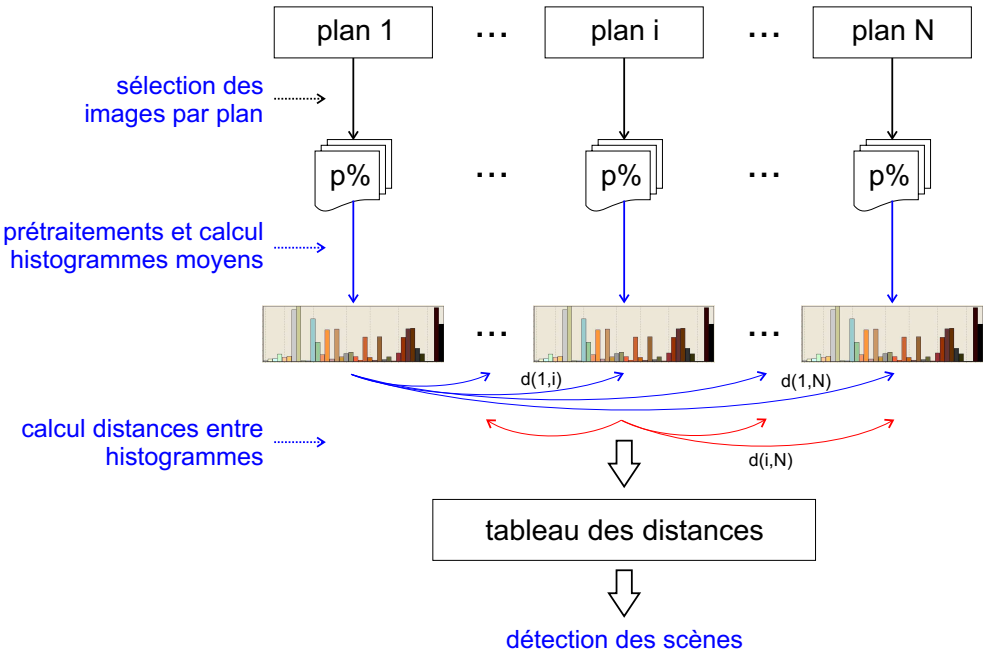


FIG. 5.3 – Le diagramme de la méthode de détection des scènes.

par un histogramme moyen couleur. Avant de calculer les histogrammes un certain nombre de prétraitements sont effectués pour réduire la complexité de calcul : sous-échantillonnage spatial, réduction des couleurs, etc.

- **calcul des distances** : pour chaque plan vidéo d'indice i on calcule les distances Euclidiennes entre l'histogramme moyen du plan i et les histogrammes moyens de tous les autres plans $j \neq i$ de la séquence. Les résultats obtenus sont représentés sous la forme d'un tableau de distances.
- **détection des scènes** : les scènes sont ensuite déterminées en seuillant les valeurs des distances ainsi obtenues.

Le résumé de chaque plan

Dans un premier temps la séquence est *découpée en plans* en utilisant les méthodes proposées dans le Chapitre 2. Chaque plan obtenu est *résumé* par un certain pourcentage, $p\%$, des images le composant. Les $p\%$ images extraites sont uniformément distribuées à l'intérieur du plan.

La précision de la détection de la frontière entre plans est de quelques images car elle est liée au pas d'analyse temporel utilisé. Avec une sélection uniforme, il est donc possible de retrouver à l'intérieur d'un plan des images qui appartiennent au plan d'avant et au plan d'après. Pour éviter cela, la sélection des plans est réalisée en introduisant une marge de sécurité en début et en fin de plan. Si le plan courant analysé est le $plan_i = [a, b]$, contenant les images comprises entre les instants a et b, les $p\%$ images retenues seront distribuées uniformément dans l'intervalle $[a + T_e, b - T_e]$, où T_e est la valeur de marge de sécurité déterminée empiriquement à 5 images.

En utilisant un nombre d'images rapporté à la taille de chaque plan, cela nous permet

de retenir plus d'informations pour les plans longs que pour les plans courts, leur contenu étant plus important dans la séquence.

Prétraitements et calcul des histogrammes

Nous allons représenter le contenu des couleurs de chaque plan à l'aide *de l'histogramme couleur moyen* défini dans l'équation 4.3 de la Section 4.2.4. Avant le calcul de ces histogrammes un certain nombre de prétraitements sont effectués. Chaque image retenue de chaque plan est dans un premier temps *sous échantillonnée* spatialement : seul un pixel de chaque bloc de 4×4 pixels (sans recouvrement) est retenu, ceci permettant de réduire le temps de calcul. Comme la méthode de détection proposée utilise des mesures statistiques (des histogrammes) l'influence du sous échantillonnage spatial des images sur les résultats est négligeable.

Le deuxième prétraitement concerne la couleur. Comme nous l'avons déjà évoqué (voir la Section 2.4.2), pour calculer les histogrammes couleurs moyens nous avons besoin de *réduire le nombre de couleurs* de l'image. La méthode de réduction des couleurs utilisée doit nous fournir une représentation fidèle du contenu visuel de l'image. C'est pour cela que nous avons décidé d'utiliser la méthode de réduction des couleurs qui utilise la diffusion d'erreur sur la palette "Webmaster".

En ce qui concerne le calcul des histogrammes nous avons testé deux stratégies différentes :

- **histogrammes couleurs "classiques"** : la première stratégie consiste à calculer des *histogrammes couleurs "classiques"*. L'histogramme moyen d'un plan i, noté $\bar{h}_i(c)$, où c est l'index de couleur, est calculé en faisant la moyenne des histogrammes couleur "classiques" de chaque image retenue (ce qui représente $p\%$ du nombre total des images du plan).
- **histogrammes pondérés** : comme les histogrammes classiques ne prennent pas en compte l'information spatiale de l'image, la *deuxième stratégie* consiste à intégrer cette information spatiale à travers le calcul de ce que nous avons appelé les *histogrammes pondérés*, notés $h_i^*(c)$ dans la suite. L'histogramme moyen pondéré d'un plan i, noté $\bar{h}_i^*(c)$, sera ensuite calculé en faisant la moyenne des histogrammes couleur pondérés de chaque image retenue.

L'histogramme couleur pondéré de l'image I (représentée sur une palette de couleurs réduite), $h_I^*(c)$ est défini de la manière suivante :

$$h_I^*(c) = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} \delta(I(m,n) - c) \cdot (1 - g(m,n)) \tag{5.3}$$

où c est l'indice de la couleur, $c = 1, ..., L$, avec L le nombre de couleurs utilisées, $M \times N$ est la taille de l'image I, $g(m,n)$ est la carte des amplitudes du gradient dans l'image I, et $\delta(x-y) = 1$ si $x = y$ et 0 autrement. Les valeurs de l'amplitude du gradient sont normalisées entre 0 et 1, la valeur 1 correspond au contour le plus important de l'image.

La carte $g(m,n)$ de l'amplitude du gradient est calculée en utilisant les masques de Sobel. Ces masques sont appliqués sur la composante de luminance Y, de l'espace $YCbCr$ (voir l'équation 2.27 de la Section 2.6.2). Dans l'image $g(m,n)$, les points de contour de l'image sont représentés par des valeurs élevées de $g(m,n)$ et les régions uniformes par des valeurs faibles proches de 0. Dans l'image inverse $1 - g(m,n)$ on retrouve bien sûr le contraire : les

régions uniformes de l'image ont des valeurs élevées et les points de contour ont des valeurs faibles. Un exemple est présenté dans la Figure 5.4.

Image originale

Image inverse de l'amplitude du gradient

FIG. 5.4 – Image originale et Amplitude du gradient (les valeurs ont été normalisées entre 0-Noir et 255-Blanc).

Dans l'histogramme couleur pondéré, les zones où le gradient est faible (régions uniformes) vont donc avoir un poids important, alors que les zones de contours n'auront que très peu de poids. L'histogramme pondéré apporte donc une information plus pertinente que l'histogramme classique, puisqu'il accentue les régions uniformes (contenant l'information principale sur les couleurs) au détriment des régions de contours de l'image.

Calcul de distances entre plans

La mesure de similarité entre plans est évaluée par la distance Euclidienne entre les histogrammes moyens de ces plans, que ce soit avec l'histogramme moyen classique ou avec l'histogramme moyen pondéré. Comme les histogrammes sont calculés sur une même palette fixe de 216 couleurs (palette "Webmaster" présentée en Section 2.4.2), cette distance est tout à fait suffisante pour donner une bonne estimation de l'écart entre les plans.

Nous analysons, sur la séquence entière, la similarité entre les plans en calculant, pour chaque paire de plans, les distances Euclidiennes entre les histogrammes moyens associés. Ainsi, nous aboutissons à un tableau croisé de distances, dans lequel chaque distance est définie par :

$$D(i,j) = \{d_E(\bar{h}_i, \bar{h}_j) \ / \ i \neq j \ et \ i,j = 1, ..., N_{plans}\} \tag{5.4}$$

où d_E désigne la distance Euclidienne et N_{plans} est le nombre total de plans dans la séquence. La même relation est utilisée dans le cas des histogrammes moyens pondérés, $\bar{h}_i^*$.

Le détection de scènes

En utilisant l'hypothèse que les plans qui font partie de la même scène ont des distributions de couleurs similaires (*unité du contenu*), la détection de scènes est effectuée en regroupant les plans les plus ressemblants, ceux pour lesquels la distance Euclidienne entre les histogrammes moyens est faible. La mise en œuvre est réalisée par simple seuillage : si la valeur courante analysée, $D(k,l)$, est inférieure à un certain seuil τ_{scene} les plans k et l seront déclarés similaires. Un label sera alors associé aux plans similaires. La valeur du seuil a été fixée empiriquement à $\tau_{scene} = 0.1$, valeur similaire au seuil utilisé pour la détection des "cuts", τ_{cut}, dans la Section 2.4.3 car les deux méthodes s'appuient sur l'analyse des histogrammes couleurs en utilisant la même réduction des couleurs.

Une scène sera composée de tous les plans associés au même label et voisins du point de vue temporel (*unité temporelle*). Ceci se traduit par le fait que ces plans doivent avoir une distance inférieure à 10 plans (valeur déterminée empiriquement après l'analyse de plusieurs séquences de film d'animation). Pour regrouper les plans ayant le même label nous appliquons le principe de transitivité qui peut s'exprimer par : si le $plan_i$ est similaire au $plan_j$, tous deux associés au même label A, et si le $plan_j$ est similaire au $plan_k$ associés au même label B, alors le $plan_i$ est similaire au $plan_k$ et les trois plans seront associés au label $A = B$. Les résultats obtenus avec les deux stratégies proposées auparavant sont présentés dans la suite.

5.3 Résultats expérimentaux

La méthode de détection de scènes proposée a été testée sur plusieurs films d'animation du CICA [CICA 06]. Nous allons présenter un exemple de détection réalisé sur un extrait du film "A Crushed World". La séquence de test a une durée totale de 2min21s et contient 3546 images. Elle a été segmentée manuellement en 11 plans, en introduisant parfois, volontairement, une sur-segmentation pour pouvoir tester pleinement notre méthode. Les 11 plans de l'extrait du film sont présentés dans la Figure 5.5 (chaque plan est représenté par l'image correspondant au milieu du plan).

$plan_0$ (scène 0) $plan_1$ (scène 1) $plan_2$ (scène 1) $plan_3$ (scène 2)

$plan_4$ (scène 3) $plan_5$ (scène 4) $plan_6$ (scène 5) $plan_7$ (scène 6)

$plan_8$ (scène 7) $plan_9$ (scène 6) $plan_{10}$ (scène 6)

FIG. 5.5 – La séquence de test (chaque plan est résumé par une image).

Les nombres d'images contenues dans chacun des 11 plans sont les suivants : $plan_0 = 281$ images, $plan_1 = 300$ images, $plan_2 = 265$ images, $plan_3 = 253$ images, $plan_4 = 297$ images, $plan_5 = 568$ images, $plan_6 = 489$ images, $plan_7 = 203$ images, $plan_8 = 255$ images, $plan_9 = 300$ images et $plan_{10} = 340$ images. La labellisation en scènes est la suivante : $plan_1$ et $plan_2$ font partie de la même scène (scène 1), de même que les plans $plan_7$, $plan_9$ et $plan_{10}$ qui font partie de la même scène (scène 6).

Pour la détection nous avons utilisé les paramètres suivants : le pourcentage d'images

retenues de chaque plan a été fixé à $p = 20\%$ et le seuil de détection est $\tau_{scene} = 0.1$. Les tableaux des distances $D(i, j)$ $(i, j = 0, ..., 10)$ obtenues par les deux stratégies de calcul de l'histogramme moyen sont présentés dans les Tableaux 5.1 et 5.2 (les distances inférieures au seuil de détection sont marquées en gras).

plan	0	1	2	3	4	5	6	7	8	9	10
0	0	0,75	0,75	0,66	0,42	0,5	0,49	0,69	0,69	0,69	0,69
1	∼	0	**0,01**	0,54	0,58	0,55	0,54	0,53	0,55	0,52	0,53
2	∼	∼	0	0,54	0,58	0,55	0,55	0,53	0,55	0,53	0,53
3	∼	∼	∼	0	0,3	0,26	0,28	0,32	0,2	0,31	0,31
4	∼	∼	∼	∼	0	0,12	0,12	0,34	0,29	0,34	0,33
5	∼	∼	∼	∼	∼	0	**0,07**	0,25	0,22	0,25	0,24
6	∼	∼	∼	∼	∼	∼	0	0,24	0,23	0,24	0,23
7	∼	∼	∼	∼	∼	∼	∼	0	0,21	**0,06**	**0,07**
8	∼	∼	∼	∼	∼	∼	∼	∼	0	0,20	0,2
9	∼	∼	∼	∼	∼	∼	∼	∼	∼	0	**0,01**
10	∼	∼	∼	∼	∼	∼	∼	∼	∼	∼	0

TAB. 5.1 – Le tableau $D(i, j)$ des distances entre les histogrammes moyens classiques, $\bar{h}_i(c)$ où i est l'indice du plan.

plan	0	1	2	3	4	5	6	7	8	9	10
0	0	0,72	0,72	0,63	0,4	0,47	0,46	0,65	0,65	0,65	0,65
1	∼	0	**0,01**	0,51	0,55	0,53	0,52	0,5	0,53	0,5	0,5
2	∼	∼	0	0,51	0,55	0,53	0,52	0,5	0,53	0,5	0,5
3	∼	∼	∼	0	0,29	0,25	0,27	0,28	0,18	0,28	0,28
4	∼	∼	∼	∼	0	**0,09**	0,1	0,32	0,27	0,3	0,31
5	∼	∼	∼	∼	∼	0	**0,06**	0,24	0,2	0,23	0,23
6	∼	∼	∼	∼	∼	∼	0	0,23	0,22	0,22	0,22
7	∼	∼	∼	∼	∼	∼	∼	0	0,19	**0,06**	**0,06**
8	∼	∼	∼	∼	∼	∼	∼	∼	0	0,19	0,18
9	∼	∼	∼	∼	∼	∼	∼	∼	∼	0	**0,01**
10	∼	∼	∼	∼	∼	∼	∼	∼	∼	∼	0

TAB. 5.2 – Le tableau $D(i, j)$ des distances entre les histogrammes moyens pondérés, $\bar{h}_i^*(c)$ où i est l'indice du plan.

En utilisant les histogrammes moyens classiques (voir le Tableau 5.1) nous avons obtenu les plans similaires suivants :

- $plan_1$ et $plan_2$ avec une distance *0.01*,
- $plan_5$ et $plan_6$ avec une distance *0.07*,
- $plan_7$ et $plan_9$ avec une distance *0.06*,
- $plan_7$ et $plan_{10}$ avec une distance *0.07*,
- $plan_9$ et $plan_{10}$ avec une distance *0.01*.

Une seule erreur de détection est survenue à cause de la proximité des couleurs de cer-

tains plans : le $plan_5$ et le $plan_6$ ont été considérés comme étant similaires alors qu'ils sont différents. Les plans similaires se regroupent en trois scènes : scène A contenant les plans 1 et 2, scène B (fausse détection) contenant les plans 5 et 6 et scène C contenant les plans 7, 9 et 10.

En utilisant les histogrammes moyens pondérés (voir le Tableau 5.2) nous avons obtenu les plans similaires suivants :

- $plan_1$ et $plan_2$ avec une distance *0.01*,
- $plan_4$ et $plan_5$ avec une distance *0.09*,
- $plan_5$ et $plan_6$ avec une distance *0.06*,
- $plan_7$ et $plan_9$ avec une distance *0.06*,
- $plan_7$ et $plan_{10}$ avec une distance *0.06*,
- $plan_9$ et $plan_{10}$ avec une distance *0.01*,

Deux erreurs de détection sont survenues : $plan_4$ et $plan_5$, puis $plan_5$ et $plan_6$ ont été considérés comme étant similaires alors qu'ils sont différents. Les plans similaires se regroupent en trois scènes : scène A contenant les plans 1 et 2, scène B (fausse détection) contenant les plans 4, 5 et 6 et scène C contenant les plans 7, 9 et 10.

Globalement, les histogrammes moyens pondérés ont donné des distances plus faibles que les histogrammes moyens classiques, les écarts de couleur étant souvent plus importants sur les contours des objets. Les techniques d'animations utilisent parfois des matériaux tels que le papier, la pâte à modeler, des figurines, du sable, etc., ce qui limite la distribution des couleurs à une certaine palette qui change peu pendant toute la durée du film. Avec ces techniques, les images sont généralement très texturées (fort gradient), et les histogrammes pondérés, en mettant un poids important sur les zones homogènes, accentuent la ressemblance entre les plans. C'est pourquoi nous obtenons donc de meilleurs résultats en utilisant les histogrammes moyens classiques. Ces résultats demanderaient à être confirmés par des tests plus variés.

L'utilité de l'approche proposée ne se limite pas seulement au découpage en scènes de la séquence. D'autres applications de l'analyse des similarités entre plans vidéo peuvent être envisagées, comme nous allons le voir dans la section suivante.

5.4 Les applications du découpage en scènes

5.4.1 La technique de la caméra "shot-reverse-shot"

Dans un premier temps la détection des scènes et particulièrement l'analyse des similarités entre les plans vidéo, permet la détection d'une technique de caméra appelée *"shot-reverse-shot"* [Bimbo 99]. Ce type de technique est particulièrement utilisé dans des scènes plutôt statiques de la séquence, comme par exemple les scènes de conversation, les scènes spécifiques au jeu de poker ou de billard (où la caméra se focalise alternativement sur la table de jeu, sur les joueurs ou le public).

En général, plusieurs caméras sont utilisées pour filmer une scène et ces caméras sont utilisées alternativement. La scène est ainsi constituée d'un plan filmé par la caméra 1, puis d'un plan filmé par la caméra 2, puis l'on revient sur la caméra 1 puis sur la caméra 2 et ainsi de suite. La façon classique de produire cet effet est d'utiliser une caméra filmant chaque acteur ou chaque point d'intérêt de la scène. A chaque fois qu'un acteur prend la parole, la caméra qui se situe devant lui commence à le filmer et la caméra qui était auparavant active

est alors arrêtée.

Nous retrouvons ce type de technique dans certaines scènes de films d'animation. Par exemple, dans la séquence de test que nous avons utilisée (voir Figure 5.5), nous retrouvons un "shot-reverse-shot" entre les plans 7 et 9. La caméra est positionnée sur un décor d'une scène d'intérieur dans le $plan_7$. Dans le $plan_8$ elle change de scène en filmant un décor d'extérieur, et revient sur le décor précédent dans le $plan_9$.

Les "shot-reverse-shot" introduisent une périodicité dans la séquence. La succession des plans est caractérisée par une structure de type $ABABAB$, où A et B sont deux labels de plans différents du point de vue du contenu. En utilisant le tableau des distances, $D(i, j)$, nous pouvons facilement détecter les "shot-reverse-shot" en retrouvant les structures caractérisées par des valeurs des distances entre plans du type : *valeur faible, valeur forte, valeur faible, valeur forte.*

Dans les films d'animation ce type de technique de caméra a une signification sémantique particulière. Si elle est répétée plusieurs fois, cela se traduit par une action à l'intérieur de la même scène, et peut servir à la caractérisation du rythme de la séquence. De plus, les "shot-reverse-shot" peuvent être utilisés pour déterminer les passages de dialogues entre les personnages où la caméra se focalise alternativement sur les différents personnages.

5.4.2 La correction de la détection des "cuts"

La détection de scènes peut également servir à améliorer le découpage en plans de la séquence et particulièrement pour la détection des "cuts". En effet, lors de la détection des transitions vidéo, nous obtenons certaines fausses détections. Une fausse détection se traduit par la détection d'un ou plusieurs changements de plan à l'intérieur d'un même plan. Elle a comme effet une sur-segmentation temporelle du plan concerné. Les plans ainsi obtenus sont similaires, et appartiennent donc au même plan.

En utilisant la détection des scènes, ces situations se traduisent par des scènes composées de plans voisins temporellement. Donc, si nous trouvons que deux plans voisins font partie de la même scène nous pouvons alors les fusionner car ils résultent d'une sur-segmentation temporelle. Dans la séquence de test, nous rencontrons cette situation à deux reprises (voir Figure 5.5), pour les plans 1 et 2, et les plans 9 et 10, qui ont été divisés, artificiellement ici pour les besoins du test, par le découpage en plans. La détection des scènes a détecté que ce sont des plans similaires, et par conséquent ils constituent un seul et même plan.

L'intérêt de cette correction vient du fait que les mesures de dissimilarité sont maintenant effectuées entre segments, et non plus entre images. Ainsi, tout en gardant la même mesure de dissimilarité et le même seuil, on aboutit à une détection plus robuste.

5.4.3 Répresentation hierarchique du contenu

La méthode de détection de scènes que nous avons proposée peut servir comme point de départ pour générer automatiquement des résumés de films sur plusieurs niveaux de détails. Le résumé d'un film est une représentation compacte de son contenu. Les méthodes de calcul de résumés utilisent souvent comme information de départ la segmentation en plans vidéo du film. Le principe s'appuie sur la représentation de chaque plan par un certain nombre d'images représentatives, appelées d'images clés (les techniques de calcul des résumés seront présentées dans le chapitre suivant). Ce type d'approche ne propose pas de hiérarchie, si bien que l'utilisateur ne peut pas choisir le niveau de détails du résumé qu'il souhaiterait obtenir.

En utilisant le tableau des distances entre les plans (tableau que nous avons utilisé pour la détection des scènes) nous pouvons arriver à une représentation compacte multi-échelle du contenu de la séquence. Dans cette représentation, le contenu du film est présenté sur plusieurs niveaux de détails (représentation qui est similaire à la structure temporelle présentée dans la Figure 2.1 de la Section 2.1). Les deux premiers niveaux sont le niveau des plans vidéo et le niveau des scènes. Le niveau de détails supérieur est obtenu en regroupant les scènes similaires, en utilisant le même mécanisme que celui qui a permis le regroupement des plans similaires. Il suffit pour cela d'assouplir le critère de similarité. Ce processus peut être réitéré de façon à obtenir la séquence entière (voir Figure 5.6).

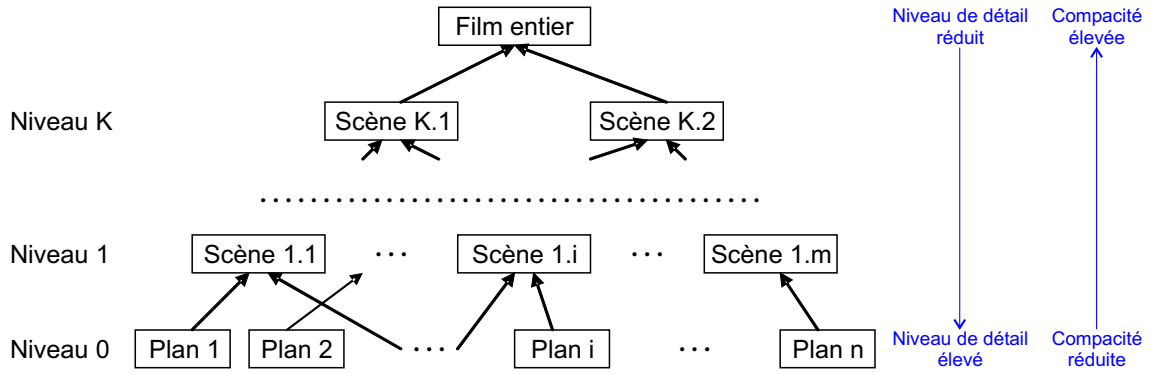


FIG. 5.6 – Représentation hiérarchique du contenu de la séquence à travers le découpage en scènes.

Cette représentation hiérarchique du contenu nous permet de résumer le film de façon à répondre conjointement aux différentes contraintes sur le niveau de détails et la durée de résumé désirée. Par exemple, le film peut être résumé en gardant uniquement un certain nombre d'images clés de chaque scène du niveau de détails K. En fonction de l'application nous pouvons ainsi choisir entre des résumés ayant un niveau de détails élevé mais au détriment d'une durée plus longue (beaucoup d'images clés), ou au contraire un niveau de détails plus faible afin d'obtenir un résumé très compact (juste quelques images clés résumant le film entier). Cette approche, en utilisant une méthodologie différente, rejoint les techniques de construction automatique de résumés présentées dans la Section 6.

5.5 Conclusions générales

Dans ce chapitre nous avons présenté les différentes techniques qui permettent la détection de scènes à partir des films. Le concept de scène est défini comme étant l'ensemble des plans qui respectent la règle des trois unités, l'unité *de temps*, l'unité *de lieu*, et l'unité *d'action*. Ces propriétés sont liées à la perception sémantique de la séquence, et pour le moment, les méthodes existantes ont beaucoup de mal à accéder à la compréhension du contenu des scènes.

Les approches trouvées dans la littérature sont essentiellement basées sur des mesures de similarités entre les différentes caractéristiques des plans vidéo et sur la définition d'un certain nombre de règles de production qui permettent l'identification des scènes à l'intérieur de la séquence. Elles n'utilisent pas l'information sémantique du contenu. D'autre part, le

concept de scène est lui même relatif à la perception, car pour la même séquence plusieurs distributions de plans en scènes sont possibles, ce qui rend l'évaluation des résultats subjective.

La méthode que nous proposons réalise la détection des scènes en se basant sur l'unité de contenu, traduite par la ressemblance des plans au niveau des couleurs, et sur l'unité de temps, concrétisée par la proximité temporelle des plans appartenant à la même scène. La distribution des couleurs de chaque plan est résumée par un histogramme couleur moyen. Deux stratégies ont été testées, la première utilise des histogrammes couleurs classiques et la deuxième utilise des histogrammes pondérés, la pondération étant inversement liée au gradient de l'image. On peut noter cependant que les histogrammes couleurs pondérés sont plus sensibles aux textures (zones de fort gradient) et donnent des résultats moins bons dans le cas de films d'animation utilisant des techniques telles que la pâte à modeler, le papier, le sable, etc. (où la texture des matériaux ne change pas significativement pendant le film).

En ce qui concerne la complexité de calcul, une implémentation en temps réel est envisageable car les traitements les plus importants à effectuer se limitent à la réduction des couleurs dans l'image et au calcul de l'histogramme couleur. En pratique le découpage en scènes peut être réalisé en même temps que le découpage en plans (Section 2.4). Les deux méthodes utilisent en effet les mêmes mesures qui sont les distances entre les histogrammes couleurs des images.

Globalement, cette méthode de détection de scènes est encore loin de la perception humaine mais elle peut servir à la segmentation temporelle de plus haut niveau du film. De plus la détection de scènes permet diverses applications parmi lesquelles nous pouvons citer la détection de la technique "shot-reverse-shot", l'amélioration du découpage en plan ou la représentation hiérarchique du contenu des films.

Chapitre 6

La construction des résumés

Résumé : *Accéder au contenu d'une vaste collection de films est une tâche difficile. Cela demande des méthodes de représentation efficace du contenu de chacun des films, qui se traduisent par la construction de résumés. Parmi les approches existantes, on trouve deux catégories de résumé : les résumés statiques, construits à partir d'images, et les résumés dynamiques, construits à partir de sous-séquences d'images. Dans ce chapitre nous présentons différentes techniques de construction automatique de résumés de séquences d'images. Pour les résumés statiques, nous étudions différentes méthodes basées sur l'analyse de l'activité à l'intérieur des plans vidéo. Ensuite, à travers l'analyse du contenu de l'action de la séquence nous proposons une méthode de construction d'un résumé dynamique de type "bande-annonce". De par sa subjectivité, l'évaluation d'un résumé est généralement une tâche délicate. La qualité de nos approches a été évaluée par la mise en place d'une campagne d'évaluation.*

Les bases de séquences d'images contiennent un volume de données gigantesque, typiquement quelques milliers de films. Une seule minute de film étant équivalente à 1500 images (à une cadence de 25 images/s), on peut imaginer la masse de données que représentent une base de films. Pour accéder au contenu de la base, l'utilisateur a souvent besoin de visualiser les séquences. Le problème est simple si on dispose de quelques séquences, mais visualiser des milliers de séquences pour retrouver l'information recherchée (par exemple l'extrait d'un film particulier) est une tâche presque impossible. Il est donc nécessaire de disposer d'une représentation compacte et efficace du contenu des séquences. La solution est la construction de *résumés*.

Le résumé d'une séquence est une représentation compacte du contenu, d'une durée beaucoup plus courte [Li 01], ou, plus précisément, *un résumé vidéo est une séquence réduite d'images fixes (collection d'images) ou en mouvement (collection de sous-séquences) représentant le contenu de la séquence de telle manière que l'utilisateur soit informé rapidement et de façon concise sur le contenu, l'essentiel de ce contenu étant préservé dans le résumé* [Pfeiffer 96].

L'intérêt d'avoir une représentation compacte du contenu ne se limite pas seulement à

la navigation (visualisation du contenu) et à la recherche dans une base de films. Le résumé permet aussi de diminuer le temps de calcul des méthodes de caractérisation et d'analyse, en réduisant la masse de données à traiter. En effet, la plupart des techniques d'analyse de séquences d'images n'utilisent pas la séquence dans son intégralité, mais seulement un certain nombre d'images représentatives. Ces images sont choisies de manière à ce que l'information exploitée ne soit pas trop altérée. Par exemple, pour analyser la distribution des couleurs dans la séquence, l'utilisation d'une seule image représentative de chaque plan vidéo peut suffire (voir la Section 4.2.4).

Un résumé peut être généré *manuellement*, de manière *semi-automatique*, ou encore *automatiquement*. Étant donné l'énorme volume de données et la main d'œuvre souvent limitée, il est de plus en plus important de développer des outils complètement automatisés de manière à réduire l'implication humaine dans ce processus.

6.1 État de l'art

Dans la littérature spécialisée nous retrouvons essentiellement deux catégories de résumé :

- le **résumé en images** ou *résumé statique* : sorte de "storyboard" simplifié, c'est une collection d'images représentatives, appelées images clés, de la séquence. Le résumé en images est aussi appelé "video summary".
- le **résumé en mouvement** ou *résumé dynamique* : le résumé dynamique, ou "video skim", est une collection de sous-séquences d'images formant une nouvelle séquence d'une durée fortement réduite par rapport à la séquence originale. Dans ce dernier type de résumé, le son est souvent présent.

Ces deux types de résumés présentent chacun des avantages et des inconvénients. Les principaux avantages *d'un résumé en images* sont les suivants :

- il peut être généré rapidement car il utilise seulement l'information visuelle (le son et le texte ne sont pas directement présents dans ce résumé).
- il peut être visualisé facilement car il n'a pas besoin de synchronisation ou de temporisation des données (entre le son et l'image par exemple).
- il peut faciliter la représentation du contenu de la séquence à travers des images "mosaïques" [Aner 01].
- il peut être imprimé, pour servir par exemple de "storyboard" de la séquence.
- il permet la réduction de la complexité de calcul pour certaines méthodes d'analyse en n'effectuant les traitements que sur les images du résumé.

En comparaison, *le résumé dynamique* présente quelques avantages essentiels :

- il possède généralement plus de sens car il contient une information de mouvement, information perdue dans le résumé en images. Et dans certains films, l'information de mouvement est l'information clé de la séquence, en particulier dans le domaine des films d'animation où le rythme de déroulement des événements est une caractéristique essentielle.
- des informations audio (musique, dialogue, etc.) peuvent être conservées dans le résumé
- sa visualisation est plus naturelle et plus attractive. Il est en effet plus agréable pour l'utilisateur de regarder la "bande-annonce" d'un film que de regarder une succession saccadée d'images statiques [Li 01].

En contrepartie, il est souvent d'une complexité de construction plus élevée et sa génération est plus délicate, en particulier du fait de la nécessité de synchroniser le son et l'image.

En fait, les deux types de résumés sont nécessaires pour la caractérisation du contenu de la séquence, chacun apportant des informations distinctes : le résumé en image permet d'avoir une représentation rapide (en quelques images) du contenu visuel de la séquence et le résumé dynamique permet d'avoir une représentation compacte et efficace du contenu dynamique de la séquence.

Enfin, même si ce n'est pas forcément la meilleure stratégie, chaque type de résumé peut être généré à partir de l'autre. Un résumé dynamique peut être créé à partir d'un résumé en images en récupérant autour de chaque image clé une sous-séquence d'images. De la même façon, un résumé en images peut être généré à partir d'un résumé dynamique en prélevant certaines images des sous-séquences.

Dans la littérature du domaine, différents états de l'art sur les techniques d'extraction de résumés ont été proposés [Li 01] [Truong 06].

6.1.1 Les résumés en images

Comme nous l'avons déjà mentionné dans la section précédente, un résumé statique est une collection d'images considérées comme représentatives du contenu de la séquence. Ces images sont les images *clés*. Formellement ce résumé est défini de la manière suivante :

$$R_{img}(S) = \{image_1, image_2, ..., image_N\} \tag{6.1}$$

où R_{img} est le résumé de la séquence S, contenant les images clés $image_i$, avec $i = 1, ..., N$, N est le nombre total d'images du résumé.

La valeur du paramètre N joue un rôle important sur la qualité du résumé. Si N est connu *a priori*, la taille du résumé est utilisée comme contrainte de départ pour l'algorithme d'extraction. Par contre, si N n'est pas fixé a priori, c'est à l'algorithme de choisir automatiquement le nombre d'images du résumé. Ce nombre sera adapté au contenu de chaque séquence (plus d'images sont nécessaires pour représenter un contenu riche en action).

Les méthodes existantes d'extraction de résumés en images peuvent être classifiées en fonction de la manière dont les images clés sont extraites de la séquence. [Li 01] propose quatre catégories :

- *l'extraction par échantillonnage*,
- *l'extraction au niveau des plans*,
- *l'extraction au niveau des segments*,
- *les autres approches*.

L'extraction par échantillonnage

L'extraction d'images clés par échantillonnage consiste à sélectionner des images en effectuant un prélèvement uniforme ou aléatoire dans la séquence initiale [Taniguchi 95].

L'avantage de cette méthode est sa simplicité, mais, en revanche, le résumé obtenu n'est pas forcément représentatif des moments importants de la séquence. Par exemple, certains plans, d'une durée réduite, mais importants pour le contenu de la séquence, risquent de ne pas être représentés par une image clé, tandis que certains plans longs auront plusieurs images clés avec un contenu similaire.

L'extraction au niveau des plans

Une méthode plus élaborée est l'extraction d'images clés au niveau des plans vidéo. Dans une version simplifiée de cette approche, l'extraction des image clés s'effectue en sélectionnant arbitrairement une des images de chaque plan (par exemple la première, la dernière, l'image centrale, etc.). Cette approche est suffisante pour des plans dont le contenu varie peu, mais une seule image clé ne permet pas forcément une bonne représentation visuelle d'un plan présentant une forte variabilité. Aussi, la plupart des travaux existants dans le domaine ont fait le choix d'interpréter le contenu en employant différentes caractéristiques de bas niveau extraites des plans vidéo.

L'extraction d'images clés par *analyse statistique couleur* a été largement utilisée du fait de l'efficacité et de la robustesse des histogrammes couleurs. Par exemple [Zhang 97] utilise le seuillage des différences entre les histogrammes couleurs des images successives pour extraire les images les moins semblables d'un plan, [Zhuang 98] propose d'extraire les images clés en utilisant une technique de classification non-supervisée où la mesure de dissimilarité est l'écart entre histogrammes couleurs. Cependant, la plupart de ces méthodes ont du mal à capturer la dynamique lorsqu'il y a beaucoup de mouvements d'objets et de caméra. De plus, elles dépendent de seuils sur les mesures d'écart entre histogrammes.

Les approches basées sur *l'analyse du mouvement* sont plus adaptées pour contrôler le nombre d'images clés en fonction de la dynamique temporelle de la scène. Un exemple est la méthode proposée dans [Wolf 96] où les images clés sont sélectionnées comme les minima locaux d'une fonction temporelle mesurant le mouvement de la séquence. Cependant, cette approche présente quelques inconvénients. D'abord, il est difficile de juger l'importance d'un passage en se basant uniquement sur des critères de dynamique du mouvement. Ensuite, d'un point de vue technique, on risque de fournir des images floues de transition d'un mouvement rapide.

Une autre approche qui utilise l'information de mouvement est la construction *d'images "mosaïques"*. Ce sont des images panoramiques représentant le contenu entier d'un plan ou d'un segment de la séquence [Irani 95][Aner 01]. Cette approche a l'avantage de capturer la dynamique de la scène dans une seule image. Par contre, la construction des images "mosaïques" n'a de sens que sur des passages contenant un mouvement global de la caméra. De plus elle ne peut pas être appliquée aux segments statiques de la séquence ou aux segments comportant des mouvements complexes de caméra. La solution est alors d'utiliser une image "mosaïque" lorsque cela est possible et d'extraire des images clés représentatives dans les autres cas.

On trouve enfin un certain nombre d'*autres approches* mixtes qui utilisent la collaboration entre différentes caractéristiques de bas niveau (couleur, mouvement, etc.), en profitant de l'avantage offert par chaque catégorie d'information utilisée. Un exemple est la méthode proposée en [Doulamis 00a], où les images clés sont extraites en estimant les points appropriés de la courbe formée à partir de vecteurs de caractéristiques de chaque image. Les vecteurs de caractéristiques sont obtenus lors d'un processus de segmentation appliqué au niveau des couleurs et du mouvement. ces méthodes, plus lourdes à mettre en œuvre, fournissent généralement de meilleurs résultats.

L'extraction au niveau des segments

Le problème avec l'utilisation d'une ou plusieurs images clés sélectionnées pour chaque plan est que cette approche n'est pas appropriée pour des séquences longues comportant beaucoup de plans. Regarder un résumé comportant des centaines d'images reste une tâche longue et fastidieuse. De plus en plus de chercheurs travaillent donc sur des unités vidéo de plus haut niveau, appelés segments vidéo : les scènes, les épisodes, les événements, ou encore la séquence vidéo entière [Sun 00][Doulamis 00b].

L'approche proposée dans [Sun 00] utilise ainsi une segmentation uniforme de la séquence en unités longues. Pour chaque unité une mesure de changement est calculée pour servir à la classification de toutes les unités en deux catégories : les unités avec changements faibles et les unités avec changement forts. Les images clés sont ensuite extraites en utilisant des règles adaptées à la catégorie et au contenu de chaque unité.

Un autre exemple est l'approche proposée dans [Doulamis 00b] où l'extraction d'images clés est effectuée au niveau de la séquence entière. Elle utilise une méthode de classification floue en classes prédéfinies selon des caractéristiques de couleur et de mouvement de chaque image de la séquence. Les images clés sont alors extraites à l'aide d'un algorithme génétique.

Le problème majeur avec ce type d'approche est la complexité élevée des algorithmes de classification utilisés. D'autre part, dans les résumés ainsi obtenus la chronologie temporelle est perdue car les images extraites sont le résultat d'une classification globale des images de la séquence. Ce type d'approche est cependant bien adapté à la construction d'un résumé global compact de la séquence avec un nombre réduit d'images.

Autres travaux

D'autres méthodes d'extraction d'images clés utilisent la localisation de certains passages importants, comme la présence de visages [Dufaux 00] ou d'un nombre important d'objets [Kim 00a]. Les images clés sont alors extraites à partir de ces passages. La principale contrainte de ce type d'approche est sa dépendance aux particularités de chaque domaine d'application, ou même de chaque film. En effet, les images clés sont extraites en utilisant des règles spécifiques au genre des séquences analysées, ce qui ne confère pas à ces approches de caractère générique.

6.1.2 Les résumés dynamiques

Le résumé dynamique ou "video skim", est une collection de segments audio-visuels extraits de la séquence. Il est lui-même un document vidéo. Formellement le résumé dynamique, $R_{seq}(S)$, de la séquence S est défini par :

$$R_{seq}(S) = seg_1 \cup seg_2 \cup ... \cup seg_M \tag{6.2}$$

où seg_i est un segment vidéo, $i = 1, ..., M$ avec M le nombre total de segments du résumé.

Par rapport au résumé en images, le résumé dynamique a une complexité de construction généralement plus élevée car il demande une analyse de plus haut niveau du contenu. Dans ce cas l'unité de base à traiter n'est pas l'image mais le segment vidéo (sous-séquence d'images).

Comme nous l'avons déjà mentionné au début du paragraphe 6.1, le résumé en images peut servir à l'extraction du résumé dynamique. Une méthode immédiate consiste à remplacer chaque image clé du résumé par un intervalle d'images, centré par exemple autour de l'image

clé. Le résumé ainsi obtenu est une représentation compacte de la séquence mais il est dépendant de la qualité du résumé en images et, comme pour le résumé en images, il n'est pas forcément très représentatif du contenu dynamique de la séquence. Il est souvent préférable d'extraire le résumé dynamique directement de la séquence originale.

Une autre méthode d'extraction du résumé dynamique est l'utilisation directe de la segmentation en plans vidéo. Dans ce cas le contenu de la séquence peut être résumé en gardant de chaque plan vidéo une sous-séquence d'images. Le résumé ainsi obtenu est une représentation de l'ensemble de la séquence, incluant aussi bien des plans importants que des plans moins importants, ce qui aboutit souvent à un résumé de durée trop élevée.

Un "bon" résumé dynamique nécessite une compréhension de plus haut niveau du contenu de la séquence, une compréhension sémantique. Les limitations des méthodes d'analyse sémantique des images ont dédié les méthodes d'extraction de résumés dynamiques à des domaines spécifiques, comme *le sport* [Coldefy 04], *les documentaires* [Yu 03], *les vidéos personnelles* [Zhao 03], etc. La difficulté d'analyse est simplifiée par l'utilisation d'informations *a priori* sur le domaine visé.

Enfin, il faut noter que la littérature sur l'extraction des résumés dynamiques est moins riche que celle sur le résumé en images [Li 01]. Pour un état de l'art complet, sur les techniques existantes d'extraction de résumés, on peut se rapporter à [Truong 06].

L'information à préserver dans le résumé

En premier lieu, la construction d'un résumé dynamique demande de définir l'information que l'on veut préserver dans ce résumé. Ce choix dépend du domaine d'application visé, et va déterminer la façon dont le résumé sera généré.

Parmi les méthodes existantes, en fonction du contenu désiré, on retrouve trois catégories distinctes [Truong 06] :

- les résumés qui couvrent *tout le contenu* de la séquence,
- les résumés qui ne reproduisent que certains *événements importants* de la séquence,
- les résumés *personnalisés* par interaction avec l'utilisateur.

Couverture totale du contenu. Les résumés qui *couvrent tout le contenu* de la séquence ont comme but de transmettre à l'utilisateur des informations générales sur le contenu global de la séquence [Sundaram 02] [Gong 03]. Dans ce cas, la compréhension du contenu original n'est pas altérée par le résumé. Ce type de résumé répond aux situations où l'utilisateur recherche un aperçu dynamique complet et efficace de la séquence entière. Le temps nécessaire pour la visualisation de ce type de résumé, généralement important, est compensé par le caractère complet de l'information fournie.

Les événements importants. Les résumés ne reproduisant qu'un certain nombre d'*événements importants* de la séquence ("video highlights") sont les plus utilisés. Il sont généralement adaptés aux particularités d'un domaine d'application. Les différents travaux proposés pour la construction des "video highlights" sont groupés selon le type d'événements à préserver [Truong 06] :

- les événements entraînant des réactions particulières de l'audience : applaudissements et encouragements [Xiong 03],
- les passages de la séquence provoquant l'enthousiasme du narrateur [Coldefy 04],

- les passages de la séquence mis en évidence par le producteur à travers des techniques de montage spécifiques : une fréquence élevée de "cuts", la présence de texte ou la reprise de certaines scènes de la séquence [Pan 01],
- les événements correspondant à un modèle prédéfini (par exemple un événement rare) [Radhakrishnan 04],
- les passages de la séquence préférés par l'utilisateur (par exemple ceux visualisés plusieurs fois) [Yu 03].

Parmi ces résumés "video highlights", il en existe un, la bande-annonce ("movie trailer"), qui ne résume que certain passages particulièrement captivants ou riches en action. Trouver ces passages est un processus subjectif et difficile, et les techniques existantes utilisent souvent des hypothèses *a priori* liées au domaine d'application considéré. Par exemple dans les matchs de football, on sait que les événements les plus captivants sont les buts.

Personnalisation du contenu. Une autre catégorie de résumés est celle qui utilise *la personnalisation du contenu*. Ces résumés sont générés en fonction de la préférence de l'utilisateur sur le contenu à préserver. Cette préférence est manifestée par l'utilisateur sous la forme d'une demande ("query") ou en choisissant un modèle de contenu choisi dans une liste prédéfinie. Par exemple, dans [Lu 03] (domaine des séquences d'informations) les options disponibles sont *la présence de visages, de parole, le zoom de la caméra, la présence de texte*. Dans [Li 03] les événements utilisés sont *les dialogues entre deux personnes, les dialogues entre plusieurs personnes* ou *les scènes hybrides*.

Dans ce cas le processus d'extraction de segments pertinents est simplifié. Le résumé est créé en ne retenant que les passages de la séquence qui sont en concordance avec les demandes de l'utilisateur. Ce type de résumé peut être considéré comme un résumé semi-automatique car l'intervention de l'utilisateur est demandée. Cette approche, de par son lien avec un domaine d'application bien précis, ne présente pas de caractère générique et sera difficilement utilisable dans un contexte non connu.

Le processus d'extraction

D'une manière générale le processus d'extraction d'un résumé dynamique comporte cinq étapes [Truong 06] :

- *le découpage* en segments de la séquence,
- *la sélection de segments*,
- *la réduction* de la taille des segments retenus,
- *l'intégration* multimodale,
- *la construction* du résumé.

Dans la pratique il arrive que certaines étapes ne soient pas présentes ou soient fusionnées entre elles.

Le découpage en segments. La première étape, indispensable pour la construction d'un résumé dynamique est *le découpage en segments*. Un *segment* peut être un plan, une scène, un épisode ou un passage de la séquence contenant un événement présentant un intérêt particulier. La notion de segment n'est pas limitée à l'information visuelle mais elle fait également référence aux autres modalités de la séquence : le son et le texte.

La sélection des segments. L'étape suivante est la *sélection des segments* qui seront utilisés pour le résumé. Cette sélection est faite à partir de l'ensemble de tous les segments de la séquence. La technique de sélection influencera la *cohérence*, la *couverture* et le *contexte* du résumé.

Différentes techniques ont été proposées. Par exemple, dans [Gong 03] une décomposition en valeurs singulières (SVD - "singular value decomposition") d'une matrice de caractéristiques couleurs extraites des images est utilisée pour caractériser les propriétés temporelles et spatiales de la séquence. En utilisant cette décomposition, il est possible d'extraire des plans à partir du degré de changement visuel, de l'uniformité de la distribution des couleurs et d'une mesure de similarité.

Une autre approche est proposée dans [Ngo 03]. Dans celle-ci, les plans et les classes sont obtenues à partir d'un algorithme généralisé de détection des "cuts". Puis ils sont analysés en utilisant un modèle de la perception humaine : "motion attention model". Ensuite, les valeurs obtenues de l'attention humaine sont structurées sous la forme d'un graphe, et sont traitées de la même façon que les chaînes de Markov. Le graphe est utilisé pour regrouper les classes similaires en scènes et l'attention est utilisée comme critère de sélection des sous-séquences servant à la construction du résumé.

La réduction de la taille des segments. Généralement les segments ainsi obtenus contiennent une information redondante et leur durée est souvent trop élevée pour le résumé. Une étape d'optimisation par la *réduction de la taille des segments* garantit un résumé plus concis tout en gardant l'essentiel de l'information. Cette étape peut cependant introduire des points de discontinuité visuelle dans le résumé final.

Différentes solutions ont été proposées pour réduire les segments. Par exemple, la méthode proposée dans [Cooper 02] utilise une matrice d'autosimilarité. De chaque segment, seul le passage continu le plus semblable au segment entier est retenu. D'autres approches s'appuient sur la compression des segments par effacement de certaines images redondantes, mais en modifiant en même temps l'information audio, tout en restant dans des limites de compréhension [Li 03].

L'intégration multimodale. Les segments dont on dispose sont typiquement unimodaux. Ce sont des segments visuels, audio ou textuels. *L'intégration multimodale* a comme but de fusionner toutes les modalités disponibles en respectant l'évolution temporelle de la séquence pour construire le résumé final. Dans l'intégration multimodale l'alignement des segments joue un rôle important. Les frontières des segments sélectionnés sont ajustées pour garder le flux, la cohérence et le contexte de la séquence. De plus la continuité du son est assurée en évitant les interruptions au milieu des phrases.

En fonction de la manière dont l'intégration audio et vidéo est effectuée, on retrouve deux catégories de résumés dynamiques : *synchrones* ou *asynchrones* [Truong 06].

Dans les résumés synchrones l'information visuelle est synchronisée avec le son en utilisant comme référence l'axe temporel de la séquence. Ce type de résumé est plus adapté aux films, car le son y est en correspondance directe avec les images qu'on visualise. Dans le cas où les segments on été générés séparément pour chaque modalité de la séquence, leur intégration peut être effectuée en utilisant les opérateurs *ou* et *et* [Erol 03] (par exemple on sélectionne les passages importants de la séquence qui sont contenus dans un segment audio *ou* dans un segment visuel).

Les résumés asynchrones sont plus adaptés au cas des documentaires et des informations de manière à maximiser la couverture de la séquence. Ce type de résumé est généré en utilisant seulement une des modalités de la séquence : l'image, le son ou le texte. Les autres modalités sont ajoutées ensuite. On trouve un exemple dans la méthode proposée dans [Smith 98] où le résumé est extrait en utilisant le son. Les éléments visuels sont ajoutés après en utilisant des règles heuristiques concernant les mouvements objets/caméra, la présence de visages ou la présence de texte incrusté dans l'image.

La construction du résumé. En ce qui concerne *la construction du résumé final*, la méthode généralement utilisée est l'agrégation de tous les segments en respectant l'évolution temporelle de la séquence. Un cas particulier est le résumé "bande-annonce" qui est un résumé contenant seulement les parties les plus spectaculaires de la séquence et qui a comme objectif de susciter un intérêt pour le film. Dans ce cas l'ordre temporel n'est pas toujours respecté.

6.1.3 L'évaluation des résumés

Un problème important est l'évaluation des résumés. La question qui se pose est : *le résumé que nous avons obtenu est-il pertinent ? Est-il cohérent pour la personne qui le regarde ?* Typiquement pour une même séquence plusieurs résumés corrects sont possibles. Le fait que le résumé ait été généré en conformité avec un certain nombre de critères objectifs, inspirés de la perception humaine, n'assure pas toujours la cohérence visuelle du résumé. Pour le moment, dans le domaine, *il n'y a pas encore de méthodologie efficace* d'évaluation des résumés. Chaque approche proposée a sa propre méthode d'évaluation, stratégie qui ne facilite pas les comparaisons avec les autres approches.

Si dans les autres domaines de recherche, comme par exemple la détection et la reconnaissance d'objets, on peut disposer de vérités terrain permettant l'évaluation quantitative des méthodes proposées, dans le cas de la construction de résumés *la vérité terrain est très subjective* et pratiquement inexistante. De plus, même si on dispose d'une référence, un second problème est la difficulté associée à la comparaison des résumés. Même pour l'être humain, il est souvent difficile de décider si un résumé est meilleur qu'un autre.

Les solutions proposant une méthodologie d'évaluation des résumés se divisent en trois catégories [Truong 06] :

- *l'analyse descriptive du résultat*,
- *l'emploi d'une certaine métrique*,
- *les tests d'évaluation*.

L'analyse descriptive du résultat

La méthode la plus simple, qui ne demande pas de comparaison avec d'autres résultats, est l'analyse descriptive du résultat obtenu. Dans [Yu 04], une technique d'extraction de résumé est ainsi appliquée sur un certain nombre de séquences et les résumés obtenus sont présentés et leur pertinence est discutée. D'autres approches essaient, toujours d'une manière descriptive, d'expliquer et d'illustrer les avantages ou la supériorité de telle méthode par rapport à d'autres méthodes [Vermaak 02] (voir Figure 6.1).

D'une manière générale, ce type d'évaluation est largement subjectif car il n'y a pas de justifications montrant que la technique proposée est efficace si elle est appliquée à d'autres

(a)

(b)

FIG. 6.1 – Exemple d'évaluation des résultats proposés dans [Vermaak 02] : (a) Les images clés obtenues avec le critère proposé (BIC - Bayes Information Criterion), (b) Les images obtenues par sous échantillonnage uniforme de la séquence (les lignes rouges marquent les changements de plan).

séquences que celles testées ou présentées. Les résultats expérimentaux proposés dans ce cas sont insuffisants pour un jugement général. De plus cette méthode est peu utilisable dans le cas de résumés dynamiques car le volume de données est trop élevé pour permettre une analyse descriptive.

L'emploi d'une métrique

Dans le cas des *résumés en images*, la métrique utilisée pour l'évaluation est typiquement une fonction de fidélité calculée entre l'ensemble des images clés du résumé d'une part et la séquence d'autre part. Cette métrique est utilisée pour comparer les résultats obtenus avec différentes approches. Cependant, comme pour l'évaluation par la description du résultat, rien n'assure que cette mesure soit en conformité avec le jugement humain sur la qualité du résumé.

Comme exemple, on peut mentionner les travaux proposés dans [Liu 04] où la semi-distance de Hausdorff appliquée sur le score SRE (erreur de reconstruction de la séquence à partir du résumé) est utilisée pour comparer les méthodes proposées à six autre approches. [Liu 02b] propose le concept "d'image clé bien distribuée" pour l'évaluation des résumés. Selon les auteurs une image clé est "bien distribuée" si elle n'est pas redondante et si elle n'appartient pas à une transition vidéo. Pour évaluer la qualité d'un résumé, le nombre d'images "bien distribuées" est représenté graphiquement en fonction du nombre total d'images clés du résumé.

Pour les *résumés dynamiques* de type "video highlight", basés sur la localisation des événements importants de la séquence, l'évaluation de la qualité du résumé à plus de sens car une vérité terrain est plus facile à construire. En effet, la localisation d'événements importants est généralement consensuele. Alors un événement placé dans le résumé sera déclaré comme correct s'il se retrouve dans la vérité terrain [Xiong 03] [Ariki 03]. Dans ce cas la performance du résumé peut être évaluée en utilisant les taux de précision et rappel,

largement utilisés dans toutes les méthodes de détection.

D'autres approches, pour améliorer l'objectivité de l'évaluation, utilisent comme vérité terrain un certain nombre de résumés qui ont été créés manuellement par des spécialistes. Par exemple, [He 99] propose une technique d'extraction de résumé pour les séquences de conférences en utilisant comme référence les résumés qui ont été créés par les auteurs. [Miura 03], dans le domaine des programmes télévisés dédiés à la cuisine, utilise comme référence les commentaires fournis par le producteur et qui accompagnent le programme.

Généralement, ces méthodes d'évaluation sont plus objectives que les méthodes qui décrivent les résultats. En effet, elles tentent de s'appuyer sur une vérité terrain construite par un spécialiste et permettent une comparaison avec d'autres approches. La comparaison est réalisée en utilisant des opérateurs de distance entre résumé de référence et résumés proposés. Il faut noter que cette approche ne fait pas intervenir directement la perception humaine.

Les tests d'évaluation

Dans ce cas, l'évaluation de la pertinence d'un résumé est réalisée par l'homme. Un certain nombre de personnes (spécialistes ou non) sont désignées pour regarder les résumés proposés et pour donner leur avis sur la qualité de leur contenu, en répondant généralement à un questionnaire. Cette évaluation, malgré sa subjectivité, est *probablement la plus réaliste* car elle implique le "consommateur" du produit lui-même. Malheureusement, une campagne d'évaluation est difficile à mettre en place, aussi bien du point de vue logistique (préparation des films, mise en place du protocole d'évaluation, etc.) que du temps nécessaire pour la visualisation d'un ensemble de séquences et de leurs résumés.

Différentes approches ont été proposées. Pour les résumés en images on peut mentionner la méthode proposée dans [Dufaux 00] où chaque image clé du résumé est classée comme *bonne*, *correcte* ou *faible* par les utilisateurs. On aboutit ainsi à un score de satisfaction global pour chaque résumé. Une approche similaire, mais qui implique un examen plus complexe, est proposée dans [Liu 03]. Les images clés sont analysées au niveau de chaque plan de la séquence. Pour chaque plan, un certain nombre d'utilisateurs donnent un score de satisfaction, *bien*, *acceptable* ou *mauvais*, pour les images clés retenues.

Dans le cas des résumés dynamiques plusieurs approches sont utilisées. Dans la plupart des situations l'utilisateur doit donner directement son appréciation sur la qualité du résumé proposé. D'autres approches, plus élaborées, essayent d'apprécier dans quelle mesure les résumés proposés ont aidé dans des tâches interactives, comme la navigation ou la recherche [Ngo 03]. On peut également analyser les performances de l'utilisateur sur l'identification du contenu de la séquence à partir du résumé [Erol 03].

6.2 Les méthodes proposées

Dans ce chapitre nous proposons et analysons différentes méthodes d'extraction de résumés en images ou dynamiques. Chacun des résumés proposés joue un rôle précis dans le système d'analyse de séquences d'images proposé dans cette thèse. Le *résumé en images* est utile pour représenter d'une manière compacte le contenu visuel global de la séquence. Le *résumé dynamique* est une représentation compacte du contenu dynamique de la séquence, information perdue dans le résumé en images. Il permet de donner à l'utilisateur une idée de l'action

contenue dans la séquence.

6.2.1 Les résumés en images

Nous avons analysé et testé plusieurs techniques d'extraction de résumés statiques. D'abord nous avons étudié l'efficacité de l'approche classique utilisant une image clé par plan. Ensuite, nous avons envisagé une technique plus complexe (voir le rapport [Ott 05]) qui adapte le nombre d'images clés extraites de chaque plan en fonction de l'action qu'il contient. Enfin, nous proposons une technique de construction d'un résumé compact de la séquence entière avec seulement quelques images.

L'approche "une image par plan"

Le premier résumé en images analysé est construit à partir du découpage en plans vidéo, en ne retenant qu'une seule image clé par plan. Cette approche assure que les images extraites ne sont pas issues d'une transition vidéo et ensuite qu'elles suivent l'évolution temporelle de la séquence.

Pour choisir l'image clé de chaque plan, nous avons testé plusieurs stratégies :

- **l'image centrale** : en gardant l'image du milieu du plan, la probabilité de tomber sur la partie la plus représentative du plan est élevée. Cependant, il est possible de tomber sur un effet de couleurs (comme par exemple les SCC - "changements brefs de couleurs") ou sur une image de transition d'un mouvement rapide de caméra,
- **l'image de début/ l'image de la fin** : l'image de début du plan est généralement une image représentative car elle marque le début du changement de contenu. Pour prendre en compte l'éventuelle imprécision de détection des changements de plans et s'assurer que cette première image ne soit pas prise dans le plan précèdent ou dans une transition, l'image sélectionnée est choisie au delà d'un intervalle de sécurité. On peut utiliser la même stratégie avec l'image de fin, mais l'image de début est généralement plus intéressante,
- **une image aléatoire** : cette stratégie s'appuie sur la définition d'un plan, ensemble homogène d'images présentant une continuité spatiale, temporelle et de l'action. Dans ces conditions toute image du plan peut, en théorie, jouer le rôle d'image clé. Cette stratégie qui, sur le fond, n'a aucune validité est utilisée pour évaluer les résultats des autres stratégies.

Du point de vue de la mise en œuvre ces résumés sont très intéressants car ils ne nécessitent pas de calculs, sous réserve de disposer du découpage en plans de la séquence. Cependant, même si ces résumés peuvent convenir dans un certain nombre de situations, les résultats obtenus souffrent de la non prise en compte du contenu des plans.

Différentes solutions ont été proposées pour prendre en compte ce contenu [Truong 06]. Nous présentons ci dessous une de ces solutions :

- **l'image médiane** : l'image médiane d'un plan P se définit comme l'image la plus proche, au sens d'une certaine distance, de l'ensemble des autres images du plan, principe inspiré par le filtrage médian vectoriel [Chanussot 98]. D'un point de vue formel, cette médiane s'exprime par :

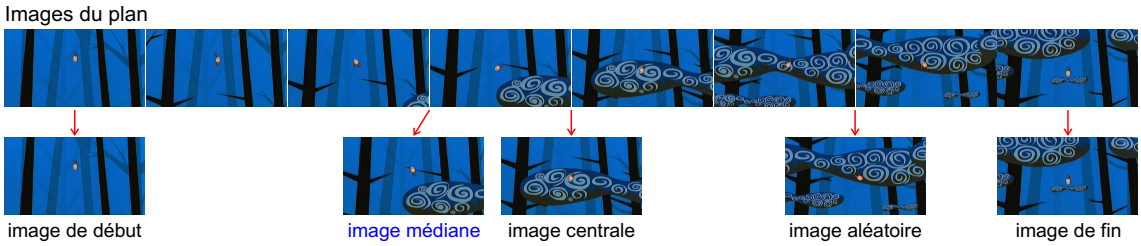
$$I_{méd} = argmin_{I_i \in P}\{D(I_i)\} \tag{6.3}$$

où I_i est une image du plan P et $D(I_i)$ est la distance cumulée de l'image d'indice i à toutes les autres images du plan, définie par :

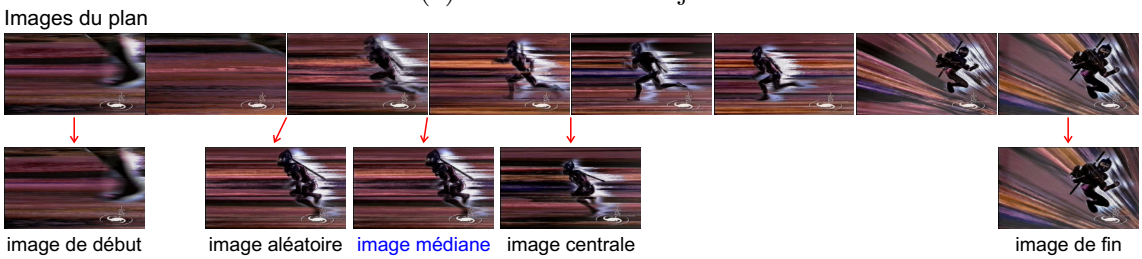
$$D(I_i) = \sum_{I_j \in P, j \neq i} d_{sim}(I_i, I_j) \tag{6.4}$$

où $d_{sim}(I_i, I_j)$ est une mesure de distance calculée entre les images I_i et I_j.

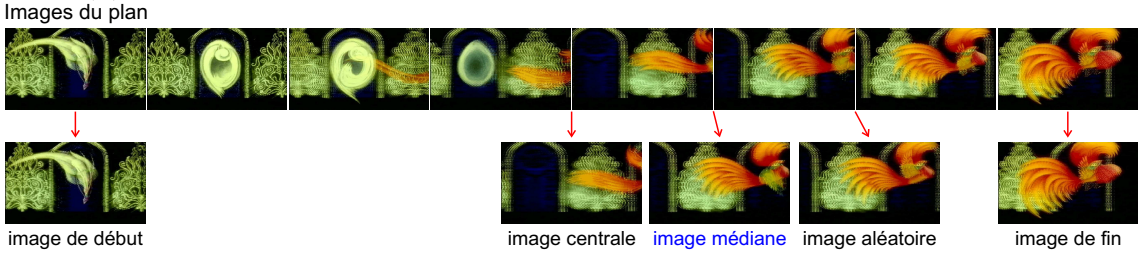
Dans la Figure 6.2, nous proposons quelques exemples utilisant ces stratégies à "une image par plan" (pour l'image médiane nous avons utilisé la distance de Manhattan [Jain 99]).



(a) mouvement d'objets



(b) mouvement très rapide de la caméra avec le changement de point de vue



(c) mouvements et changements complexes de la scène

FIG. 6.2 – Exemple d'images clés : (a) film "François le Vaillant", plan [9249 – 9308], (b) film "The Buddy System", plan [4907 – 5034], (c) film "Paradise", plan [4950 – 5191].

D'une manière générale l'approche "une image par plan" sans l'analyse du contenu convient pour les plans plutôt homogènes, comme par exemple la situation présentée par la Figure 6.2.a, où n'importe quelle image convient pour le résumé. D'autre part, pour les plans dont le contenu est plus complexe, cas des plans illustrés par les figures 6.2.b et 6.2.c, l'image médiane est plus adaptée car elle capture l'image la plus courante du plan.

Le résumé de la séquence est alors constitué par l'ensemble de toutes les images clés extraites de chaque plan :

$$R_{img}(S) = \{I_{plan_1} \cup I_{plan_2} \cup ... \cup I_{plan_N}\} \tag{6.5}$$

où S est la séquence, I_{plan_i} est l'image retenue du plan i et N le nombre total de plans de la séquence.

Dans la réalité, le contenu d'un plan comporte des changements visuels importants apportés par les déplacements d'objets ou les mouvements de caméra. Garder une seule image par plan avec les stratégies proposées ci-dessus n'est pas la meilleure solution. Il se peut que l'image retenue ne soit pas une image très significative, par exemple une image de transition dans un mouvement rapide (voir l'image de début dans la Figure 6.2.b ou l'image milieu dans la Figure 6.2.c). De plus, certains plans ne peuvent pas être résumés avec une seule image. C'est le cas des plans comportant un mouvement de caméra important (voir Figure 6.2.c). Plusieurs images sont alors nécessaires pour bien représenter le contenu du plan.

Le résumé adaptatif

Nous avons testé une technique d'extraction adaptative de résumé en images (voir le rapport [Ott 05] ou [Ionescu 06d]). Le contenu de chaque plan est résumé avec un *nombre d'images clés qui est adapté à l'action contenu* dans le plan.

Cette technique utilise la distance cumulée définie ci-dessus dans l'équation 6.4. L'histogramme de ces distances cumulées traduit de manière compacte et significative l'action contenue dans le plan. Le nombre d'images clés extraites pour chaque plan est alors déterminé en fonction de la forme de l'histogramme (unimodal, multimodal, etc.). L'évaluation globale de cette approche a été obtenue à travers la mise en place d'une étude présentée dans la Section 6.3.

Nous allons ici montrer quelques résultats obtenus pour un certain nombre de plans extraits de 2 films d'animation : "The Buddy System" et "Gazoon" [CICA 06]. Pour chaque plan analysé nous montrons *l'histogramme des distances cumulées* et les *images clés* extraites. Le résumé du plan ainsi obtenu est comparé avec le résumé utilisant l'image centrale du plan (voir l'Annexe D). Dans ces figures l'axe temporel situé à gauche précise les intervalles correspondant à chaque plan (image de début et image de la fin).

Nous pouvons remarquer que les résumés obtenus pour chaque plan sont en accord avec le contenu du plan. Dans la plupart des situations l'image du milieu, mais ce serait la même chose avec toutes les approches ne gardant qu'une seule image par plan, n'est pas suffisante pour représenter le contenu du plan quand celui-ci contient une forte activité. Dans ce cas le résumé adaptatif fournit plusieurs images en fonction de l'action contenue dans le plan. Par exemple, dans la Figure D.1, le premier plan (images $[19, 749]$) contient un mouvement 3D continu de la caméra avec plusieurs zooms sur des zones d'intérêt. L'image clé du milieu du plan est une image de transition, alors que les images clés obtenues avec la méthode adaptative correspondent à chacun des instants intéressants du plan.

Que ce soit avec la stratégie "une image par plan" ou la méthode adaptative, le nombre d'images du résumé est *en général trop élevé* pour une visualisation rapide. Ainsi, pour une séquence de 20 minutes dont la durée moyenne des plans serait de 6 secondes, le résumé aurait au moins 200 images. Des méthodes plus performantes sont nécessaires pour réduire ce nombre d'images. Cependant, bien que volumineux, ces résumés sont utiles car ils réduisent considérablement la masse des données contenues dans la séquence initiale et peuvent ainsi servir de point de départ pour d'autres stratégies de résumé plus complexes ou d'autres analyses (comme par exemple l'analyse de la distribution des couleurs proposée dans le Chapitre 4).

Le résumé compact

Pour certaines applications comme la navigation dans une base de films, il est indispensable de disposer d'un résumé très concis constitué de quelques images seulement. Les résumés présentés ci-dessus ne peuvent alors convenir. Ce résumé concis, noté *résumé compact* dans la suite, permettra de disposer des informations succinctes sur le contenu visuel global de la séquence.

Dans cette section nous proposons une amélioration de la méthode d'extraction du *résumé compact* proposée dans [Ott 05]. Dans ce résumé le nombre d'images est spécifié par l'utilisateur. Le résumé compact est calculé à partir d'un ensemble de départ constitué d'images clés obtenues par n'importe quelle méthode d'extraction d'images clés (par exemple une des méthodes présentées ci-dessus). Le nombre d'images de l'ensemble de départ, N_{init}, doit être supérieur au nombre d'images du résumé compact, N_{comp}, ($N_{init} > N_{comp}$). Par rapport à la méthode proposée dans [Ott 05], notre approche évite la sélection, dans le résumé compact, d'images clés visuellement trop similaires.

L'algorithme d'extraction du résumé compact consiste en la réduction itérative du nombre d'images de l'ensemble de départ, R_{init}, en utilisant comme critère la similarité visuelle entre les images. Il est présenté ci-dessous (Algorithme 4) :

Algorithm 4 Le calcul du résumé compact

initialisation(R_{init}, N_{init}) {la construction de l'ensemble des images de départ R_{init}, où N_{init} représente le nombre d'images}
$R_{comp} \leftarrow \{\emptyset\}$ {initialisation du résumé}
$N \leftarrow 0$ {N représente le nombre d'images du résumé}
pour $i = 1$ à N_{init}
  $D(I_i) \leftarrow \sum_{j=1}^{N_{init}} d_M(R_{init}[i], R_{init}[j])$ {$D(I_i)$ est la distance cumulée de l'image d'indice i à l'ensemble R_{init}, $R_{init}[i]$ est l'image d'indice i et $d_M()$ est la distance de Manhattan, voir [Ott 05]}
  $\bar{D}(I_i) \leftarrow D(I_i)/(N_{init} - 1)$ {calcul de la distance cumulée moyenne}
fin pour
$R^t_{init} \leftarrow$tri_croissant(R_{init}, $\bar{D}$) {tri des images de l'ensemble R_{init} selon les valeurs croissantes de la distance cumulée moyenne}
faire
  $i \leftarrow N_{init}$
  $N \leftarrow N + 1$
  $R_{comp}[N] \leftarrow R_{comp}[N] + R^t_{init}[i]$ {les images clés du résumé sont extraites d'une manière itérative dans le sens de la décroissance des valeurs de la distance cumulée moyenne}
  $i \leftarrow i - 1$
tant que ($N < N_{comp}$) {fin du calcul, le nombre d'images du résumé compact, N, est égal au nombre désiré N_{comp}}

Dans l'algorithme proposé dans [Ott 05], la première image du résumé est toujours l'image médiane (la plus semblable aux autres). Puis les autres images sélectionnées itérativement sont les images les plus éloignées de celles déjà retenues dans le résumé. Notre stratégie diffère fondamentalement de cette approche, en particulier parce qu'elle n'utilise pas l'image médiane. Elle assure que les images retenues dans le résumé sont toujours les images les plus dissemblables de l'ensemble initial R_{init}.

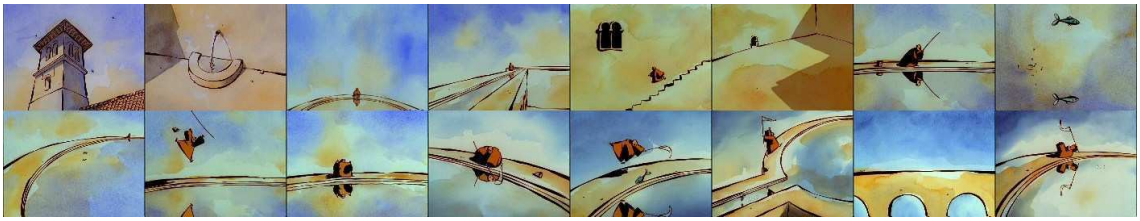
Nous avons pu tester le résumé compact sur plusieurs films d'animation de [CICA 06]. Dans la suite nous allons présenter et commenter quelques exemples de résumés obtenus pour certains passages pris dans 3 films d'animation : "François le Vaillant" (Figure 6.3) "Le Moine et le Poisson" (Figure 6.4) et "Le Roman de Mon Ame" (Figure 6.5). Pour l'ensemble initial d'images clés, R_{init}, nous avons utilisé le résumé en une image par plan (image du milieu) présenté ci-dessus.

(a) une image par plan (image de milieu)

(b) résumé compact en 5 images

(c) résumé compact en 3 images

FIG. 6.3 – Exemple de résumé compact pour un extrait de 16 plans du film "François le Vaillant" [Folimage 06b].

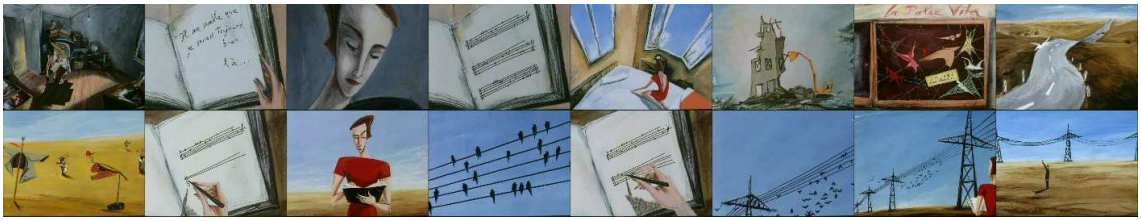


(a) une image par plan (image de milieu)

(b) résumé compact en 5 images

(c) résumé compact en 3 images

FIG. 6.4 – Exemple de résumé compact pour un extrait de 16 plans du film "Le Moine et le Poisson" [Folimage 06b].



(a) une image par plan (image de milieu)

(b) résumé compact en 5 images

(c) résumé compact en 3 images

FIG. 6.5 – Exemple de résumé compact pour un extrait de 16 plans du film "Le Roman de Mon Âme" [Folimage 06b].

En analysant les résultats nous avons observé que si l'ensemble d'images clés initial, R_{init}, contient un petit groupe composé d'images semblabes entre elles mais très différentes de la plupart des autres images de l'ensemble, le résumé compact aura tendance à les retenir. Par exemple, dans la Figure 6.4 le résumé compact en 3 images contient deux images assez similaires mais qui sont différentes de la plupart des images initiales. Ce défaut, peu fréquent, pourrait être contourné en utilisant l'information temporelle en imposant par exemple un écart temporel minimum entre les images du résumé.

Il faut noter que le choix des images du résumé est dépendant de la méthode de réduction des couleurs employée et du choix de la mesure de distance entre images utilisée. La distance que nous avons employée est la distance de Manhattan entre les histogrammes des images en couleurs réduites (voir [Ott 05]). La réduction couleur que nous avons retenue est la quantification uniforme de l'espace RVB en $5 \times 5 \times 5$ couleurs. Cette technique, que nous avons déjà présentée (voir Section 2.4.2), est sensible aux variations de l'intensité lumineuse dans l'image. Par exemple, dans la Figure 6.4.c les deux dernières images sont similaires, mais la présence de l'ombre dans la troisième image change la distribution des couleurs ce qui augmente la distance entre les deux images. Néanmoins, la quantification de l'espace RVB a l'avantage d'avoir une complexité de calcul réduite et donne des résultats globalement satisfaisants.

Les meilleurs résultats ont été obtenus pour le film "Le Roman de Mon Ame", film qui comporte beaucoup de changements visuels (voir Figure 6.5). Les images du résumé compact sont toutes différentes entre elles.

Conclusions

Globalement la durée d'un résumé obtenu en gardant une image par plan est importante. Si 10 minutes de film correspondent à peu près à 100 plans, alors ce type de résumé comportera 100 images. Dans le cas du résumé adaptatif le nombre d'images est encore plus

élevé, plusieurs images pouvant être retenues pour chaque plan en fonction de l'activité du plan. Dans la plupart des applications la visualisation de toutes ces images est une opération lourde.

Cependant, ce type de résumé est une représentation fidèle de la totalité du contenu de la séquence. Il est donc utile dans le cas où l'utilisateur cherche à connaître tout le contenu visuel de la séquence, mais sans prendre le temps de la regarder entièrement. Par exemple, les 100 images peuvent être visualisées, à une cadence de 0.5 images/s, en 50 secondes, ou même seulement en quelques secondes si elles sont organisées sous la forme d'une planche (voir Figure 6.3.a). En conclusion, bien que volumineux, ce résumé "par plan" peut être intéressant pour une visualisation approfondie de la séquence.

Le second intérêt de ce résumé est sa capacité à réduire la redondance temporelle de l'information contenue dans la séquence. En effet, le contenu visuel d'une séquence y est préservé d'une manière efficace et compacte, avec un rapport de compression très élevé (par exemple un film de 10 minutes, à 25 images/s, contient 15000 images qui sont résumées en une centaine d'images). Un tel résumé peut alors constituer les données initiales pour des analyses du contenu. Nous avons ainsi utiliser ce résumé pour calculer la distribution globale des couleurs d'une séquence (approche proposée dans la Section 4.2) et nous avons constaté que la restriction à quelques images par plan n'altère pas beaucoup les résultats.

Dans des tâches de recherche ou de navigation dans une base de données de séquences d'images, le temps de consultation du contenu d'une séquence doit être très court. Le résumé par plan est très mal adapté à cette situation alors que le résumé compact (voir Figure 6.3.c) peut être très efficace. Par exemple, l'outil d'exploration de Microsoft Windows utilise une vignette (typiquement la première image de la séquence) associée à chaque fichier vidéo pour donner une idée du contenu. En utilisant le résumé compact, on peut envisager de proposer quelques vignettes (par exemple 2, 3, ...) fournissant une meilleure représentation que la première image ou même qu'une image aléatoire.

6.2.2 Les résumés dynamiques

Comparé aux résumés en images, les résumés dynamiques apportent une information complémentaire sur le mouvement contenu dans la séquence. L'information audio peut également y être présente. Dans les méthodes que nous proposons dans la suite, le son ne sera pas exploité. Dans le cadre de ce travail, nous nous sommes limités à l'utilisation de l'image. Le son est un élément important qui devra être pris en compte dans l'avenir. Notons cependant que, dans la plupart des films d'animation utilisés dans nos expérimentations (42 sur 52) il n'a pas de dialogues ou de commentaires, mais uniquement de la musique (voir l'Annexe F).

L'approche par plan

La première approche proposée est semblable au résumé en images avec une seule image par plan. L'idée est de résumer le contenu dynamique de chaque plan vidéo en ne retenant qu'un passage du plan. Le découpage en plans nous assure que les images moins pertinentes de la séquence, celles correspondant par exemple à des transitions lentes ou à des plans très courts, sont éliminées.

En admettant que la probabilité de tomber sur des images représentatives du contenu d'un plan est très élevée pour les images proches du milieu du plan, nous proposons de représenter chaque plan de la séquence par une sous-séquence continue d'images centrée au

milieu du plan, et contenant $p\%$ du nombre total d'images du plan. Avec cette stratégie, le résumé dynamique de la séquence est défini par :

$$R_{mouv}(S) = \{seq^c_{1,p} \cup seq^c_{2,p} \cup ... \cup seq^c_{N,p}\} \tag{6.6}$$

où S est la séquence, N est le nombre total de plans vidéo, $seq^c_{i,p}$ est une sous-séquence centrée sur le milieu du plan i, contenant $p\%$ du nombre total d'images du plan.

En retenant de chaque plan un pourcentage du nombre total d'images du plan, les plans longs, contenant aussi plus d'information, seront bien évidemment mieux représentés que les plans courts.

En ce qui concerne le choix du paramètre p, nous avons fait un compromis entre *la préservation de la continuité visuelle du résumé* et *la longueur du résumé*. Après un certain nombre de tests effectués sur plusieurs séquences d'animation, nous avons trouvé que $p \in [15, 25]\%$ est le meilleur compromis de continuité visuelle/longueur du résumé.

Une valeur de $p = 15\%$ assure une continuité visuelle et une préservation satisfaisante du rythme de la séquence. La réduction du contenu ainsi obtenue est supérieure à $[\frac{100}{p}] = 6$ (et ce coefficient ne prend pas en compte les images de transition éliminées dans l'étape d'agrégation en plans, voir la Section 2.7). Ainsi, pour le film "Fini Zayo" [Folimage 06b] d'une durée de 7min, en retenant 15% de chaque plan on obtient un résumé d'un durée de 1min 3s. Pour un film plus long, comme par exemple le film "The Hill Farm" [Folimage 06b], d'une durée totale de 17min on obtient un résumé de 2 min 33s. Pour obtenir une préservation plus fidèle du rythme de la séquence (accéléré dans le résumé par le prélèvement d'images de chaque plan), on peut envisager de prendre une valeur de p supérieure à 15%, mais cela aboutit à des résumés de plus longue durée. Déjà pour $p = 20\%$ dans le cas du film "The Hill Farm", le résumé obtenu a une durée de 3min 24s, durée importante pour une visualisation rapide.

L'approche par plan est plutôt efficace dans le cas de films courts (d'une durée inférieure à 12 minutes), car le résumé obtenu est visualisable en moins de 2 minutes. Ce type de résumé est bien adapté aux courts métrages d'animation tels que ceux du festival d'Annecy [CICA 06]. De plus, cette approche est intéressante car elle a une complexité de calcul réduite : elle ne demande pas de calculs supplémentaires pour extraire le résumé.

Le résumé "bande-annonce"

En considérant que les parties de la séquence contenant de l'action correspondent aux plages présentant une fréquence de changements de plan élevée (hypothèse souvent utilisée dans le domaine, voir la Section 6.1.2 sur l'état de l'art), nous proposons un résumé dynamique qui prend en compte l'action contenue dans la séquence [Ionescu 06d].

Dans l'approche par plan nous avons utilisé comme unité de base les plans vidéo. Dans le résumé que nous proposons, nous utilisons une unité de plus haut niveau qui est *le segment d'action*. Un segment d'action est défini comme un passage de la séquence comprenant plusieurs plans et présentant un nombre élevé de changements de plan. L'algorithme de construction des segments d'action a été présenté dans la Section 2.9.1.

Le résumé dynamique proposé s'appuie sur le résumé du contenu de chaque segment d'action de la séquence. Construit de cette manière *il contiendra seulement les parties de la séquence contenant de l'action*. Ceci permet d'aboutir à un résumé que nous appellerons "bande-annonce", par analogie avec les bandes annonces des films qui tentent d'attirer le spectateurs en ne montrant que les passages riches en action.

Ce résumé "bande-annonce" est construit de la manière suivante : *pour chaque segment d'action de la séquence nous prélevons un court extrait.* Ce court extrait est obtenu en concaténant les résumés dynamiques de chaque plan constituant le segment d'action. Le résumé dynamique de chaque plan est une sous-séquence centrée sur le milieu du plan et contenant $p\%$ images du plan (voir résumé par plan présenté ci-dessus). On a ainsi :

$$R_{ba}(S) = \{pass_1 \cup pass_2 \cup ... \cup pass_M\} \tag{6.7}$$

où S la séquence, $pass_i$ est le court extrait provenant du segment d'action i, $i = 1, ..., M$ avec M le nombre total de segments d'action de la séquence. $pass_i$ est défini par :

$$pass_i = \{seq^c_{i,1,p} \cup seq^c_{i,2,p} \cup ... \cup seq^c_{i,N_i,p}\} \tag{6.8}$$

où $seq^c_{i,j,p}$ est une sous-séquence d'images, centrée sur le milieu du plan j du segment d'action i, contenant $p\%$ du nombre total d'images, et $j = 1, ..., N_i$ où N_i est le nombre total de plans vidéo contenus dans le segment d'action i.

En ce qui concerne la valeur du p, les remarques faites pour le résumé dynamique par plan sont aussi valables pour ce résumé (voir la sous section précédente). Nous allons donc utiliser une valeur $p \in [15, 25]\%$ pour assurer une continuité visuelle du résumé. Une étude comparative des durées des deux approches est présentée dans le Tableau 6.1.

Film	Durée	T_{ba}	T_{dyn}	N_{plans}	R_{action}
"François le Vaillant"	8min56s	**1min25s**	2min15s	164	70%
"La Bouche Cousue"	2min48s	**16s**	42s	39	52.5%
"Ferrailles"	6min15s	**1min31s**	1min34s	138	98%
"A Viagem"	7min32s	**1min**	1min48s	54	71%
"David"	8min12s	**23s**	1min58s	27	40%
"Greek Tragedy"	6min32s	**24s**	1min36s	29	48%

TAB. 6.1 – Etude comparative des durées des résumés : T_{ba} est la durée du résumé "bande-annonce", T_{dyn} est la durée du résumé dynamique par plan, N_{plans} est le nombre total de plans vidéo, R_{action} est le rapport d'action du film ($p = 25\%$).

Le résumé "bande-annonce" est beaucoup plus court que celui obtenu en gardant une sous-séquence par plan, même pour une valeur élevée de p (25%), car seuls les passages de la séquence riches en action seront résumés. La seule situation où le résumé "bande-annonce" a une durée comparable à celle de l'approche par plan est le cas de films contenant beaucoup d'action (valeur du rapport R_{action} élevée, voir la Section 2.9.1), comme par exemple le film "Ferrailles" [Folimage 06b] où du fait de la fréquence élevée des changements de plan la plupart des plans ont été considérés comme importants pour le résumé.

L'évaluation de la qualité du résumé "bande-annonce" est présentée dans le chapitre suivant.

Conclusions

En résumant chaque plan/segment de la séquence par une sous-séquence d'images de durée proportionnelle à la durée des plans, comme dans les approches proposées, le résumé obtenu donne l'impression d'une d'accélération du rythme visuel. Cet effet est encore plus

prononcé pour les passages de la séquence contenant une succession de plans de courte durée. Par exemple, pour des plans d'une durée de 3s, contenant 75 images (à 25 images/s), en ne retenant que 15% des images on aboutit à une succession de sous-séquences d'une durée d'environ 0.5s, durée à peine suffisante pour avoir une bonne perception des plans.

Il y a des situations où cet effet d'accélération change la perception que l'on peut avoir de la séquence. Par exemple, dans le domaine des films d'animation, on trouve souvent des films pour lesquels le rythme de déroulement des événements est lié au contenu de la séquence, l'artiste ayant choisi volontairement une certaine vitesse de déroulement de l'action pour transmettre une sensation particulière.

On peut envisager différentes solutions pour améliorer la qualité du résumé obtenu et éviter ce phénomène d'accélération. On peut par exemple augmenter le nombre d'images retenues pour chaque unité de la séquence, mais ceci augmente la taille du résumé. On peut également ne retenir qu'un faible nombre de plans, plans qui ne sont pas résumés mais présentés en intégralité dans le résumé. La difficulté est alors de sélectionner judicieusement les plans conservés.

Le résumé "bande-annonce" proposé s'appuie sur le fait que les zones d'action sont liés à une cadence de changements de plan élevée. Cette hypothèse n'est cependant pas toujours valable. Il y a des situations où l'action se déroule à l'intérieur d'un même plan. Dans ce cas l'action de la séquence provient des mouvements des personnages dans la scène ou de changements visuels. Dans le cas des films d'animation, on trouve également des films ne contenant qu'un nombre réduit de plans vidéo (inférieur à 5) ce qui rend impossible une analyse du rythme des changements de plan. Dans cette catégorie on peut mentionner des films comme "Amerlock", "Sculptures", "The Wall" [CICA 06] qui utilisent une technique particulière d'animation : la pâte à modeler (voir Figure 6.6). Dans ce cas l'action du film se déroule dans une ou deux scènes seulement et est entièrement contenue dans les images et le son, mais pas ne provient pas du rythme des changements de plan.

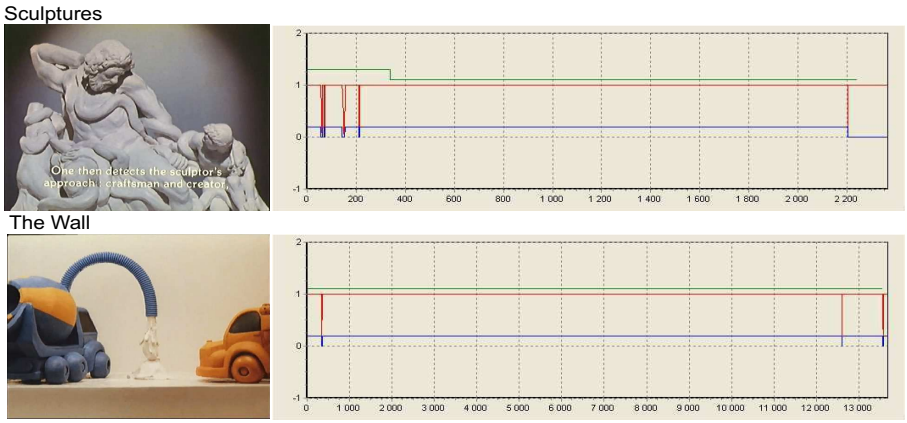


FIG. 6.6 – Exemple de films d'animation [CICA 06] contenant un nombre réduit de plans. Chaque film est représenté par une image et l'annotation visuelle des transitions (l'axe oX est l'axe temporelle et les lignes rouges verticale indiquent un changement de plan, voir Section 2.8).

Pour améliorer le résumé "bande-annonce" du film une stratégie consiste à utiliser des informations extraites d'une analyse intra-plan de la séquence, comme le mouvement de la

caméra ou d'objets et plus généralement l'activité spatiale fournie par exemple par l'histogramme des distances cumulées proposé dans [Ott 05] (voir aussi l'Algorithme 4 Section 6.2.1).

6.3 L'évaluation des résumés

L'évaluation d'un résumé est un processus subjectif car le concept de résumé est lui-même subjectif. Il est donc difficile d'extraire automatiquement un indicateur numérique mesurant la qualité d'un résumé. Parmi les méthodes d'évaluation existantes (voir la Section 6.1.3 sur l'état de l'art), les tests d'évaluation semblent être les méthodes les plus pertinentes pour juger la qualité des résumés [Truong 06]. En effet, dans ces méthodes, le "consommateur" du produit, c'est-à-dire l'utilisateur, évalue lui-même le résumé à travers un test de qualité. Après visualisation des résumés proposés, c'est à lui de décider de la cohérence et de la qualité de chaque résumé.

L'évaluation qualitative de la méthode d'extraction du résumé adaptatif (résumé en images, voir le rapport [Ott 05]) et du résumé " bande-annonce " [Ionescu 06d] (résumé dynamique) a été effectuée en organisant une campagne d'évaluation. Parmi les films d'animation du festival d'Annecy ([CICA 06]), nous avons sélectionné pour la campagne 10 séquences représentatives : *"Casa"* 6min5s, *"Circuit Marine"* 5min35s, *"Ferrailles"* 6min15s, *"François le Vaillant"* 8min56s, *"Gazoon"* 2min47s, *"La Bouche Cousue"* 2min48s, *"La Cancion du Microsillon"* 8min29s, *"Le Moine et le Poisson"* 5min59s, *"Paroles en l'Air"* 6min50s et *"The Buddy System"* 6min19s, d'une durée totale de 60min3s.

Les tests d'évaluation des résumés proposés ont été effectués sur un groupe de *27* personnes composé pour la plupart d'étudiants, de quelques enseignants et de trois spécialistes du domaine de l'animation. Les ages des participants allaient de 21 à 49 ans.

6.3.1 Le protocole d'évaluation

Le protocole d'évaluation consiste en la visualisation des résumés proposés aux 27 participants en même temps, dans une salle de projection. Pour chaque film nous disposons de *la séquence originale*, du *résumé adaptatif* en images et du *résumé "bande-annonce"*. Pour la visualisation du résumé en images nous avons opté pour une présentation de type "slide-show". Les images du résumé sont montrées progressivement, l'une après l'autre, avec une cadence de deux images par seconde. Chaque film est accompagné d'*un questionnaire* sur la qualité des résumés proposés (voir l'Annexe E). L'évaluation d'un film se déroule de la façon suivante :

- présentation de la la séquence originale dans son intégralité. En particulier, le son a été conservé,
- présentation du premier résumé (le résumé adaptatif),
- réponse aux questions concernant le premier résumé,
- présentation du deuxième résumé (le résumé "bande-annonce"),
- réponse aux questions concernant le deuxième résumé,
- présentation de la la séquence originale suivante,

Le processus est répété sur l'ensemble des 10 films.

6.3.2 Les questionnaires

Les questionnaires sont construits de la manière suivante : pour chaque film la fiche d'évaluation contient une première partie pour le résumé adaptatif, une seconde partie pour le résumé "bande-annonce" et une dernière partie permettant au participant de faire des remarques (voir Figure E.1). Les questions posées concernent le contenu et la durée du résumé.

Le résumé adaptatif. Pour *le résumé adaptatif* en images nous avons posé les questions suivantes :

- **question 1** : *"Estimez-vous que le résumé en images représente bien le contenu du film ?"*. L'évaluation est donnée en utilisant une échelle de 0 à 10 (principe inspiré des sondages d'opinion) accompagnée des appréciations suivantes : *je ne sais pas* - 0 (ou symbole X), *pas du tout* - 1 ou 2, *très peu* - 3 ou 4, *partiellement* - 5 ou 6, *en grande partie* - 7 ou 8 et *totalement* - 9 ou 10. Pour chaque niveau d'appréciation l'utilisateur dispose de deux niveaux sur l'échelle.
- **question 2** : *"Comment estimez-vous le nombre d'images pour le résumé en images ?"* avec comme réponses possibles : *je ne sais pas* - 0 (ou symbole X), *trop petit* - 1 ou 2, *petit* - 3 ou 4, *suffisant* - 5 ou 6, *élevé* - 7 ou 8, *trop élevé* - 9 ou 10.

Le résumé "bande-annonce". En ce qui concerne le *résumé "bande-annonce"* les questions que nous avons posées sont les suivantes :

- **question 1** : *"Pensez-vous que le résumé "bande-annonce" contient les passages les plus importants du film ?"*. Pour l'évaluation des réponses nous avons utilisé la même échelle que pour la question 1 précédente.
- **question 2** : *"Comment trouvez-vous la durée du résumé proposé ?"*. Dans ce cas nous avons proposé les réponses suivantes : *trop courte*, *courte*, *correcte*, *longue*, *trop longue*.

Les résultats obtenus pour cette campagne d'évaluation sont présentés dans la suite.

6.3.3 Les résultats de la campagne

Nous avons évalué les réponses aux questionnaires en analysant les scores obtenus. Pour chaque séquence et pour chaque question posée nous avons calculé le score moyen et l'écart type. La valeur moyenne donne une idée globale sur l'appréciation générale du résumé et l'écart type nous indique la dispersion des réponses.

Le résumé adaptatif. Les résultats obtenus pour l'évaluation *du résumé adaptatif* sont présentés en Figure 6.7 (pour la question 1) et Figure 6.8 (pour la question 2).

Pour la question 1 : *"Estimez-vous que le résumé en images représente bien le contenu du film ?"* nous avons obtenu pour les 10 films un score moyen global de *6.9* avec un écart type globale de *1.7*. Pour la question 2 : *"Comment estimez-vous le nombre d'images pour le résumé en images ?"* le score moyen global est de *6.1* avec un écart type globale de *1.5*.

D'une manière générale, le résumé adaptatif a été apprécié comme représentant *en grande partie* le contenu de la séquence et ayant un *nombre d'images suffisant*. Le résumé adaptatif a été moins bien apprécié pour certains films ayant un contenu complexe. Par exemple, dans le cas du film "La Cancion du Microsillon" nous avons obtenu un nombre élevé de réponses

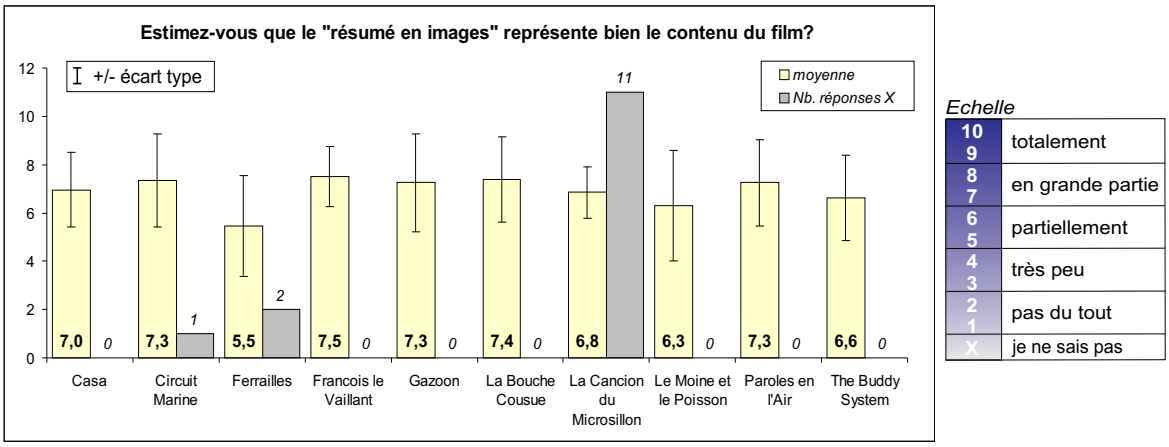


FIG. 6.7 – Les statistiques des réponses pour la question 1 du *résumé adaptatif*.

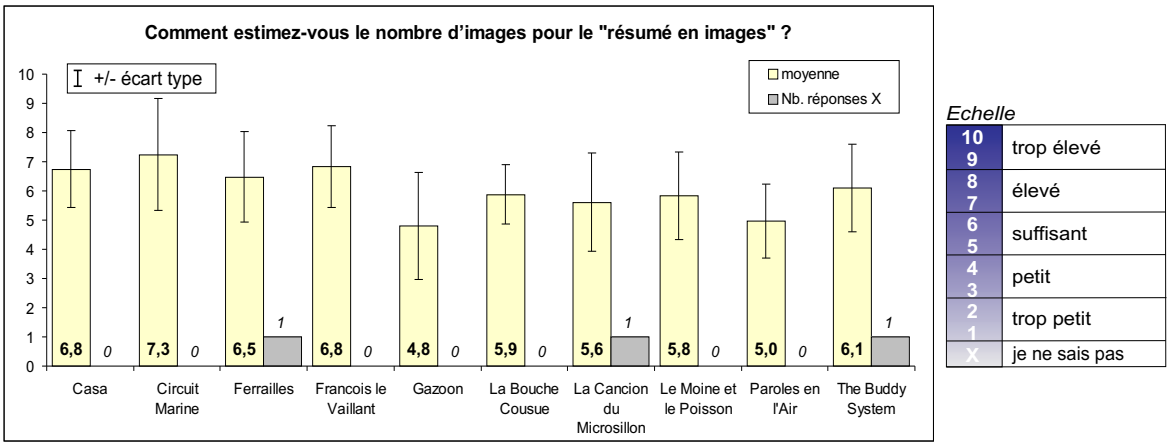


FIG. 6.8 – Les statistiques des réponses pour la question 2 du *résumé adaptatif*.

je ne sais pas (11) pour la question 1, car le contenu du film est difficilement compréhensible. De même, pour le film "Le Moine et le Poisson", nous avons obtenu une valeur élevée de la dispersion des réponses (écart type 2.3), ce qui montre que le film a été perçu différemment selon les personnes. Le plus mauvais score a été obtenu pour le film "Ferrailles" qui est un film complexe avec beaucoup de changements de scènes. Pour ce film, le résumé obtenu a été considéré comme représentant *partiellement* le contenu de la séquence. En ce qui concerne la durée des résumés, généralement le nombre d'images du résumé a donné des appréciations situées entre *correct* et *élevé*.

Le résumé "bande-annonce". Les résultats obtenus pour le *résumé "bande-annonce"* sont présentés en Figure 6.9 (pour la question 1) et Figure 6.10 (pour la question 2).

Pour la question 1 : *"Pensez-vous que le résumé "bande-annonce" contient les passages les plus importants du film ?"* nous avons obtenu pour les 10 films un score moyen global de *7.7* avec un écart type global de *1.3*. Pour la question 2 : *"Comment trouvez-vous la durée du résumé proposé ?"* le score moyen global est de *2.6* avec un écart type global de *0.6*.

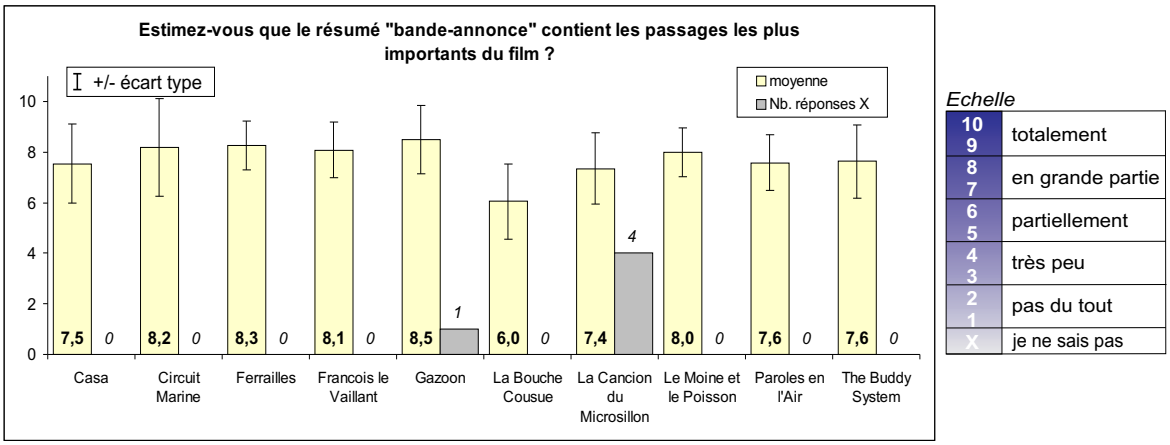


FIG. 6.9 – Les statistiques des réponses pour la question 1 du *résumé "bande-annonce"*.

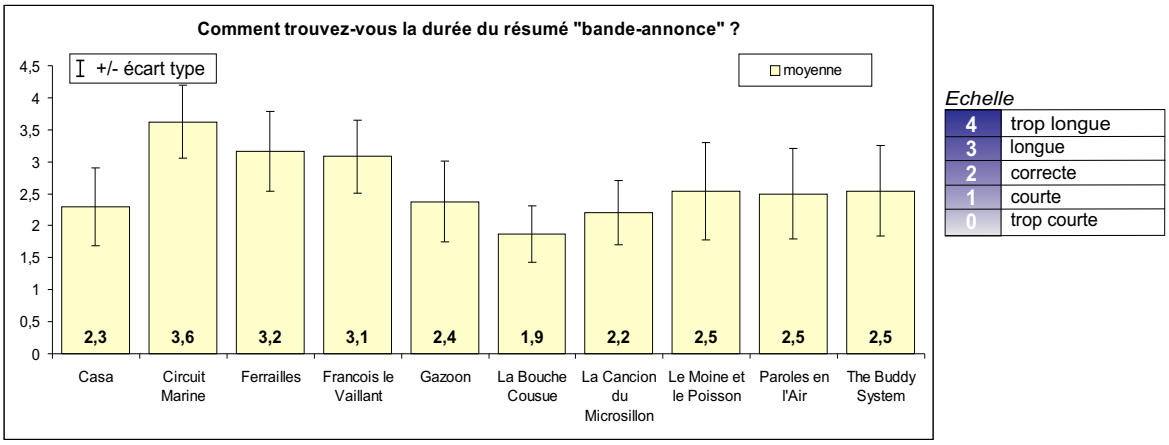


FIG. 6.10 – Les statistiques des réponses pour la question 2 du *résumé "bande-annonce"*.

D'une manière générale, le résumé "bande-annonce" a été apprécié comme représentant *en grande partie* le contenu de la séquence et ayant *une durée correcte*. Grâce à son contenu dynamique, le résumé "bande-annonce" a été mieux apprécié que le résumé adaptatif. La dispersion des réponses est plus faible que dans le cas du résumé adaptatif.

La méthode proposée pour l'extraction du résumé "bande-annonce" a donné de bons résultats pour chacun des films testés. Les résumés obtenus ont été globalement appréciés comme représentant en grande partie les passages les plus importants de la séquence. Cependant la durée du résumé a parfois été appréciée comme étant longue. C'est le cas de films contenant beaucoup d'action, comme par exemple "François le Vaillant", "Circuit Marine" ou "Ferrailles" [Folimage 06b]. Le score le plus mauvais a été obtenu pour le film "La Bouche Cousue" où l'action est pratiquement entièrement contenue dans le son.

6.3.4 Conclusions sur l'évaluation

La mise en place d'un test d'évaluation pour l'appréciation des résumés est une tâche très difficile du point de vue pratique. Par rapport aux évaluations faites en indexation ou compression d'images, le fait de travailler avec des séquences d'images augmente la difficulté du protocole d'évaluation.

Le premier problème est lié au *temps élevé* nécessaire à la visualisation des films et des résumés proposés. Pour les 10 films utilisés, il faut déjà une heure pour la seule projection des séquences originales. En ajoutant la durée de présentation des résumés, les pauses entre les visualisations, le temps pour répondre aux questions, le temps total nécessaire pour l'évaluation des 10 films a dépassé deux heures. Il est donc difficile de trouver un nombre suffisant de personnes acceptant de passer quelques heures en continu pour évaluer des résumés.

Pour résoudre cette difficulté, une solution est de réaliser l'évaluation à distance, en utilisant par exemple Internet. Actuellement nous réfléchissons à la mise en place d'un site web pour l'évaluation des résumés dans le contexte des films d'animation (projet en collaboration avec CITIA - Cité de l'Image en Mouvement d'Annecy [CITIA 06] et CICA - Centre International du Cinéma d'Animation [CICA 06]) qui soit accessible en ligne pour tous les utilisateurs. Chaque personne visualisera comme elle le veut et quand elle le veut une séquence et répondra ensuite à un questionnaire électronique sur la qualité des résumés. De cette façon l'évaluation se fait au fur à mesure en évitant la fastidieuse visualisation en continu de tous les films. L'inconvénient de cette façon de procéder est de ne pas contrôler les conditions de visualisation des films et des résumés.

Un autre problème, lié également à la durée du test, est *la difficulté de réaliser un comparatif* entre des résumés obtenus avec différentes approches ou différents réglages d'une même approche. L'évaluation d'un trop grand nombre de résumés risque de provoquer *la lassitude* ou *la fatigue* des participants pouvant fausser les réponses. De plus, après une seule visualisation les participants ont souvent un peu de mal à apprécier les différences entre deux résumés résultant de deux approches différentes. Pour donner un avis objectif il est souvent nécessaire de faire une seconde visualisation, mais ceci allonge la durée de l'évaluation.

Un autre effet gênant est *l'apprentissage du contenu* des résumés. Après la visualisation de la première version d'un résumé l'utilisateur a tendance à apprendre son contenu et ensuite à le mélanger avec le contenu des autres versions visualisées. Il va alors qualifier comme satisfaisants la plupart des résumés proposés sans faire la différence entre eux. La pertinence de l'évaluation des autres versions du résumé en est bien sûr affectée. Une solution est d'évaluer séparément chaque version de résumé proposée et de les comparer à la fin.

Cependant, le principal inconvénient de l'évaluation de résumés par des tests d'évaluation reste la subjectivité liée à la perception de chaque participant au test. La qualité attribuée à un résumé est liée à la *façon de le percevoir*. Chaque personne en fonction de son jugement, de son émotion, de sa formation professionnelle, etc. évaluera d'une manière différente le contenu d'une séquence (un ingénieur et un artiste vont sûrement utiliser des éléments d'appréciations différents). Cet effet est visible dans les tests que nous avons effectués. Par exemple dans le cas du film "Ferrailles" le résumé adaptatif a été apprécié par les participants avec des scores allant de 0 à 10 (voir dans la Figure 6.7 les valeurs élevées de la dispersion des réponses).

Un problème particulier aux résumés en images est le mode de visualisation de ce type de résumé. Pour l'évaluation nous avons montré les résumés en images en utilisant une présentation de type "slideshow ". Ce type de présentation donne l'impression de voir une copie saccadée et de mauvaise qualité de la séquence originale. Le résumé n'est pas perçu

comme un ensemble d'images mais comme une séquence. Une solution à ce problème est de présenter les images sous la forme d'une planche contenant toutes les images du résumé (voir par exemple la Figure 6.3). Mais avec cette présentation, l'évolution temporelle du résumé est en partie perdue car l'utilisateur a tendance à se focaliser sur le centre de la planche et à regarder les images d'une manière aléatoire.

Dans le protocole d'évaluation utilisé, nous avons choisi de visualiser en premier la séquence originale et ensuite les résumés proposés. Une autre solution intéressante est d'inverser le processus : les résumés d'abord, la séquence originale ensuite. On demande alors aux participants dans quelle mesure ils ont compris le contenu de la séquence à partir des résumés. Cette approche évite aux participants la nécessité de comprendre dans le détail le contenu de la séquence.

6.4 Conclusions générales

Dans ce chapitre nous avons proposé et testé un certain nombre de techniques d'extraction de résumés d'une séquence d'images. Les approches existantes sont orientées vers deux directions distinctes : les résumés en images, résumés statiques, et les résumés dynamiques. Nous avons proposé plusieurs techniques dans chaque catégorie :

- **résumés en images** : d'abord nous avons étudié l'intérêt de l'approche *une image par plan* (chaque plan est résumé par une image clé). Cette approche se révèle intéressante du point de vue de la complexité de calcul (pratiquement négligeable si on dispose du découpage en plans de la séquence), cependant le résumé obtenu ne prend pas en compte le contenu dynamique de la séquence et est souvent trop long pour certaines applications.

  Ensuite, nous avons testé une *approche adaptative* pour laquelle le nombre d'images extraites de chaque plan est proportionnel à l'activité visuelle du plan. Les contraintes de cette approche sont le temps de calcul élevé et la taille du résumé, souvent trop longue (plus longue que dans l'approche par plan car plusieurs images peuvent être extraites de chaque plan). Cependant, cette approche a l'avantage d'être adaptée au contenu visuel de chaque plan.

  Enfin, nous avons proposé un *résumé compact* en un nombre d'images spécifié par l'utilisateur et calculé à partir d'un ensemble initial d'images clés. Cette approche nous permet d'avoir un résumé visuel constitué de seulement quelques images extraites de la totalité de la séquence, très utile dans des tâches telle que la navigation dans une base de séquences d'images,

- **résumés en mouvement** : d'abord nous avons testé une *approche par plan* (chaque plan est résumé par une sous-séquence d'images). Semblable à l'approche par plan du résumé en images, le résumé dynamique ainsi obtenu a généralement une durée trop longue. Néanmoins, cette approche est très efficace pour les films courts (moins de 12 minutes, situation fréquente pour les films d'animation) car la durée du résumé obtenu est alors satisfaisante. De plus la complexité du calcul est négligeable si on dispose du découpage en plans de la séquence. Le résumé par plan ne prend pas en compte l'action contenu dans la séquence.

  Aussi avons-nous proposé un résumé plus compact qui reproduit seulement les passages où l'action est importante : le *résumé "bande-annonce"*. Cette approche s'appuie

sur la mesure de la fréquence des changements de plan. La durée du résumé dans ce cas sera beaucoup plus courte que le résumé par plan, tout en ne gardant que le contenu intéressant de la séquence.

Les méthodes proposées ont été appliquées au cas particulier des films d'animation du festival d'Annecy ([CICA 06]). L'évaluation des résumés proposés a été effectuée par la mise en place d'une campagne de tests. Les tests d'évaluation se trouvent être la meilleure méthode d'évaluation car elles engagent dans le processus d'évaluation la perception du "consommateur du produit" qu'est l'utilisateur.

Apprécier la qualité d'un résumé est une tâche difficile et subjective. Elle est liée à la manière de percevoir de chacun d'entre nous. La qualité d'un résumé est aussi liée à l'objectif visé. Par exemple, le jugement d'un résumé devant représenter le contenu global de la séquence ne peut pas être le même que pour celui composé des événements importants de la séquence. De la même façon, un résumé en images ne peut pas être comparé avec un résumé dynamique car les deux représentent des informations différentes. Dans le résumé en images il manque l'information dynamique, mais sa durée est beaucoup plus courte que le résumé dynamique et il donc plus facile à visualiser. En ce qui concerne le contenu, le résumé dynamique est plus intéressant car la présence de mouvement apporte une information complémentaire très riche. Du point de vue de la visualisation, il est aussi plus agréable de regarder une séquence d'images que de visualiser un certain nombre d'images fixes.

Chapitre 7

La description sémantique

Résumé : *La description de données par des caractéristiques numériques n'est pas toujours compréhensible par tous, et reste souvent l'affaire de spécialistes de l'analyse des données. Par contre une description sémantique est en général accessible au plus grand nombre. Dans ce chapitre, nous allons proposer une méthodologie permettant la conversion de différents paramètres de bas niveau, extraits des séquences d'images, en des termes sémantiques. Le but est donc de définir des termes simples à comprendre reflétant le plus fidèlement possible l'impression se dégageant d'une séquence d'images. La méthodologie proposée est appliquée au domaine des films d'animation. Ces concepts symboliques/sémantiques sont extraits en utilisant une représentation floue des données construite à l'aide de connaissance a priori fournie par des experts du domaine de l'animation.*

Le but des travaux proposés dans cette thèse est la mise en place d'un *système de traitement* capable d'analyser et de comprendre d'une manière automatique le contenu des films d'animation (voir les rapports [Ionescu 04a] [Ionescu 04b]). Le système est composé d'un ensemble d'outils qui font la traduction de données de bas niveau (mesures mathématiques, statistiques caractérisant certaines propriétés de la séquence, etc.), difficilement compréhensibles pour les non spécialistes, en des données de haut niveau traduisant la perception humaine, données exprimées dans un langage accessible. Le système proposé a été appliqué au domaine particulier des films d'animation [CICA 06].

Dans ce chapitre nous proposons donc une méthodologie pour transformer des informations de bas niveaux, dont l'extraction a été détaillée dans les chapitres précédents, en des informations de plus haut niveau. Les techniques proposées dans la première partie de la thèse nous ont permis de définir un certain nombre de mesures numériques caractérisant : la *structure des plans* (Chapitre 2), les différentes *catégories de mouvement* (Chapitre 3) et la *distribution des couleurs* de la séquence (Chapitre 4).

L'ensemble des paramètres ainsi obtenus est converti en des termes sémantiques à l'aide de représentations de données basées sur les ensembles flous. L'avantage d'une telle représentation vient du fait que les ensembles flous permettent d'associer facilement des concepts linguistiques proches de la perception humaine à des valeurs numériques, tout en conservant

une gradualité à travers les degrés d'appartenance à ces symboles. Selon l'application, ce niveau de représentation peut ne pas être suffisant pour la caractérisation sémantique. Pour atteindre un niveau sémantique supérieur, nous pouvons alors utiliser des jeux de règles sémantiques opérant sur les symboles définis. Les règles que nous avons retenues sont du type *si/alors*, et sont construites à partir de la connaissance a priori d'experts.

L'intérêt de ces descriptions symboliques/sémantiques est de pouvoir être utilisées pour la navigation ou la recherche dans une base de films d'animation, ou encore être exploitées par les spécialistes comme des outils d'aide à l'analyse.

Dans ce chapitre, nous allons d'abord présenter les principes généraux de la logique floue que nous avons utilisés pour la caractérisation sémantique des séquences d'images, en mettant l'accent sur les avantages de ce type de représentation.

7.1 La logique floue : le concept d'incertitude

Dans la construction d'un système d'analyse, il arrive souvent que la complexité dépasse la capacité de traitement. Une solution consiste alors à introduire la notion d'incertitude : si on n'est pas capable de définir une information de manière parfaitement précise, du moins peut-on lui donner une représentation imprécise. *L'incertitude* devient une information précieuse et très utile si elle utilisée en relation avec d'autres caractéristiques du système. Généralement, en introduisant de l'incertitude dans un système, cela réduit sa complexité et augmente la crédibilité du modèle associé [Klir 95]. Dans les systèmes de traitement, l'incertitude est introduite au travers d'une formalisation basée sur la théorie des ensembles flous.

7.1.1 La formalisation basée sur la théorie des ensembles flous

Le processus de formalisation basé sur le flou comporte deux étapes de traitement :

- **les ensembles flous** : d'abord en fonction des paramètres de bas niveau dont on dispose, nous allons définir l'ensemble des variables floues ou *des ensembles flous* dont nous avons besoin pour modéliser le système.
- **les inférences floues** : puis *des inférences floues* sont utilisées pour représenter les relations modélisant le système.

Dans la suite nous allons détailler ces deux concepts.

Les ensembles flous

Le concept d'incertitude a été présenté pour la première fois dans les travaux publiés par Zadeh [Zadeh 65], mais déjà envisagé par le philosophe Max Black en 1937. Il proposait une nouvelle théorie basée sur la représentation des données par des *ensembles flous*.

Les ensembles flous sont des ensembles pour lesquels les frontières ne sont pas précises. L'appartenance des données à ce genre d'ensemble n'est pas une question d'affirmation ou de négation mais de *degré d'appartenance*. Alors que la théorie des probabilités est basée sur les valeurs logiques vrai (1) ou faux (0), dans la logique floue si A est un ensemble flou et x est un élément, la proposition *"x est inclus dans A"* n'est pas forcement vraie ou fausse comme dans le cas de la logique booléenne. La proposition est vraie avec un certain degré. Ce degré est typiquement exprimé comme une fonction d'appartenance floue, $\mu(x)$, prenant

ses valeurs dans l'intervalle $[0, 1]$ (où 0 est la négation totale et 1 l'affirmation totale).

La capacité des ensembles flous à exprimer la transition graduelle entre l'appartenance et la non appartenance fournit une représentation significative et puissante de l'incertitude de la mesure, mais également une représentation des concepts vagues qui peuvent être exprimés dans un *langage naturel*.

Pour mieux comprendre le concept d'ensemble flou nous allons l'illustrer par un exemple [Lescieux 06]. Considérons la grandeur d'une personne représentée numériquement par la valeur H exprimée en mètre. L'ensemble des valeurs possibles de H constitue l'univers de discours. En utilisant la formalisation par des ensembles flous nous allons associer le concept flou *la taille de la personne* à la valeur H. Ce concept peut, par exemple, prendre trois valeurs linguistiques : *"petite"*, *"moyenne"* et *"grande"*, constituant les sous-ensembles flous du paramètre H. Les valeurs du paramètre H sont associées aux trois symboles en utilisant des fonctions d'appartenance floues (voir Figure 7.1) : $\mu_{taille\_p}(H)$ (ligne bleu), $\mu_{taille\_m}(H)$ (ligne verte) et respectivement $\mu_{taille\_g}(H)$ (ligne rouge).

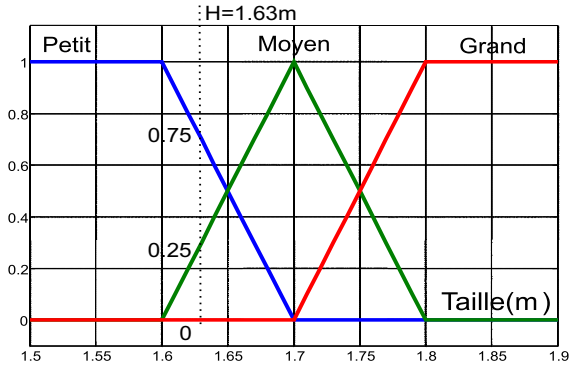


FIG. 7.1 – La partition floue de l'univers de discours pour le concept *la taille de la personne*. Exemple de degré d'appartenance à chaque valeur symbolique pour $H = 1.63m$.

Dans le langage courant, on dira d'une personne qui mesure $1.63m$ qu'elle est de *"petite"* taille, en attachant implicitement une valeur de vérité de 1 à cette information (c'est le cas de la logique booléenne). La formalisation par des ensembles flous se traduit alors par : *la taille de la personne est petite* avec un degré d'appartenance de 0.75, *la taille de la personne est moyenne* avec un degré de 0.25 et *la taille de la personne est grande* avec un degré de 0 (voir la Figure 7.1). Les degrés d'appartenance sont souvent des fonctions affines par morceaux, et le choix des seuils définissant la manière dont ces degrés d'appartenance sont associés aux valeurs numériques est très dépendante de l'application. Dans la plupart des situations les fonctions d'appartenance sont définies *empiriquement* en utilisant une connaissance a priori portant sur l'évolution des paramètres analysés, cette connaissance étant souvent fournie par des spécialistes du domaine étudié. Dans notre exemple, nous savons, par expérience, qu'une personne est de petite taille si $H < 1.60m$, de taille moyenne si $H \simeq 1.70m$, et de grande taille si $H > 1.80m$.

L'inférence floue

Disposant de symboles flous, la prise de décision ou la définition de concepts linguistiques de plus haut niveau se fait souvent en utilisant des jeux de règles du type *si/alors*.

L'inférence floue nécessaire peut alors être réalisée selon le principe de combinaison projection [Zadeh 65] qui repose sur l'utilisation d'opérateurs spécifiques de conjonction et de disjonction. La construction de ces jeux de règles, qui est la formalisation des relations entre les symboles flous, est souvent la traduction d'une expertise ou d'un apprentissage.

Une base de règles est caractérisée par certaines propriétés [Guillaume 01], [Klir 95] :

- **le respect de la sémantique** : les sous-ensembles flous doivent pouvoir être interprétés en termes linguistiques,
- **la cohérence** : les conclusions des règles utilisées simultanément ne doivent pas être contradictoires,
- **la continuité** : de petites variations sur l'entrée ne doivent pas impliquer de grandes variations sur la sortie,
- **la complétude** : assure que chacune des valeurs non nulles des entrées, active au moins une règle, et qu'une valeur de sortie est inférée dans tous les cas.

Pour plus de détails sur la génération de règles floues et sur la logique floue on pourra se rapporter à la thèse [Guillaume 01] ou au livre [Klir 95].

7.1.2 Les domaines d'application

En permettant une représentation proche de la façon dont on perçoit la réalité, la logique floue a vite gagné du terrain dans beaucoup de domaines d'activités, remplaçant partiellement ou totalement la logique booléenne.

Ainsi, on trouve la logique floue dans l'aide à la décision ou au diagnostic, les bases de données (objets flous et/ou requêtes floues), la reconnaissance de formes, la vision par ordinateur, le traitement d'images, l'agrégation multi-critères et l'optimisation, la commande des systèmes, etc. La généralisation de cette technique a aussi été introduite dans certains produits de la vie courante, comme les appareils électroménagers ou les systèmes audiovisuels, mais aussi dans des processeurs dédiés et des interfaces de développement spécifiques, comme le processeur Motorola 68HC12 ou Thomson WARP [Lescieux 06].

Dans le domaine du traitement des images et de la vision par ordinateur, l'apport de la logique floue est incontestable. Elle est le pont entre les données de bas niveau acquises après différents traitements et leur signification. Les règles floues *sont énoncées en langage naturel* ce qui donne plus de *sens* aux valeurs numériques. Par exemple, il sera plus naturel de dire que les couleurs dans une image sont foncées que de dire que le pourcentage des couleurs foncées de la séquence est supérieur à 90% du total des couleurs utilisées. Dans la suite, nous allons illustrer les avantages de la représentation floue des données.

7.1.3 Les avantages de la représentation floue

Les avantages de la formalisation à l'aide de la logique floue peuvent être synthétisés par :

- **approximation universelle** : l'univers de discours, qui peut être très vaste ou même infini, est converti en un nombre limité de concepts à travers la formalisation floue. L'histoire a montré que les systèmes d'inférence floue sont aussi performants que d'autres techniques d'approximation [Wang 92].

- **réduction de la complexité** du système : dans le cas où la quantité d'informa-

tions disponibles est trop élevée pour qu'elle soit contrôlée en totalité et lorsque la compréhension des processus est limitée, la formalisation floue permet de diminuer la complexité du système en introduisent la notion d'incertitude. Généralement, en tolérant plus d'incertitude dans le modèle du système, la complexité a tendance à diminuer et la crédibilité du modèle augmente [Klir 95].

- **représentation de la réalité** : à l'aide du concept d'incertitude les variables floues représentent la réalité, qui est typiquement incertaine, de manière plus fidèle que les variables classiques nettes. Cette propriété a été bien exprimée par Albert Einstein en 1921 : *bien que les lois des mathématiques se rapportent à la réalité, elles ne sont pas certaines. Et bien qu'elles soient certaines, elles ne se rapportent pas nécessairement à la réalité.*

- **langage naturel** : dans la logique floue les concepts vagues sont représentés dans un langage naturel. Cette propriété est une des plus importantes. La formalisation floue fait la conversion entre les mesures numériques et les concepts linguistiques proches de notre mode de perception. Si l'univers de discours est numérique, la représentation floue se fait à travers des variables linguistiques prenant des valeurs linguistiques (voir textuelles) [Lescieux 06].

- **respect de la sémantique** : le fonctionnement de la logique floue est similaire à notre façon de percevoir. Notre cerveau fonctionne en logique floue. Il apprécie les variables d'entrée continues de façon approximative (par exemple faible, élevé, loin, proche). La formalisation floue respecte donc la sémantique en permettant l'extraction de la connaissance à partir des données de bas niveau.

- **cohérence** : le fait que la formalisation floue soit construite à travers l'expertise et la connaissance a priori sur les données et leurs relations confère plus de cohérence au modèle que lors d'une représentation classique.

- **normalisation des valeurs** : les concepts linguistiques sont exprimés en utilisant des valeurs de vérité qui sont normalisées entre 0 et 1. La tâche de comparaison ou de fusion entre les différentes données se trouve alors simplifiée.

- **généralisation de la logique booléenne** : la représentation classique utilisant la logique booléenne est un cas particulier de la logique floue. En effet, la formalisation floue contient la formalisation nette.

Dans la suite nous allons utiliser la formalisation floue pour extraire des informations sémantiques à partir des paramètres de bas niveau extraits des séquences d'animations.

7.2 La sémantique des couleurs

Dans cette section, nous allons utiliser la formalisation floue des paramètres de couleur de bas niveau pour caractériser la distribution des couleurs d'une séquence d'images d'un point de vue sémantique. Les informations sémantiques extraites portent sur les techniques artistiques d'utilisation des couleurs et plus particulièrement sur *les contrastes couleur d'Itten*

[Itten 61] et *l'harmonie des couleurs* [Birren 69], techniques qui sont présentes dans les films d'animation de [CICA 06].

Dans les films d'animation, au moment de la création du film, les couleurs sont choisies et mélangées par l'artiste en utilisant différents concepts artistiques. Cela permet à l'auteur de transmettre par l'image certains sentiments ou certaines sensations particulières comme par exemple la chaleur, l'harmonie, le contraste, la joie, la tristesse, etc.

7.2.1 Le calcul des paramètres couleurs de haut niveau

Pour définir ces paramètres de plus haut niveau, nous allons adopter une démarche en trois temps. Dans un premier temps, notre point de départ est l'histogramme couleur global pondéré, $h_{seq}()$, qui synthétise les informations sur la distribution globale des couleurs de la séquence (équation 4.4, Section 4.2). Cet histogramme étant défini sur la palette "Webmaster" (voir la Section 4.2.3) où chaque couleur est définie par des noms, il est alors possible d'extraire de cet histogramme un certain nombre de caractéristiques statistiques de haut niveau sur les couleurs de la séquence.

La deuxième étape est un travail similaire mais effectué sur un nouvel histogramme, *l'histogramme des couleurs élémentaires* de la séquence, $h_{élém}$. Cet histogramme est construit à partir de l'histogramme couleur global pondéré, et il a pour objectif de donner une représentation compacte des couleurs élémentaires de la séquence. Le troisième temps est la représentation symbolique floue des caractéristiques extraites de ces deux histogrammes.

L'histogramme des couleurs élémentaires $h_{élém}()$

Comme nous l'avons déjà dit, une particularité importante des films d'animation est que chaque film possède sa propre distribution de couleurs (voir la Section 1.5). La plupart des films utilisent une palette réduite de couleurs. Ces couleurs prédominantes constituent une caractéristique discriminante de chaque film.

En utilisant le dictionnaire des noms des couleurs fournis par la palette "Webmaster" nous allons définir un *histogramme des couleurs élémentaires* de la séquence, $h_{élém}$, qui est extrait à partir de l'histogramme global pondéré de la manière suivante :

$$h_{élém}(c_e) = \sum_{c=1}^{216} h_{seq}(c)|_{\{Nom(c_e) \subset Nom(c)\}} \tag{7.1}$$

où c_e est l'indice d'une couleur issue de l'ensemble des couleurs élémentaires de la palette "Webmaster", $\Gamma_{élém} = \{$*"Orange", "Red", "Pink", "Magenta", "Violet", "Blue", "Azure", "Cyan", "Teal", "Green", "Spring", "Yellow", "Gray", "White", "Black"*$\}$[1], $c_e = 1, ..., 15$ (12 couleurs + Blanc, Noir et Gris), $c = 1, ..., 216$ est l'indice d'une couleur de la palette "Webmaster", $h_{seq}()$ est l'histogramme global pondéré et $Nom(c)$ est l'opérateur qui retourne le nom associé à la couleur d'indice c.

La palette "Webmaster" a l'avantage de proposer les mêmes couleurs élémentaires que celles de la roue couleur d'Itten (voir la Figure 4.7). Cette particularité nous permettra d'analyser les relations perceptuelles entre les couleurs (cette partie sera détaillée dans la suite de ce mémoire).

[1] le Blanc, le Noir et le Gris ont été nommés artificiellement couleur dans ce cas.

Le fait que l'histogramme $h_{élém}()$ est calculé à partir de l'histogramme $h_{seq}()$ nous assure que les valeurs de $h_{élém}()$ représentent le pourcentage d'apparition des couleurs élémentaires de l'ensemble $\Gamma_{élém}$ dans la séquence. Lorsque les couleurs de la séquence sont obtenues par un mélange de plusieurs couleurs élémentaires, leur apport dans l'histogramme $h_{élém}()$ sera sur chacune des couleurs élémentaires composant ce mélange. Par exemple la couleur "Medium Azur-Cyan" contribue aux valeurs $h_{élém}(c_a)$ et $h_{élém}(c_c)$ où c_a, c_c sont les indices des couleurs élémentaires "Azur" et "Cyan" dans l'ensemble $\Gamma_{élém}$.

Pour construire l'histogramme $h_{élém}(c_e)$, une couleur de la séquence est projetée sur sa couleur élémentaire d'origine en négligeant les variations de saturation et d'intensité de la couleur. Ce mécanisme confère à l'histogramme $h_{élém}()$ l'avantage d'être invariant aux variations de saturation ou d'intensité des couleurs. Ainsi, un rouge foncé et un rouge clair seront représentés par la couleur rouge dans $h_{élém}()$. En pratique, l'histogramme des couleurs élémentaires nous permet également de normaliser la distribution des couleurs, facilitant ainsi la recherche des séquences dans une base de films d'animation sur des critères de ressemblance entre couleurs (voir la classification des films d'animation selon les couleurs prédominantes dans le Chapitre 8).

Les paramètres de plus haut niveau proposés dans ce chapitre sont donc divisés en deux catégories (voir [Ionescu 05h] et le rapport [Ionescu 05c]) :

- les paramètres extraits à partir de l'histogramme global pondéré, $h_{seq}()$,
- les paramètres extraits à partir de l'histogramme des couleurs élémentaires, $h_{élém}()$.

Les paramètres extraits de $h_{seq}()$

Le premier paramètre calculé sur l'histogramme global pondéré, $h_{seq}()$, est la *variété des couleurs* du film, P_{var}. Ce paramètre est lié à la richesse de la palette couleur utilisée dans le film, et est défini par le nombre de couleurs *différentes* et *significatives* de la distribution des couleurs de la séquence entière :

$$P_{var} = \frac{Card\{c/\ h_{seq}(c) > \tau_c\}}{216} \tag{7.2}$$

où $c = 1, ..., 216$ est l'indice d'une couleur de la palette "Webmaster", $Card()$ est l'opérateur cardinal qui retourne le nombre d'éléments d'un ensemble et τ_c est le seuil utilisé pour définir l'importance d'une couleur par rapport à la distribution globale. La valeur du seuil a été déterminée empiriquement à $\tau_c = 0.01$, soit 1%, après avoir analysé plusieurs films d'animation de [CICA 06].

Les paramètres suivants proposés sont liés à l'intensité, à la saturation et à la teinte des couleurs de la séquence.

Le *coefficient de couleurs claires*, $P_{claires}$, est calculé comme étant la proportion des couleurs claires présentes dans la séquence. Dans la palette "Webmaster", l'attribut clarté associé à une couleur apparaît dans son nom par l'utilisation des mots *"light"* ou *"pale"*. Le cas particulier du Blanc, qui représente la couleur la plus claire possible, doit être pris en compte dans le calcul de $P_{claires}$. Le paramètre $P_{claires}$ est défini par :

$$\begin{aligned} P_{claires} &= \sum_{c=1}^{216} h_{seq}(c)|_{\{Mot_{claire} \subset Nom(c)\}} \\ Mot_{claire} &\in \{"light", "pale", "white"\} \end{aligned} \tag{7.3}$$

où c est l'indice d'une couleur et $Nom(c)$ est l'opérateur qui retourne le nom associé à la couleur d'indice c. Ainsi, une couleur est considérée comme étant claire si son nom contient l'un des mots "light", "pale" ou "white".

Les autres paramètres sont définis sur le même principe. Ainsi, le *coefficient de couleurs foncées*, $P_{foncées}$, représente la proportion des couleurs sombres présentes dans la séquence. L'attribut foncé associé à une couleur apparaît dans le dictionnaire des noms à travers les mots *"dark"* ou *"obscure"*. Le Noir, qui représente la couleur la plus foncée possible, est également pris en compte dans le calcul de $P_{foncées}$. Une couleur est considérée comme étant foncée si son nom contient l'un des mots "dark", "obscure", ou "black".

Le *coefficient de couleurs fortes*, P_{fortes} reflète la proportion de couleurs saturées présentes dans la séquence. Dans le dictionnaire des noms, les couleurs saturées utilisent les mots *"hard"* ou *"faded"*. Par définition, une couleur élémentaire ou pure est une couleur ayant une saturation de 100%. Les 12 couleurs élémentaires de la palette "Webmaster" contribuent donc également à la valeur du paramètre P_{fortes}. Une couleur est considérée comme saturée si son nom contient l'un des mots "hard" ou "faded", ou si elle est une couleur élémentaire de l'ensemble $\Gamma_{élém}$ (auquel on soustrait le Blanc, le Noir et le Gris).

Le *coefficient de couleurs faibles*, $P_{faibles}$, par opposition au paramètre P_{fortes}, représente la proportion des couleurs de la séquence, ayant une faible saturation. Dans le dictionnaire des noms des couleurs, une saturation faible est exprimée en employant les mots *"dull"* ou *"weak"*. Une couleur est considérée comme ayant une saturation faible si son nom contient l'un des mots "dull" ou "weak".

Les deux paramètres suivants permettent de mesurer la sensation de chaud ou de froid qui est associée aux couleurs. Il est bien connu, en particulier dans le domaine de la peinture, que certaines couleurs sont considérées comme dégageant une certaine chaleur, ou au contraire une sensation de froid. Sur ce principe, en utilisant les noms de la palette "Webmaster", nous allons donc définir les proportions de *couleurs chaudes* et de *couleurs froides*. En correspondance avec la roue couleur d'Itten, dans la palette "Webmaster" les couleurs chaudes sont distribuées sur la première moitié de la roue, de la couleur "Spring" à la couleur "Magenta" en passant par la couleur "Yellow" (voir la Figure 4.7 dans la Section 4.2.3). Quant aux couleurs froides, elles sont distribuées sur la deuxième moitié de la roue, de "Violet" à "Green".

Le *coefficient de couleurs chaudes*, $P_{chaudes}$, est la proportion de couleurs chaudes présentes dans la séquence. Dans l'art, les couleurs considérées comme étant chaudes sont les couleurs appartenant à l'ensemble Γ_{chaud} = {*"Yellow", "Orange", "Red", "Yellow Orange", "Red Orange", "Red Violet", "Magenta", "Pink" and "Spring"*} [Lay 04]. Une couleur de la séquence est considérée comme chaude si son nom contient l'un des mots de l'ensemble Γ_{chaud}.

Par opposition, le *coefficient de couleurs froides*, $P_{froides}$, représente la proportion de couleurs froides présentes dans la séquence. Dans l'art, les couleurs considérées comme étant froides sont les couleurs de l'ensemble Γ_{froid} = {*"Green", "Blue","Violet", "Yellow Green", "Blue Green", "Blue Violet", "Teal", "Cyan", "Azure"*} [Lay 04]. Une couleur est considérée comme étant froide si son nom contient l'un des mots de l'ensemble Γ_{froid}.

En conclusion, à partir de l'histogramme global pondéré, $h_{seq}()$, nous avons pu définir les paramètres suivants :

– P_{var} : la variété couleur,

- $P_{claires}$ et $P_{foncées}$: les proportions de couleurs claires et sombres de la séquence,
- P_{fortes} et $P_{faibles}$: les proportions de couleurs saturées et non saturées de la séquence,
- $P_{chaudes}$ et $P_{froides}$: les proportions de couleurs chaudes et froides de la séquence.

Le fait que l'histogramme $h_{seq}()$ représente la probabilité d'apparition d'une couleur dans la séquence (voir la Section 4.2.4) nous assure que toutes les valeurs des six paramètres définis ci-dessus sont normalisées entre 0 et 1. De même, la valeur du paramètre P_{var} est normalisée par le nombre total de couleurs (216) et est donc aussi normalisée entre 0 et 1.

Les paramètres extraits de $h_{élém}()$

De la même manière, nous allons extraire quelques paramètres statistiques de l'histogramme des couleurs élémentaires $h_{élém}()$.

Le premier paramètre est la *diversité couleur* de la séquence, noté P_{div}. C'est le rapport entre le nombre de couleurs élémentaires *différentes* et *significatives* de la distribution des couleurs de la séquence et le nombre total de couleurs élémentaires. Dans cette définition, le nombre total de couleurs élémentaires est égal à 13, car sur les 15 couleurs élémentaires définies précédemment (voir $\Gamma_{élém}$ de l'équation 7.1), nous avons considéré le Blanc, le Noir et le Gris comme un seul niveau de gris. Ainsi, le paramètre P_{div} est défini par :

$$P_{div} = \frac{Card\{c_e / h_{élém}(c_e) > \tau_e\}}{13} \tag{7.4}$$

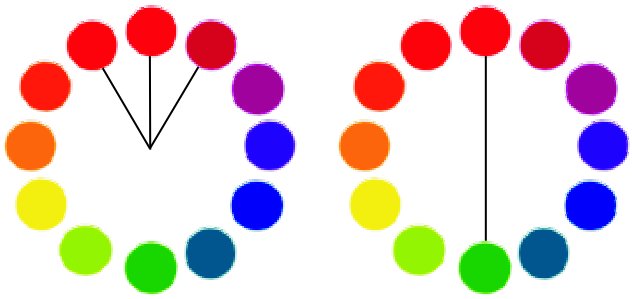
où c_e est l'indice d'une couleur élémentaire de la palette "Webmaster", $c_e = 1, ..., 13$ (12 couleurs élémentaires + Gris), $Card()$ est l'opérateur cardinal qui retourne le nombre d'éléments d'un ensemble et la valeur du seuil $\tau_e = 0.04$ a été déterminée empiriquement après avoir analysé un certain nombre de films d'animation de [CICA 06].

On peut remarquer que P_{div} apporte une information voisine de celle contenue dans le paramètre P_{var} défini dans l'équation 7.2. Néanmoins ce regroupement en 13 couleurs élémentaires apporte un nouvel éclairage sur l'analyse de la distribution des couleurs, comme nous allons le voir dans la suite.

Les deux paramètres suivants sont liés aux relations perceptuelles qui existent entre les couleurs. Comme nous l'avons dit dans la Section 4.1.2, les couleurs élémentaires sont souvent présentées sur une roue de couleurs, telle que la roue d'Itten, sur laquelle la disposition des couleurs permet une perception progressive. Ce type de représentation permet donc d'étudier les différentes relations perceptuelles qui existent entre les couleurs, en particulier les deux relations fondamentales que sont la *complémentarité* et *l'adjacence* entre les couleurs.

La relation de *complémentarité* est liée au concept de contraste entre les couleurs et caractérise les relations entre les teintes de couleurs opposées. Deux couleurs sont considérées comme complémentaires si elles sont opposées du point de vue de la perception. Lorsqu'elles sont utilisées ensemble, cela accentue le contraste entre les couleurs (voir la Section 4.1.2). En utilisant la roue des couleurs d'Itten, une ligne droite passant par le centre de la roue est utilisée pour déterminer les paires de couleurs complémentaires (voir la Figure 7.2.b). Les *couleurs analogues* ou adjacentes sont, au contraire, des couleurs qui sont proches du point de vue de la perception. Sur la roue des couleurs d'Itten, les couleurs adjacentes sont des couleurs voisines deux à deux (voir la Figure 7.2.a).

En utilisant le fait que la palette des couleurs "Webmaster" est structurée de la même façon que la roue des couleurs d'Itten (voir la Figure 4.7 dans la Section 4.2.3) nous allons



(a) (b)

FIG. 7.2 – Les relations entre les couleurs sur la roue des couleurs d'Itten : (a) Relation d'adjacence (adjacence de Rouge), (b) Relation de complémentarité (Rouge-Vert).

définir les deux paramètres suivants. Le *rapport des couleurs adjacentes* de la séquence, P_{adj}, et le *rapport des couleurs complémentaires* de la séquence, P_{compl}, représentant respectivement la proportion des couleurs adjacentes et complémentaires présentes dans la séquence. P_{adj} et P_{compl} sont définis par :

$$P_{adj} = \frac{Card\{c_e/\ \exists\ c'_e \neq c_e;\ c'_e, c_e\ adjacentes;\ h_{élém}(c_e), h_{élém}(c'_e) > \tau_e\}}{2 \cdot Card\{c_e/h_{élém}(c_e) > \tau_e\}} \quad (7.5)$$

$$P_{compl} = \frac{Card\{c_e/\ \exists\ c'_e \neq c_e;\ c'_e, c_e\ complémentaires;\ h_{élém}(c_e), h_{élém}(c'_e) > \tau_e\}}{2 \cdot Card\{c_e/h_{élém}(c_e) > \tau_e\}} \quad (7.6)$$

où c_e et c'_e sont les indices des couleurs élémentaires adjacentes pour P_{adj} et complémentaires pour P_{compl}. Dans ces définitions, $c_e, c'_e = 1, ..., 12$, car le Blanc, le Noir et le Gris ne sont pas pris en compte. De plus, les couleurs c_e et c'_e sont considérées comme étant significatives pour la distribution des couleurs de la séquence, si elles ont un pourcentage d'apparition supérieur à 4% dans la séquence, donc si $\tau_e = 0.04$.

Dans la définition de ces proportions, pour des raisons techniques, la normalisation est effectuée par le double du nombre de couleurs élémentaires significatives. En effet, si les couleurs c_{e1} et c_{e2} sont complémentaires, l'algorithme utilisé détectera deux complémentarités, une entre c_{e1} et c_{e2} et une seconde entre c_{e2} et c_{e1}.

En conclusion, à partir de l'histogramme des couleurs élémentaires, $h_{élém}()$, nous avons pu définir les paramètres suivants :

- P_{div} : la diversité couleur,
- P_{adj} et P_{compl} : le rapport des couleurs adjacentes et respectivement complémentaires de la séquence.

Les paramètres de couleurs que nous avons proposés dans cette section ont été déterminés après avoir analyser le contenu de plusieurs films d'animation de la base de données de [CICA 06]. Nous avons constaté que l'intensité couleur, la saturation et la chaleur jouent un rôle important dans la caractérisation des films d'animation. Ce sont des informations qui nous permettront de faire la différence entre différents genres ou techniques d'animation. Par exemple, les films d'animation utilisant la technique de la pâte à modeler ont typiquement une palette de couleurs sobres et plutôt froides (ceci est lié au matériel utilisé, voir Figure 8.5 dans la Section 8.2.3).

De plus, chaque film d'animation utilise une palette particulière de couleurs, donc une autre information discriminante est la variété/diversité des couleurs. Enfin, la relation perceptuelle entre les couleurs est nécessaire pour faire la différenciation entre les différents schémas de couleurs : couleur adjacentes, complémentaires, etc (voir Figure 8.5 dans la Section 8.2.3).

7.2.2 La caractérisation sémantique floue des couleurs

A partir des paramètres numériques proposés dans la section précédente nous allons extraire des informations sur la perception visuelle de la distribution des couleurs des films d'animation de [CICA 06]. La méthodologie utilisée est la représentation des données à partir des ensembles flous (voir [Ionescu 06c]).

La formalisation floue nous permet d'exprimer des données numériques sous forme de *concepts sémantiques* proches de la perception humaine. Les concepts que nous proposons, sont définis à partir de la connaissance a priori de certains experts du domaine des films d'animation. La formalisation floue peut être effectuée de deux manières. Cela peut être une simple caractérisation symbolique floue des grandeurs numériques. Mais, pour la définition de concepts plus élaborés, cette formalisation floue peut résulter de l'utilisation de jeux de règles floues opérant sur les symboles flous.

La caractérisation symbolique floue

Globalement, la caractérisation sémantique des couleurs proposée concerne la *perception des couleurs*, *la variation/diversité des couleurs*, les *contrastes couleurs d'Itten* et *l'harmonie des couleurs* (voir la Section 4.1.2).

Pour donner une représentation symbolique floue de ces paramètres de bas niveau, nous les avons regroupés en trois catégories, selon leur contenu sémantique :

- **catégorie 1** : catégorie concernant les propriétés des couleurs (intensité, saturation et perception). Elle contient les paramètres $P_{claires}$, $P_{foncées}$, P_{fortes}, $P_{faibles}$, $P_{chaudes}$ et $P_{froides}$,
- **catégorie 2** : catégorie concernant la richesse en terme de couleurs de la séquence (variété et diversité). Elle contient les paramètres P_{var} et P_{div},
- **catégorie 3** : catégorie concernant les relations entre les couleurs (adjacence et complémentarité). Elle contient les paramètres P_{adj} et P_{compl}.

Catégorie 1 *: description de l'intensité, de la saturation et de la perception.*

Le concept linguistique *présence de couleurs claires* est associé au paramètre $P_{claires}$ qui représente la proportion de couleurs claires présentes dans la séquence. Ce concept est décrit en utilisant trois valeurs linguistiques illustrées par les symboles suivants : *"présence faible de couleurs claires"*, *"présence moyenne de couleurs claires"* et *"présence élevée de couleurs claires"*. La signification floue de chaque symbole est traduite par sa fonction d'appartenance. La partition floue de l'univers de discours, $P_{claires}$, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\alpha_{faible_{cl}}$, $\alpha_{moyen_{cl}}$ et $\alpha_{élevé_{cl}}$, et est illustrée dans la Figure 7.3.

Les fonctions d'appartenance aux symboles ont été définies d'une manière classique en utilisant des fonctions linéaires par morceaux. Cette définition est basée sur le choix de 4 valeurs de seuils, $\{33, 50, 60, 66\}$. Ces seuils ont été déterminés empiriquement après avoir

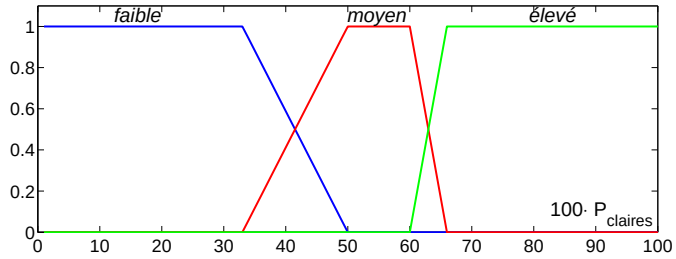


FIG. 7.3 – Partition floue de l'univers de discours $P_{claires}$ déterminée par les fonctions d'appartenance : $\alpha_{faible_{cl}}$ (bleu), $\alpha_{moyen_{cl}}$ (rouge) et $\alpha_{élevé_{cl}}$ (vert) (l'axe oY correspond au degré d'appartenance).

analysé plusieurs films d'animation de la base de données de [CICA 06]. Nous avons ainsi trouvé qu'une séquence a une faible distribution de couleurs claires (vérité 1) si $100 \cdot P_{claires} < 33\%$, une distribution moyenne de couleurs claires (vérité 1) si $100 \cdot P_{claires} > 50\%$ et $100 \cdot P_{claires} < 60\%$ et une distribution élevée de couleurs claires (vérité 1) si $100 \cdot P_{claires} > 66\%$.

En utilisant le même principe nous allons définir les concepts linguistiques suivantes :

- **présence des couleurs foncées** : associé au paramètre $P_{foncées}$ qui représente la proportion de couleurs foncées présentes dans la séquence,
- **présence des couleurs saturées** : associé au paramètre P_{fortes} qui représente la proportion des couleurs saturées présentes dans la séquence,
- **présence des couleurs faiblement saturées** : est associé au paramètre $P_{faibles}$ qui représente la proportion de couleurs faiblement saturées présentes dans la séquence,
- **présence des couleurs chaudes** : associé au paramètre $P_{chaudes}$ qui représente la proportion des couleurs chaudes présentes dans la séquence,
- **présence des couleurs froides** : associé au paramètre $P_{froides}$ qui représente la proportion des couleurs froides présentes dans la séquence.

Les fonctions d'appartenance floue associées aux valeurs linguistiques des concepts énumérés ci-dessus (3 degrés : faible, moyen et élevé) ont été définis en utilisant les mêmes seuils et le même raisonnement que pour le concept *présence de couleurs claires* (voir dans la Figure 7.3).

Catégorie 2 *: description de la richesse des couleurs.*

Les concepts proposés dans cette catégorie sont liés à la variété et à la diversité des couleurs présentes dans la séquence. Le concept linguistique *variété des couleurs* est donc associé au paramètre P_{var} qui est une mesure de la richesse des couleurs utilisées dans la séquence (du total de 216). Le concept *variété des couleurs* est décrit par trois valeurs symboliques : *"variété des couleurs faible"*, *"variété des couleurs moyenne"* et *"variété des couleurs élevée"*. La partition floue de l'univers de discours P_{var}, traduite par les trois fonctions d'appartenance floues à chaque symbole, α_{faible_v}, $\alpha_{moyenne_v}$ et $\alpha_{élevée_v}$, est illustrée dans la Figure 7.4.

Comme pour les autres partitions floues, les valeurs des seuils utilisés pour la définition des trois fonctions, $\{30, 36, 60, 66\}$ ont été déterminées empiriquement après avoir analysé plusieurs films d'animation de la base de données de [CICA 06]. Nous avons en effet pu

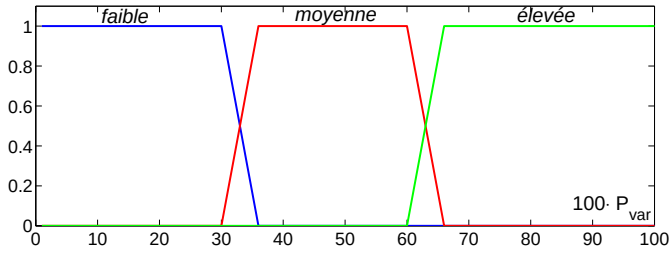


FIG. 7.4 – La partition floue de l'univers de discours P_{var} déterminée par les fonctions d'appartenance floues : α_{faible_v} (bleu), $\alpha_{moyenne_v}$ (rouge) et $\alpha_{élevée_v}$ (vert) (l'axe oY correspond au degré d'appartenance).

observer que la séquence a une variété des couleurs faible (vérité 1) si la séquence utilise moins de 30% du total de 216 couleurs disponibles (< 65 couleurs), soit $100 \cdot P_{var} < 30\%$, une variété des couleurs moyenne (vérité 1) si la séquence utilise entre 77 et 129 couleurs, soit $100 \cdot P_{var} > 36\%$ et $100 \cdot P_{var} < 60\%$, et une variété des couleurs élevée (vérité 1) si la séquence utilise plus de 142 couleurs, soit $100 \cdot P_{var} > 66\%$.

En utilisant le même principe nous avons défini le concept linguistique *diversité des couleurs*. Il est associé au paramètre P_{div}, qui est une mesure de la diversité des couleurs élémentaires utilisés par la séquence. Les valeurs symboliques décrivant ce concept sont également : *"diversité des couleurs faible"*, *"diversité des couleurs moyenne"* et *"diversité des couleurs élevée"* et sont définies de la même manière que précédemment (voir la Figure 7.4).

Catégorie 3 *: description des relations entre les couleurs.*

Les concepts proposés dans cette catégorie sont liés aux relations d'adjacence et de complémentarité qui existent entre les couleurs. Le concept linguistique *couleurs adjacentes* est associé au paramètre P_{adj}. Il représente la proportion de couleurs élémentaires, sur la séquence entière, proches du point de vue de la perception.

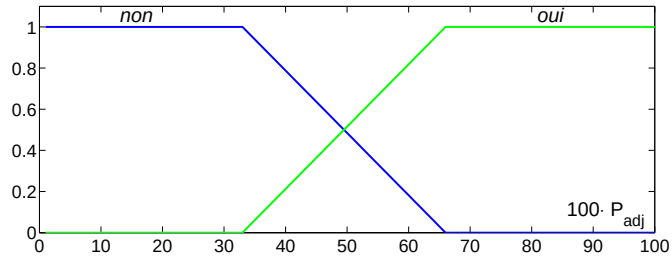


FIG. 7.5 – La partition floue de l'univers de discours P_{adj} déterminée par les fonctions d'appartenance floues : α_{non_a} (bleu) et α_{oui_a} (vert) (l'axe oY correspond au degré d'appartenance).

Dans ce cas le concept d'adjacence n'est décrit que par deux valeurs symboliques : *"oui"* ou *"non"*. L'expertise et l'expérience ont en effet montré que deux valeurs symboliques suffisaient pour décrire ce concept. La partition floue de l'univers de discours P_{adj}, déterminée par l'ensemble des fonctions d'appartenance floues associées à chaque symbole, α_{oui_a} et

respectivement α_{non_a}, est illustrée par la Figure 7.5.

Les valeurs des seuils utilisés pour la définition des deux fonctions, soient $\{33, 66\}$, ont été également déterminées empiriquement, l'univers de discours étant divisé en deux parties égales.

La relation de complémentarité (couleurs opposées du point de vue de la perception, voir Figure 7.2) est introduite par le concept linguistique *couleurs complémentaires*. Il est associé au paramètre P_{compl} en utilisant le même raisonnement que dans le cas précédent.

En conclusion, en utilisant les paramètres numériques caractérisant les propriétés des couleurs (intensité, saturation et chaleur), la richesse couleur et les relations entre couleurs, nous avons pu définir les concepts linguistiques suivants :

- **présence de couleurs claires**
- **présence de couleurs foncées**
- **présence de couleurs saturées**
- **présence de couleurs faiblement saturées**
- **présence de couleurs chaudes**
- **présence de couleurs froides**
- **variété des couleurs**
- **diversité des couleurs**
- **couleurs adjacentes**
- **couleurs complémentaires**

Ceci constitue le premier niveau symbolique de caractérisation de la distribution des couleurs d'une séquence d'images.

La caractérisation sémantique à partir des règles floues

Pour acquérir un niveau sémantique supérieur et définir de nouveaux concepts linguistiques nous allons utiliser des règles floues appliquées sur les symboles définis dans la section précédente. Les informations sémantiques envisagées sont certains contrastes de couleurs d'Itten et différents schémas sur l'harmonie des couleurs (voir la Section 4.1.2).

Tout d'abord, la séquence est caractérisée du point de vue de la prédominance de *l'intensité des couleurs*. De nouveaux concepts linguistiques sont définis à partir des concepts linguistiques *présence de couleurs claires* et *présence de couleurs foncées*, en utilisant les règles floues suivantes :

∘ ***SI*** {(**présence de couleurs claires** *est* "présence faible de couleurs claires") ***ET*** (**présence de couleurs foncées** *est* "présence élevée de couleurs foncées")}
ALORS "les couleurs prédominantes sont foncées".

∘ ***SI*** {(**présence de couleurs claires** *est* "présence élevée de couleurs claires") ***ET*** (**présence de couleurs foncées** *est* "présence faible de couleurs foncées")}
ALORS "les couleurs prédominantes sont claires".

∘ ***SI*** {(**présence de couleurs claires** *est* "présence moyenne de couleurs claires") ***ET*** (**présence de couleurs foncées** *est* "présence moyenne des couleurs foncées")}
ALORS "il y a un contraste claire-foncé".

Les descriptions floues des nouveaux concepts sont obtenus en employant le mécanisme

de combinaison/projection de Zadeh [Zadeh 65]. Nous avons utilisé l'opérateur $min()$ comme opérateur de conjonction ET et l'opérateur $max()$ comme opérateur de disjonction OU . Par exemple, la fonction d'appartenance au nouveau symbole *"il y a un contraste clair-foncé"*, $\alpha_{cont-c/f}$, est définie par :

$$\alpha_{cont-c/f}(P_{claires}, P_{foncées}) = min(\alpha_{moyen_{cl}}(P_{claires}), \alpha_{moyen_{fc}}(P_{foncées})) \qquad (7.7)$$

où $\alpha_{moyen_{cl}}$ et $\alpha_{moyen_{fc}}$ sont les degrés floues d'appartenance aux symboles *"présence moyenne de couleurs claires"* et *"présence moyenne de couleurs foncées"*. Les fonctions d'appartenance floues aux autres symboles seront définies de la même façon.

En ce qui concerne les autres combinaisons des deux concepts linguistiques, elles ne sont par prises en compte. Dans ces situations particulières, une caractérisation précise de la séquence n'est pas possible. On peut dans ce cas utiliser le symbole *"pas de description"* (noté PD). Les jeux de règles utilisés pour la caractérisation de l'intensité couleur sont résumés dans le Tableau 7.1.

$\downarrow P_{foncées} \mid P_{claires} \rightarrow$	***faible***	***moyen***	***élevé***
faible	PD	PD	prédominance des couleurs claires
moyen	PD	contraste clair-foncé	PD
élevé	prédominance des couleurs foncées	PD	PD

TAB. 7.1 – Les jeux de règles utilisés pour la caractérisation de l'intensité des couleurs (PD : pas de description).

En utilisant le même raisonnement nous allons définir des informations sémantiques liées à la caractérisation de la *saturation* des couleurs prédominantes de la séquence, à la caractérisation de la *chaleur* transmise par les couleurs prédominantes de la séquence et à la caractérisation des *relations perceptuelles* qui existent entre les couleurs prédominantes. les règles utilisées sont les suivantes :

◦ ***SI*** {(**présence de couleurs saturées** *est* "présence faible de couleurs saturées") ***ET*** (**présence de couleurs faiblement saturées** *est* "présence élevée de couleurs faiblement saturées")} ***ALORS*** "les couleurs prédominantes ont une saturation faible".

◦ ***SI*** {(**présence de couleurs saturées** *est* "présence élevée de couleurs saturées") ***ET*** (**présence de couleurs faiblement saturées** *est* "présence faible de couleurs faiblement saturées")} ***ALORS*** "les couleurs prédominantes sont saturées".

◦ ***SI*** {(**présence de couleurs saturées** *est* "présence moyenne de couleurs saturées") ***ET*** (**présence de couleurs faiblement saturées** *est* "présence moyenne de couleurs faiblement saturées")} ***ALORS*** "il y a un contraste de saturation".

◦ ***SI*** {(**présence de couleurs chaudes** *est* "présence faible de couleurs chaudes") ***ET*** (**présence de couleurs froides** *est* "présence élevée de couleurs froides")} ***ALORS*** "les couleurs prédominantes sont froides".

◦ ***SI*** {(**présence de couleurs chaudes** *est* "présence élevée de couleurs chaudes") ***ET*** (**présence de couleurs froides** *est* "présence faible de couleurs froides")} ***ALORS*** "les couleurs prédominantes sont chaudes".

◦ ***SI*** {(**présence de couleurs chaudes** *est* "présence moyenne de couleurs chaudes") ***ET*** (**présence de couleurs froides** *est* "présence moyenne de couleurs froides")} ***ALORS*** "il y a un contraste chaud-froid".

◦ ***SI*** {(**couleurs adjacentes** *est* "oui") ***ET*** (**couleurs complémentaires** *est* "non")} ***ALORS*** "les couleurs prédominantes sont des couleurs adjacentes".

◦ ***SI*** {(**couleurs adjacentes** *est* "non") ***ET*** (**couleurs complémentaires** *est* "oui")} ***ALORS*** "les couleurs prédominantes sont des couleurs complémentaires".

◦ ***SI*** {(**couleurs adjacentes** *est* "oui") ***ET*** (**couleurs complémentaires** *est* "oui")} ***ALORS*** "il y a un contraste des couleurs adjacentes-complémentaires".

Les résultats expérimentaux obtenus pour différents films d'animation de la base de données de [CICA 06] sont présentés dans la Section 7.5.

7.3 La sémantique des plans vidéo

Dans cette partie nous allons présenter des concepts linguistiques décrivant de manière sémantique la structure temporelle d'une séquence d'images. De façon similaire à ce qui a été fait pour la caractérisation des couleurs, nous proposons une représentation linguistique par des ensembles flous des paramètres de bas niveau que nous avons extraits à partir des plans vidéo. Ces paramètres ont été proposés dans la Section 2.9. Nous n'utilisons pas ici d'étape intermédiaire définissant des attributs numériques de plus haut niveau comme cela a été fait pour la couleur.

Les informations symboliques envisagées sont d'abord le *rythme de la séquence* et *le contenu en terme d'action*, caractérisations qui sont génériques et qui restent valables quelque soit le type de film. Puis nous proposons deux caractérisations plus spécifiques aux films d'animation. Certains films d'animation ont un contenu *mystérieux* qui est mis en évidence par l'utilisation fréquente de transitions vidéo de type "dissolve" et "fade". Une autre caractéristique des films d'animation est la présence d'effets spécifiques sur les couleurs, comme par exemple les SCC ("changement bref de couleurs", voir la Section 2.5). L'utilisation fréquente de cet effet donne un caractère plutôt *explosif* à la séquence. Les descriptions symboliques de ces informations seront détaillées dans les paragraphes suivants.

7.3.1 La caractérisation sémantique floue des plans

Pour chaque paramètre numérique de bas niveau caractérisant les propriétés de la structure temporelle de la séquence nous allons associer un concept linguistique à travers une représentation par des ensembles flous.

Le rythme de la séquence

D'abord le concept linguistique *rythme de la séquence* est associé au paramètre $\bar{v}_T$, représentant la vitesse moyenne de changements de plans, calculée sur la séquence entière par tranches de T secondes (valeur fixée à $T = 5s$, voir la Section 2.9.1).

A un *niveau global*, le paramètre $\bar{v}_T$ est lié au rythme du déroulement temporel du contenu de la séquence. Une valeur élevée de $\bar{v}_T$ correspond à un nombre moyen élevé de changements

de plans par unité temporelle T. Lors de la visualisation d'une telle séquence, nous notons une cadence élevée de succession des informations visuelles. *A un niveau plus local*, le rythme de la séquence est lié à l'action. Les passages contenant un nombre élevé de changements de plans et donc beaucoup de changements visuels, sont typiquement des passages de la séquence contenant beaucoup d'action (concept souvent utilisé dans les méthodes d'extraction de résumés en mouvement, voir la Section 6.1.2).

Le concept linguistique *rythme de la séquence* est décrit par trois valeurs linguistiques : *"rythme lent"*, *"rythme moyen"* et *"rythme rapide"*. La signification floue de chaque symbole est illustrée par sa fonction d'appartenance floue. La partition floue de l'univers de discours, $\bar{v}_{T=5s}$, est déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : α_{lent_r}, α_{moyen_r} et α_{rapide_r} (voir la Figure 7.6).

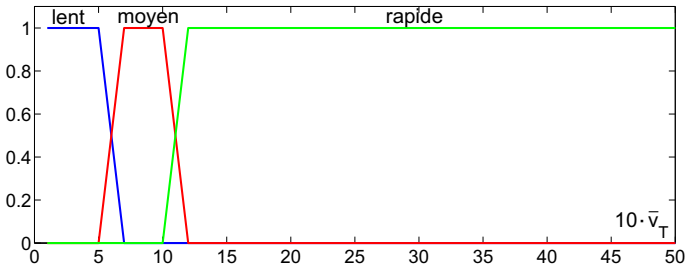


FIG. 7.6 – La partition floue de l'univers de discours $\bar{v}_T$ déterminée par les fonctions d'appartenance floues : α_{lent_r} (bleu), α_{moyen_r} (rouge) et α_{rapide_r} (vert) (l'axe oY correspond au degré d'appartenance).

Comme précédemment, les fonctions d'appartenance aux symboles ont été définies d'une manière classique en utilisant des fonctions linéaires par morceaux. Cette définition est basée sur le choix des 4 seuils $\{5, 7, 10, 12\}$, qui ont été déterminés empiriquement. En ce qui concerne la valeur maximale du paramètre $\bar{v}_{T=5s}$, théoriquement elle est située à $25 \times 5 = 125$, valeur inatteignable car cela se traduirait par un changement de plan à chaque image. L'avantage de la représentation floue dans ce cas est la *normalisation* des valeurs du paramètre $\bar{v}_{T=5s}$ entre 0 et 1 (le degré de vérité du concept linguistique associé).

Rythme lent. Pour définir le concept de *rythme lent* nous avons utilisé comme référence les films d'animation "A Crushed World" (6min42s, $\bar{v}_{T=5s} = 0.46$ qui possède une moyenne de 5.5 changements par minute), "Amerlock" (1min57s, $\bar{v}_{T=5s} = 0.04$ ayant une moyenne de 0.5 changements par minute), "David" (8min12s, $\bar{v}_{T=5s} = 0.49$ ayant une moyenne de 5.8 changements par minute) et "Greek Tragedy" (6min32s, $\bar{v}_{T=5s} = 0.44$ ayant une moyenne de 5.2 changements par minute). Quelques images représentatives de ces films sont présentées dans la Figure 7.7.

Ces films sont caractérisés par un rythme *lent*. Par exemple, dans le film "A Crushed World" l'action se déroule dans un monde restreint en utilisant un nombre limité de scènes (inférieur à 5). Ces caractéristiques sont liées à la technique d'animation utilisée (animation d'objets en papier). Le contenu du film est résumé par le synopsis : *les expériences d'un personnage de papier froissé, roulé, déformé, jeté, ballotté, se terminent par une rencontre douce et légère* [CICA 06]. De façon similaire, le film "Amerlock" est : *un jeu avec quelques grands mythes et personnages mystifiés des États-Unis d'Amérique* [CICA 06]. Dans ce cas l'action se déroule sur une seule scène, durant laquelle les différentes figurines en pâte à

FIG. 7.7 – Exemples de films d'animation ayant un rythme *lent* (de gauche à droite) : "A Crushed World", "Amerlock", "David" et "Greek Tragedy".

modeler se transforment en différents personnages. Le film "Greek Tragedy" est décrit par : *le violent réveil des cariatides d'un temple ancien qui doivent faire face aux archéologues et aux touristes* [CICA 06]. L'action se déroule au même endroit pendant tout le film, car les trois cariatides doivent soutenir la structure d'un temple. En fonction de ces exemples, nous avons considéré qu'un film a un *"rythme lent"* (vérité 1) si le nombre moyen de changements de plan est inférieur à 6 changements par minute ou à 0.5 changements toutes les 5 secondes, soit $10 \cdot \bar{v}_{T=5s} < 5$.

Rythme moyen. Pour le concept de *rythme moyen* nous avons utilisé comme référence les films suivants : "Gazoon" (2min47s, $\bar{v}_{T=5s} = 0.88$, ayant une moyenne de 10.6 changements de plan par minute), "L'Homme aux bras ballants" (3min38s, $\bar{v}_{T=5s} = 0.86$, ayant une moyenne de 10.3 changements de plan par minute), "The Sand Castle" (12min12s, $\bar{v}_{T=5s} = 1$, ayant une moyenne de 12 changements de plan par minute) et "Le Trop Petit Prince" (6min26s, $\bar{v}_{T=5s} = 0.89$, ayant une moyenne de 10.7 changements de plan par minute). Les films sont présentés dans la Figure 7.8.

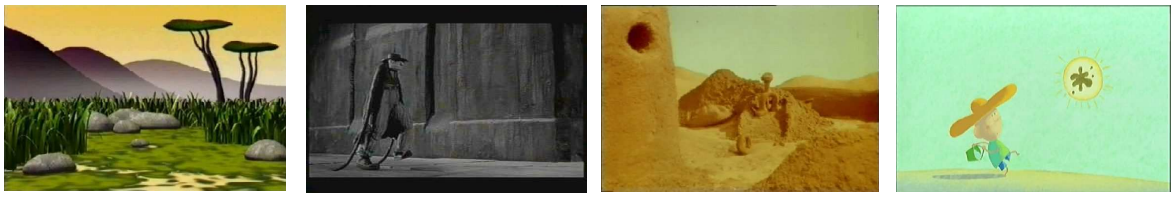


FIG. 7.8 – Exemples de films d'animation ayant un rythme *moyen* (de gauche à droite) : "Gazoon", "L'Homme aux Bras Ballants", "The Sand Castle" et "Le Trop Petit Prince".

Les films de cette catégorie sont typiquement des films pour lesquels l'action suit un cheminement classique : intrigue, le déroulement de l'action, le point culminant et le dénouement, le tout ayant un rythme plutôt constant. Par exemple le contenu du film "L'Homme aux bras ballants" est décrit de la façon suivante : *par une nuit sans lune, dans une ville endormie, un personnage aux bras démesurés marche. Précédé par son ombre, il se rend dans une arène pour accomplir un rituel.* [CICA 06]. Un autre exemple est de film "Le Trop Petit Prince" : *toute la journée, un tout petit bonhomme essaie de nettoyer le soleil qui est sale* [CICA 06], description qui aboutit à un rythme constant et qui se traduit par une action répétitive. A travers ces exemples, nous avons décidé qu'une séquence a un *"rythme moyen"* (vérité 1) si le nombre moyen de changements de plans est supérieur à 8 changements par minute mais aussi inférieur à 12 changements par minute, soit $10 \cdot \bar{v}_{T=5s} > 7$ et $10 \cdot \bar{v}_{T=5s} < 10$.

Rythme rapide. Pour définir le concept de rythme rapide nous avons utilisé comme référence les films "Ferrailles" (6min15s, $\bar{v}_{T=5s} = 1.92$, ayant une moyenne de 23 changements de plan par minute), "Le Moine et le Poisson" (6min, $\bar{v}_{T=5s} = 2.37$, ayant une moyenne de 28.4 changements de plan par minute), "François le Vaillant" (8min56s, $\bar{v}_{T=5s} = 1.57$, ayant une moyenne de 18.8 changements de plan par minute) et "Le Chat d'Appartement" (6min42s, $\bar{v}_{T=5s} = 1.32$, ayant une moyenne de 15.8 changements de plan par minute). Les films sont présentés dans la Figure 7.9.

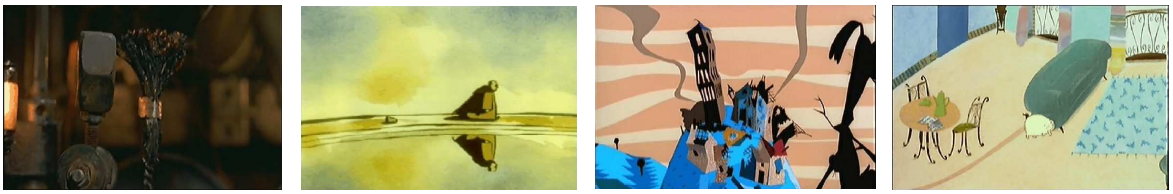


FIG. 7.9 – Exemples de films d'animation ayant un rythme *rapide* (de gauche à droite) : "Ferrailles", "Le Moine et le Poisson", "François le Vaillant" et "Le Chat d'Appartement".

Les films de cette catégorie comportent un rythme plutôt rapide. Un nombre élevé d'événements importants se succèdent durant tout le film. Les changements visuels sont fréquents. Par exemple, le rythme alerte du film "Ferrailles" se retrouve dans le synopsis du film : *ambiance d'atelier, dans une usine métallurgique. Les Chaînes grincent, les engrenages tournent, les moteurs ronronnent, [...] ,On graisse, on nettoie, on vérifie une dernière fois, puis on relance la machine qui semble bien repartie...* [Folimage 06a]. Un autre exemple est le film "Le Moine et le Poisson" où le sujet est la poursuite répétitive et de plus en plus dynamique d'un poisson : *un moine découvre un poisson dans un réservoir d'eau près d'un monastère. Il essaie de l'attraper en utilisant toutes sortes de moyens. Au cours du film, la poursuite devient de plus en plus symbolique.* [CICA 06]. En se basant sur ces exemples, nous avons considéré que la séquence a un *"rythme rapide"* (vérité 1) si le nombre moyen de changements de plans est supérieure à 14 changements par minute, soit $10 \cdot \bar{v}_{T=5s} > 12$.

Le contenu en terme d'action

La description linguistique proposée dans cette section est liée au contenu en terme d'action de la séquence. Le concept linguistique *contenu en terme d'action* est associé au paramètre R_{action}, qui représente la quantité de passages d'action de la séquence (voir la Section 2.9.1).

Une valeur élevée de R_{action} correspond à un contenu riche en action de la séquence. Les valeurs symboliques décrivant le concept, *"action faible"*, *"action moyenne"* et *"action élevée"*, seront décrites par les fonctions d'appartenance floue α_{faible_a}, α_{moyen_a} et $\alpha_{élevé_a}$. La partition floue de l'univers de discours, R_{action}, est présentée par la Figure 7.10.

Les valeurs des seuils utilisés pour la définition de ces trois fonctions, $\{30, 36, 63, 69\}$, ont été déterminées empiriquement par l'analyse manuelle de plusieurs films d'animation de la base de données de [CICA 06]. L'univers de discours a été divisé en trois intervalles approximativement égaux car nous avons pu constaté que le contenu en terme d'action est proportionnel à la valeur du paramètre R_{action}.

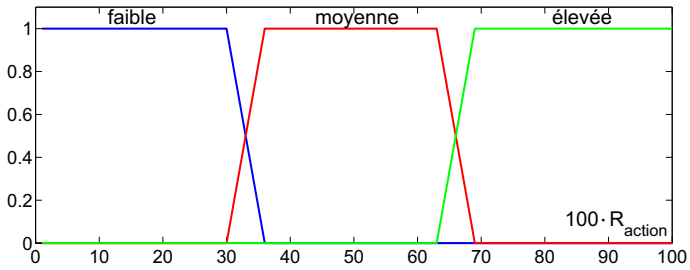
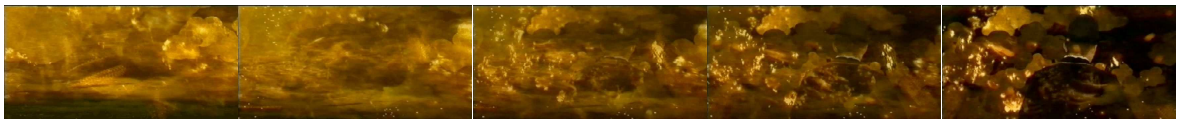


FIG. 7.10 – La partition floue de l'univers de discours R_{action} déterminée par les fonctions d'appartenance floue : α_{faible_a} (bleu), $\alpha_{moyenne_a}$ (rouge) et $\alpha_{élevée_a}$ (vert) (l'axe oY correspond au degré d'appartenance).

Le mystère se dégageant d'une séquence

Le concept linguistique *contenu mystérieux* est associé au paramètre R_{trans} représentant la quantité de transitions vidéo de type "dissolve" et "fade" présentes dans la séquence (voir la Section 2.9.2). Une valeur élevée de R_{trans} correspond à une utilisation fréquente de ce type de transitions et donc à une durée totale significative par rapport à la durée de la séquence entière.



(a) "dissolve" (changement de temps de l'action, film "A Viagem")

(c) "dissolve", (changement de l'endroit de l'action, film "Paradise")

(b) "fade in", (changement progressif de la scène, film "Coeur de Secours")

FIG. 7.11 – Exemple d'utilisation des transitions vidéo progressives dans les films d'animation de la base de données de [CICA 06].

Comme nous l'avons déjà mentionné dans la Section 2.9.2, les transitions graduelles sont utilisées dans les films avec un but précis. Elles vont se substituer aux transitions abruptes ("cuts", transitions les plus souvent utilisées) pour mettre en relief certains moments de la séquence. Par exemple, les "dissolves" sont typiquement utilisés pour changer le temps de l'action. D'autre part, les "fades" sont utilisés pour faire une transition progressive entre deux scènes différentes ou pour augmenter le suspens (voir la Figure 7.11). Dans les films d'animation l'utilisation fréquente de ce type de transitions est liée au contenu mystérieux (inexplicable, voire énigmatique) de la séquence. Dans certains films d'animation, presque toutes les transitions vidéo sont des transitions progressives de type "dissolve" ou "fade".

Ce sont par exemple les films "Paradise" et "A Viagem" [CICA 06], qui seront caractérisés comme ayant un contenu énigmatique ou mystérieux.

Le concept *contenu mystérieux* sera décrit en utilisant trois valeurs linguistiques : *"contenu mystérieux réduit"*, *"contenu mystérieux moyen"* et *"contenu mystérieux élevé"*. La signification floue de chaque symbole est illustrée par sa fonction d'appartenance floue. La partition floue de l'univers de discours, R_{trans}, est alors déterminée par l'ensemble des fonctions d'appartenance aux trois symboles : $\alpha_{réduit_m}$, α_{moyen_m} et $\alpha_{élevé_m}$ (voir la Figure 7.12). Les valeurs des seuils utilisés pour la définition de ces trois fonctions, $\{6, 8, 30, 32\}$, ont été déterminées empiriquement après avoir analysé plusieurs films d'animation de la base de données de [CICA 06].

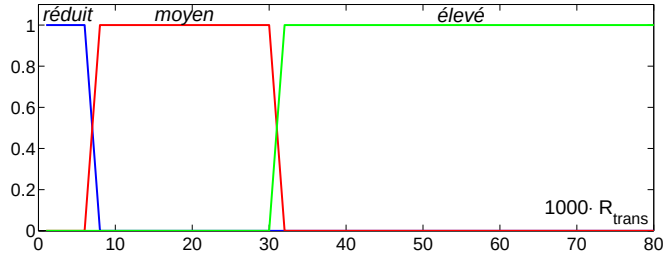


FIG. 7.12 – La partition floue de l'univers de discours R_{trans} déterminée par les fonctions d'appartenance floues : $\alpha_{réduit_m}$ (bleu), α_{moyen_m} (rouge) et $\alpha_{élevé_m}$ (vert) (l'axe oY correspond au degré d'appartenance).

Les valeurs faibles du paramètre $100 \cdot R_{trans}$ (inférieures à 4%), utilisées pour définir les fonctions d'appartenance floue, sont justifiées par le fait que la durée d'une transition vidéo progressive est d'environ quelques secondes. Le nombre de transitions de la séquence est bien sûr lié au nombre de plans vidéo, une transition faisant le lien entre deux plans différents. Ainsi, une valeur de $100 \cdot R_{trans} = 4\%$ est une valeur importante correspondant à une durée totale des transitions progressives très élevée par rapport à la durée de la séquence.

Contenu mystérieux élevé. Pour définir le concept de *contenu mystérieux élevé* nous avons utilisé comme référence les films suivants : "A Viagem" (7min32s, $100 \cdot R_{trans} = 4.2$, contenant 8 "fades" et 8 "dissolves"), "Cœur de Secours" (9min13s, $100 \cdot R_{trans} = 7.24$, contenant 26 "fades" et 63 "dissolves"), "Le Moine et le Poisson" (6min, $100 \cdot R_{trans} = 4.62$, contenant 10 "fades" et 61 "dissolves") et "Le Pas" (8min57s, $100 \cdot R_{trans} = 3.48$, contenant 7 "fades" et 117 "dissolves"). Les films sont présentés dans la Figure 7.13.

FIG. 7.13 – Exemples de films d'animation ayant un contenu mystérieux/énigmatique élevé (de gauche à droite) : "A Viagem", "Cœur de Secours", "Le Moine et le Poisson" et "Le Pas".

Les films de cette catégorie sont des films ayant un contenu énigmatique prédominant. Par exemple, le film "A Viagem" est caractérisé par : *vision déformée, amusée, parfois inquiète, d'un monde trépidant, qui navigue entre histoire et légende, entre farce et drame, entre peinture et réalité* [CICA 06]. Le film "Cœur de Secours" est un film philosophique ayant un contenu qui n'est pas facilement compréhensible : *l'amour désespéré d'un clown lunaire pour une demoiselle à l'intérieur d'une horloge arrêtée* [Production 06]. De façon similaire, le contenu du film "Le Moine et le Poisson" est plutôt symbolique : *un moine découvre un poisson dans un réservoir d'eau près d'un monastère [...] Il essaie de l'attraper [...] la poursuite devient de plus en plus symbolique.* [CICA 06]. Nous avons ainsi décidé qu'une séquence a un *"contenu mystérieux élevé"* (vérité 1) si les transitions graduelles représentent plus de 3.2% de la durée totale du film, soit $100 \cdot R_{trans} > 3.2$.

Contenu mystérieux moyen. Le concept *contenu mystérieux moyen* a été défini en utilisant comme référence les films : "Greek Tragedy" (6min32s, $100 \cdot R_{trans} = 2$, contenant 4 "fades" et 2 "dissolves"), "Le Rêve du Diable" (10min9s, $100 \cdot R_{trans} = 2.61$, contenant 103 "dissolves"), "Paradise" (14min3s, $100 \cdot R_{trans} = 2$, contenant 14 "fades" et 53 "dissolves") et "Repete" (7min52s, $100 \cdot R_{trans} = 2.9$, contenant 2 "fades" et 20 "dissolves"). Les films sont présentés dans la Figure 7.14.

FIG. 7.14 – Exemples de films d'animation ayant un contenu mystérieux/énigmatique moyen (de gauche à droite) : "Greek Tragedy", "Le Rêve du Diable", "Paradise" et "Repete".

Les films de cette catégorie ont un contenu différent par rapport au contenu des films classiques. Ce sont typiquement des films avec un contenu qui combine des passages plutôt difficiles à comprendre avec des passages facilement accessibles. Par exemple, le contenu du film "Paradise" est décrit par : *un merle ordinaire, très malheureux dans son paradis tropical, est fasciné par l'oiseau divin et sa vie luxueuse. Il veut lui ressembler et aspire à l'éclat et au confort du palais de cristal.* [CICA 06]. Le film "Repete" : *on a tous des réflexes intérieurs qui guident nos vies. La routine quotidienne est ennuyeuse mais, en même temps, on s'y sent en sécurité. On rêve de changement, mais on n'a pas le courage de faire le pas décisif.* [CICA 06]. A travers ses exemples typiques, nous avons décidé qu'une séquence a un *"contenu mystérieux moyen"* (vérité 1) si $100 \cdot R_{trans} > 0.8$ et $100 \cdot R_{trans} < 3$.

Contenu mystérieux réduit. Pour définir le concept de *contenu mystérieux réduit* nous avons utilisé comme référence les films suivants ayant un contenu classique et facilement compréhensible, où l'action est bien définie : "At the Ends of the Earth" (7min28s, $100 \cdot R_{trans} = 0.3$, contenant 2 "fades"), "Fini Zayo" (7min2s, $100 \cdot R_{trans} = 0.29$, contenant 3 "fades"), "François le Vaillant" (8min56s, $100 \cdot R_{trans} = 0.69$, contenant 6 "fades") et "La Bouche Cousue" (2min48s, $100 \cdot R_{trans} = 0.24$, contenant un seul "fade") (voir la Figure 7.15).

Par exemple, l'action du film "At the Ends of the Earth" est résumée par : *posée sur le*

FIG. 7.15 – Exemples de films d'animation ayant un contenu mystérieux/énigmatique réduit (de gauche à droite) : "At the Ends of the Earth", "Fini Zayo", "François le Vaillant" et "La Bouche Cousue".

pic d'une colline, une maison balance alternativement de droite à gauche, au grand dam de ses habitants [CICA 06]. Un autre exemple est le film "François le Vaillant" : *une armée médiévale emmenée par un chef cruel et sanguinaire fait régner la terreur. François le Vaillant, la fleur à la lance, traverse, avec un certain détachement, le théâtre des ravages de la guerre.* [Folimage 06b]. Aussi, nous avons décidé qu'une séquence a un *"contenu mystérieux réduit"* (vérité 1) si $100 \cdot R_{trans} < 0.6$.

Le contenu explosif

Le concept linguistique *contenu explosif* est associé au paramètre R_{SCC} représentant la quantité de "changements bref de couleurs" (SCC) présents dans la séquence (voir la Section 2.9.2). Un SCC est un effet spécifique aux films d'animation (voir la Section 2.5) qui correspond à des changements de couleurs de courte durée dans la séquence. Dans les films d'animation il est typiquement associé aux éclairs, aux explosions, aux flashs, etc. La présence fréquente de SCC donne à la séquence un caractère particulier, que nous nommerons "*explosif*".

Le concept *contenu explosif* sera décrit en utilisant seulement deux valeurs linguistiques : *"oui"* ou *"non"*. La signification floue de chaque symbole est illustrée par sa fonction d'appartenance floue. La partition floue de l'univers de discours, R_{SCC}, est déterminée par l'ensemble des fonctions d'appartenance aux deux symboles : $\alpha_{non_{ex}}$ et $\alpha_{oui_{ex}}$ (voir la Figure 7.16).

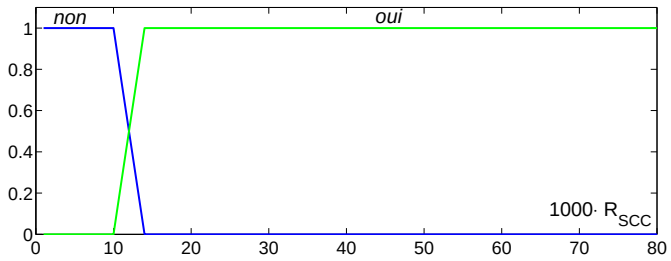


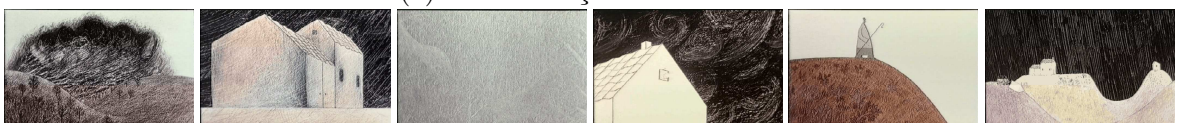
FIG. 7.16 – La partition floue de l'univers de discours R_{SCC} déterminée par les fonctions d'appartenance floue : $\alpha_{non_{ex}}$ (bleu) et $\alpha_{oui_{ex}}$ (vert) (l'axe oY correspond au degré d'appartenance).

Les valeurs des seuils utilisés pour la définition de ces deux fonctions, $\{10, 14\}$, ont été déterminées empiriquement après une analyse manuelle de plusieurs films d'animation de la

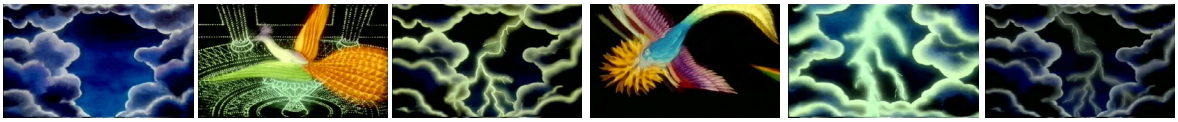
base de données de [CICA 06]. L'univers de discours a été divisé en deux intervalles car les films d'animation sont ou ne sont pas explosifs.

Pour définir le concept d'un *contenu explosif* nous avons utilisé comme référence les films d'animation suivants : "François le Vaillant" (8min56s, $100 \cdot R_{SCC} = 1.78$, contenant 39 SCC), "The Hill Farm" (16min39s, $100 \cdot R_{SCC} = 0.86$, contenant 45 SCC) et "Paradise" (14min3s, $100 \cdot R_{SCC} = 0.37$, contenant 7 SCC) (voir la Figure 7.17). Nous avons caractérisé la séquence comme explosive (valeur *"oui"*) avec un degré d'appartenance de 1, si la durée totale de tous les effets SCC de la séquence est supérieure à 1.4% de la durée totale de la séquence, dont $100 \cdot R_{SCC} > 1.4$, et non explosif (vérité 1) si $100 \cdot R_{SCC} < 1$.

(a) film "François le Vaillant"



(c) film "The Hill Farm"



(b) film "Paradise"

FIG. 7.17 – Exemples d'utilisation des effets couleur SCC dans les films d'animation de données de [CICA 06].

En conclusion, en utilisant les paramètres de bas niveau extraits des plans proposés dans la Section 2.9 nous avons pu définir les concepts linguistiques suivants :

- **rythme de la séquence**,
- **contenu en terme d'action**,
- **contenu mystérieux**,
- **contenu explosif**.

Des résultats expérimentaux obtenus pour différents films d'animation de la base de données de [CICA 06] seront présentés dans la Section 7.5.

7.4 La sémantique du mouvement

En ce qui concerne le mouvement, des travaux sont en cours pour définir une caractérisation symbolique/sémantique à partir des paramètres de bas niveau définis dans la Section 3.3. Dans ce paragraphe nous allons présenter les principes de ce que nous envisageons de développer, le mécanisme étant similaire à celui que nous avons utilisé pour la caractérisation des couleurs et des plans. Les concepts proposés sont :

- **contenu calme** : ce concept est lié au paramètre R_{pm} qui représente la durée totale

de tous les passages sans mouvement (de caméra ou de personnages) présents dans la séquence. Une valeur élevée de R_{pm} correspond à un contenu calme de la séquence, avec peu d'action, donc plutôt statique.

- **activité de la séquence** : ce concept est lié au paramètre R_{mo} qui représente la durée totale de tous les passages de la séquence comportant un mouvement d'objets. Une valeur élevée de R_{mo} correspond à un contenu plutôt actif à l'intérieur des scènes de la séquence. Dans ce cas le mouvement prédominant de la séquence est surtout lié au mouvement des personnages ou des objets.

- **changement de point de vue** : ce concept est lié au paramètre $R_{m.trans}$ qui représente le nombre de mouvements de translation de caméra présents dans la séquence. Une valeur élevée de $R_{m.trans}$ reflète la prédominance de tels mouvements. Typiquement, un mouvement de translation est utilisé pour changer le point de vue de la scène. L'utilisation fréquente de ce type de mouvement reflète donc de nombreux changements de scène.

- **présence de détails** : ce concept est lié au paramètre R_{zoom} qui représente le nombre de mouvements de type "zoom in/out" de la séquence. Dans les films, de tels mouvements sont utilisés soit pour se focaliser sur un certain point d'intérêt de la scène ("zoom in"), soit pour donner une vue d'ensemble sur la scène ("zoom out"). L'utilisation fréquente de ce type de mouvement est donc liée aux détails des informations présentées dans la séquence.

- **effets de caméra** : ce concept est lié au paramètre R_{rot} qui représente le nombre de mouvements de rotation présents dans la séquence. Dans les films, le mouvement de rotation n'est utilisé que dans des situations particulières. On peut le voir comme un effet visuel pour mettre en relief certains événements importants de la séquence, comme par exemple : la chute d'un personnage, une poursuite de voitures, etc. La présence fréquente de ce type de mouvement de caméra implique un film riche en événements ayant beaucoup d'action.

7.5 Résultats expérimentaux

Pour valider les descriptions sémantiques proposées dans les sections précédentes (caractérisation des couleurs et des plans), nous avons utilisé un extrait de la base numérique des films d'animation du CICA [CICA 06]. L'extrait contient 52 courts métrages présentant une vaste diversité de techniques d'animation et d'une durée totale de 6 heures et 6 minutes.

Les films utilisés sont présentés dans l'Annexe F. Les résultats obtenus pour un certain nombre de films représentatifs parmi les films analysés ("Amerlock", "Casa", "Circuit Marine", "Le Moine et le Poisson", "Och, och", "Tamer of Wild Horses", "La Cancion du Microsillon", "Le Chateau des Autres" et "François le Vaillant") sont présentés et commentés dans l'Annexe G.

Nous présentons ici la caractérisation sémantique du film d'animation "Le Trop Petit Prince". Les informations présentées sont les suivantes : quelques *images représentatives*

pour se faire une impression globale du contenu de la séquence[2], *le synopsis* du film qui est un court résumé textuel du film (typiquement créé par l'auteur) fournissant des informations sur le contenu, *l'histogramme des couleurs élémentaires*, $h_{élém}()$, représentant la distribution des couleurs élémentaires de la séquence (voir la Section 7.2.1) [3], *l'histogramme couleur global pondéré*, $h_{seq}()$, (calculé pour $p = 15\%$, voir la Section 4.2), qui représente la distribution des couleurs de la séquence, *l'annotation visuelle* des transitions vidéo (voir la Section 2.8) qui nous donne des informations sur la distribution des transitions vidéo dans la séquence et sur la densité des changements de plans et enfin les *descriptions sémantiques/symboliques* des couleurs et des plans vidéo obtenues par la formalisation floue des paramètres de bas niveau de la séquence.

Pour le film "Le Trop Petit Prince" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure 7.18) :

- **synopsis :** *"Toute la journée, un tout petit bonhomme essaie de nettoyer le soleil qui est sale."* [CICA 06].
- **couleurs élémentaires :** "Yellow" 24,74%, "Green" 21,93%, "Orange" 13,79%, "Gray" 13,70%, "Cyan" 7,01%, "Black" 6,99%, "Azure" 6,22%, "Blue" 4,74%, "Teal" 4,25%, "Spring" 3,59%, "Red" 2,53%, "Magenta" 0,31%, "White" 0,15%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 0.89$	*"rythme moyen"* (**1**)
$100 \cdot R_{action} = 84.11\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 1\%$	*"contenu mystérieux moyen"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 62.03\%$	*"présence moyenne de couleurs claires"* (**0.65**)
$100 \cdot P_{foncées} = 37.63\%$	*"présence faible de couleurs foncées"* (**0.73**)
$100 \cdot P_{fortes} = 3.17\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 66.28\%$	*"présence élevée de couleurs faiblement saturées"* (**1**)
$100 \cdot P_{chaudes} = 38.72\%$	*"présence faible de couleurs chaudes"* (**0.66**)
$100 \cdot P_{froides} = 40.42\%$	*"présence faible de couleurs froides"* (**0.56**)
$100 \cdot P_{var} = 70.83\%$	*"variété des couleurs élevée"* (**1**)
$100 \cdot P_{div} = 61.54\%$	*"diversité des couleurs moyenne"* (**0.74**)
$100 \cdot P_{adj} = 90\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 80\%$	*"couleurs complémentaires : oui"* (**1**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0.35**) *"les couleurs prédominantes sont foncées"* (**0**) *"il y a un contraste claire-foncé"* (**0.27**)

[2] dans la plupart des situations le contenu du film ne peut pas être résumé en quelques images ; pour un aperçu plus exact du contenu se reporter à la séquence originale.

[3] $h_{élém}()$ est présenté sous la forme d'un "camembert" ; les couleurs utilisées dans la représentation ne suivent pas l'intensité et la saturation des couleurs élémentaires réelles utilisées dans la séquence, ce sont des couleurs de saturation maximale qui sont présentées à titre indicatif.

Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible* (**1**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0.34**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0**) *"les couleurs prédominantes sont des couleurs complémentaires* (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires* (**1**)

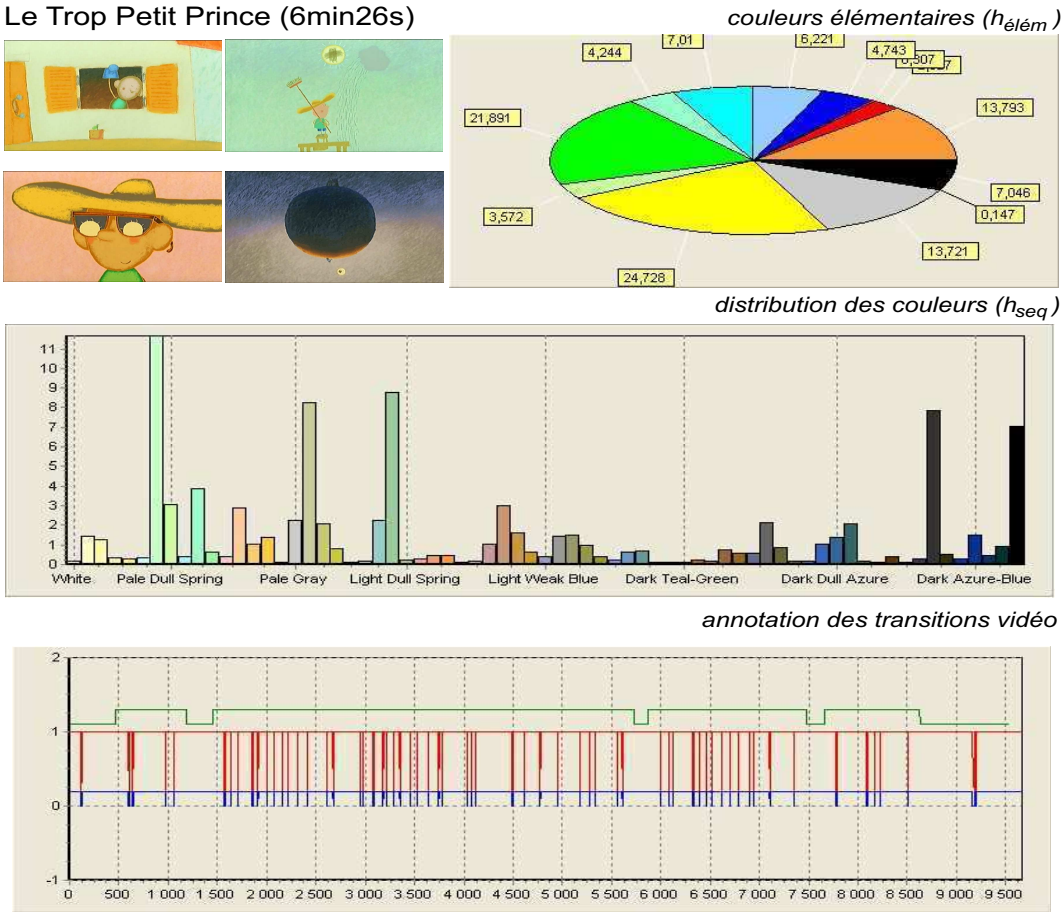


FIG. 7.18 – Film "Le Trop Petit Prince" : histogramme des couleurs élémentaires, histogramme global pondéré et l'annotation visuelle des transitions.

Le film "Le Trop Petit Prince" a un *rythme moyen* constant pendant la durée du film, paramètre lié à la répétivité des actions du personnage principal : "toute la journée, un tout petit bonhomme essaie de nettoyer le soleil". Comme le montre la Figure 7.18, l'action est

répartie tout au long de la séquence ce qui dénote un *contenu en terme d'action élevé*. De ce fait il est difficile de prévoir l'action future et la fin du film ce qui lui donne un *caractère mystérieux moyen*.

En ce qui concerne la distribution des couleurs, les couleurs élémentaires prédominantes sont le *"Yellow"*, *"Green"*, *"Orange"*, *"Gray"* et *"Cyan"*. La séquence est caractérisée par une proportion de couleurs claires de 62%. Les couleurs prédominantes du film ont une faible saturation et se divisent à égale proportion entre les couleurs chaudes et froides, qui sont également à la fois adjacentes et complémentaires. Notons que la variété des couleurs est élevée puisque 153 couleurs différentes sont utilisées sur un total de 216. D'autre part le film utilise un nombre moyen de couleurs élémentaires ce qui fait que la diversité des couleurs est moyenne.

L'évaluation des résultats est une tâche subjective. Comme nous n'avons pas réellement de vérité terrain, pour valider les résultats nous avons utilisé quatre types d'informations :

- **les synopsis** des films, qui sont de courts résumés textuels du contenu fournis par les auteurs,
- **les fiches techniques** (voir Animaquid [CICA 06]),
- **d'autres informations** obtenues sur les sites des différents auteurs,
- **l'expertise** de quelques utilisateurs, en ce qui concerne plus spécifiquement la couleur.

A travers l'exemple analysé ci-dessus et ceux donnés en Annexe G, nous avons obtenu une bonne concordance entre les caractérisations proposées et les informations disponibles. Néanmoins, cette bonne concordance est une impression subjective demanderait à être quantifiée de manière objective.

7.6 Représentation et comparaison des films d'animation : le gamut sémantique

Dans cette section nous proposons une *méthodologie de représentation visuelle* des caractéristiques des films. Cette représentation peut, par exemple, être utilisée pour comparer les contenus de différents films et ensuite permettre de trouver d'une manière efficace des caractéristiques communes à plusieurs films, tâche nécessaire dans un moteur de recherche d'une base de données vidéo. Les caractérisations acquises de chaque film sont représentées en utilisant une représentation graphique inspirée de la construction des gamuts de couleurs d'un dispositif de restitution d'images couleurs (écran ou imprimante). Dans la suite de ce mémoire, nous appellerons cette représentation *le gamut sémantique* de la séquence.

7.6.1 La construction des gamuts

La méthode de construction d'un gamut sémantique caractérisant certaines propriétés sémantiques d'un film est la suivante : les valeurs de tous les paramètres d'une catégorie, caractérisant le contenu du film sont représentées sur des axes différents dans l'espace 2D XoY. Le point de référence (à savoir l'origine) de la représentation est le centre du graphique. Le gamut sémantique est déterminé par la surface formée en joignant les différentes valeurs des paramètres représentés sur les différents axes. Un exemple est présenté dans la Figure 7.19.

Nous avons divisé les caractérisations des séquences en trois catégories. Pour chacune

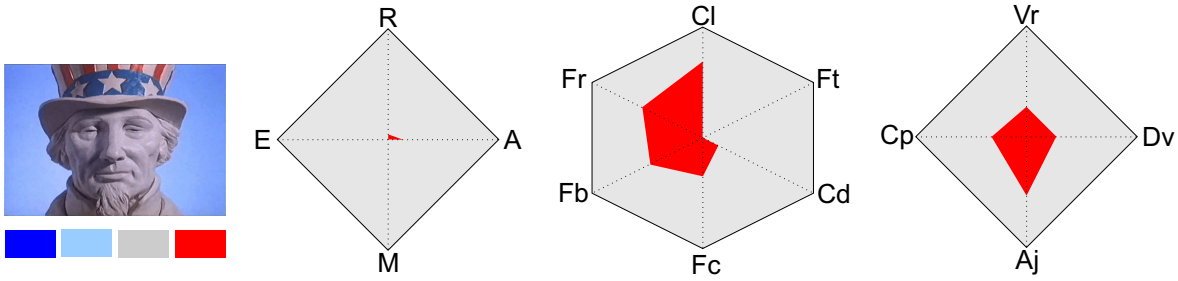


FIG. 7.19 – Les gamuts sémantiques obtenus pour le film "Amerlock" : G^p, G^c et G^{rl} (les couleurs élémentaires prédominantes sont illustrées en dessous de l'image).

nous avons associé un gamut sémantique : *le gamut des plans* (G^p), *le gamut des propriétés couleurs* (G^c) et *le gamut de la richesse couleur et des relations entre couleurs* (G^{rl}). Les paramètres utilisés pour la construction des gamuts sont les suivants :

- **gamut des plans (G^p)** : **R**- rythme (paramètre $\bar{v}_{T=5s}$, valeur maximale $2.4 cuts/5s$), **A**- action (paramètre $100 \cdot R_{action}$, valeur maximale 100%), **M**- mystère (paramètre $100 \cdot R_{trans}$, valeur maximale 5%), **E**- explosivité (paramètre $100 \cdot R_{SCC}$, valeur maximale 2%),
- **gamut des propriétés couleurs (G^c)** : **Cl**- couleurs claires (paramètre $100 \cdot P_{claires}$, valeur maximale 100%), **Ft**- couleurs saturées (paramètre $100 \cdot P_{fortes}$, valeur maximale 100%), **Cd**- couleurs chaudes (paramètre $100 \cdot P_{chaudes}$, valeur maximale 100%), **Fc**- couleurs foncées (paramètre $100 \cdot P_{foncées}$, valeur maximale 100%), **Fb**- couleurs faible saturées (paramètre $100 \cdot P_{faibles}$, valeur maximale 100%), **Fr**- couleurs froides (paramètre $100 \cdot P_{froides}$, valeur maximale 100%),
- **gamut de la richesse couleur et des relations entre couleurs (G^{rl})** : **Vr**- variété des couleurs (paramètre $100 \cdot P_{var}$, valeur maximale 100%), **Dv**- diversité des couleurs (paramètre $100 \cdot P_{div}$, valeur maximale 100%), **Aj**- couleurs adjacentes (paramètre $100 \cdot P_{adj}$, valeur maximale 100%), **Cp**- couleurs complémentaires (paramètre $100 \cdot P_{compl}$, valeur maximale 100%).

Ce type de *représentation visuelle compacte* permet de se faire une *idée globale* sur l'ensemble des caractéristiques de la séquence. Ainsi, la tâche de comparaison des différents films s'en trouvera simplifiée car l'utilisateur n'a plus besoin de comparer indépendamment les valeurs des paramètres extraits. Il suffit de comparer visuellement *les formes* des gamuts sémantiques obtenus pour trouver les caractéristiques communes, partagés par les films analysés. Les films ayant des caractéristiques sémantiques différentes auront des formes de gamuts sémantiques différentes et inversement.

7.6.2 Résultats expérimentaux

Pour montrer l'efficacité de cette représentation nous l'avons testé sur plusieurs films d'animation. Les gamuts sémantiques obtenus pour les 10 films d'animation étudiés dans la section précédente (Section 7.5) sont illustrés dans l'Annexe H.

En comparant les résultats obtenus nous avons trouvé qu'il y a des similarités entre les films qui sont facilement repérables à partir des gamuts. Par exemple, les films "Casa" et

"Le Moine et le Poisson" utilisent des techniques de couleurs similaires et leurs gamuts ont des formes similaires. Les deux films utilisent en réalité la même technique d'animation qui est le dessin sur cellulose. De plus la distribution des couleurs est orientée vers une seule couleur prédominante contrastée par la présence d'un niveau de gris (Noir, Blanc ou Gris).

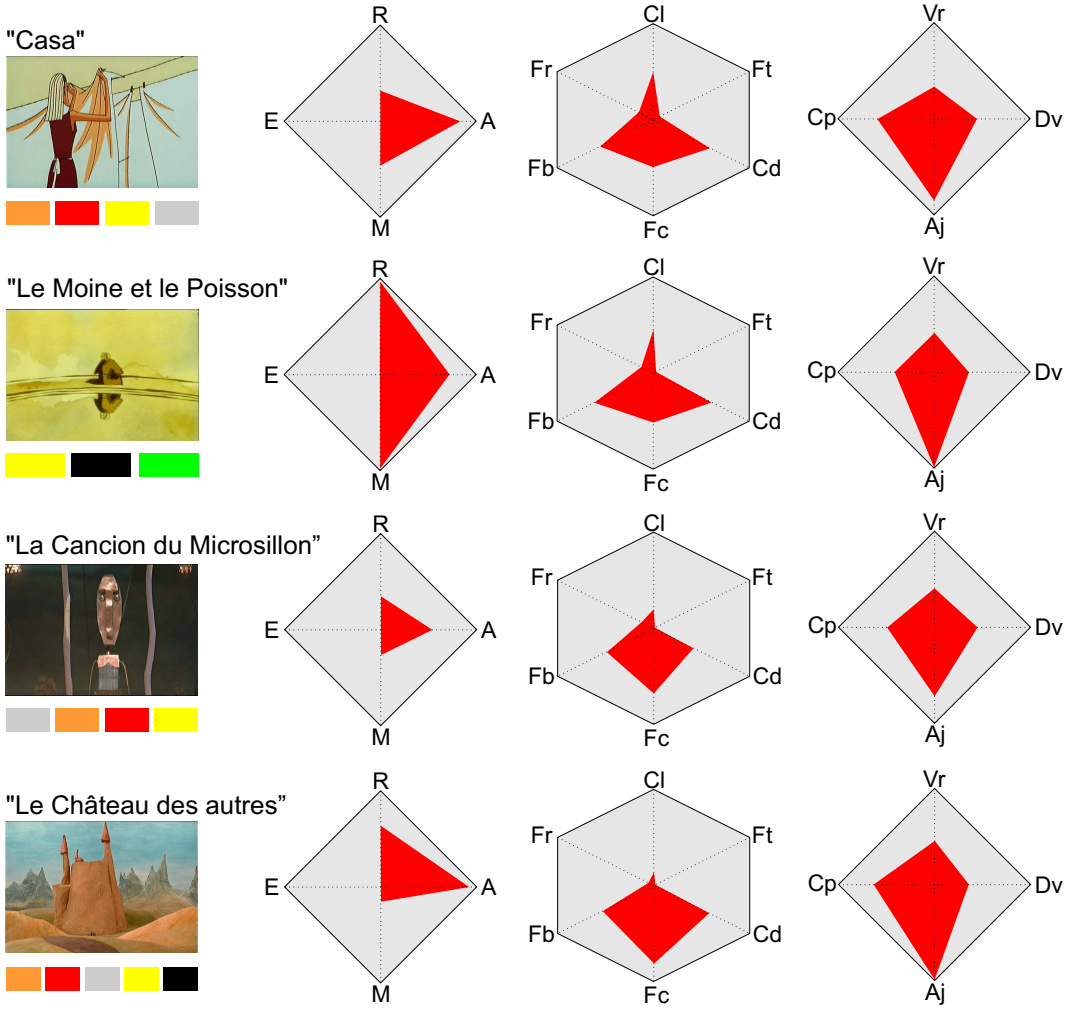


FIG. 7.20 – Les gamuts sémantiques G^p, G^c et G^{rl} (voir la Section 7.6) obtenus pour les films cités ci-dessus (les couleurs élémentaires prédominantes de chaque film sont illustrées en dessous de l'image).

Un autre exemple peut être donné avec les films "La Cancion du Microsillon" et "Le Château des autres", qui utilisent les mêmes couleurs élémentaires : "Orange", "Red", "Yellow" et "Gris" et des techniques couleurs similaires. Les gamuts des propriétés couleurs sont ressemblants (voir la Figure 7.20).

7.6.3 Les applications

Une application immédiate de ce type de représentation graphique peut être faite dans un *outil de navigation* d'une base de données vidéo. En associant à chaque film les gamuts proposés (comme ils sont présentés dans la Figure 7.19), cela permettra à l'utilisateur de se

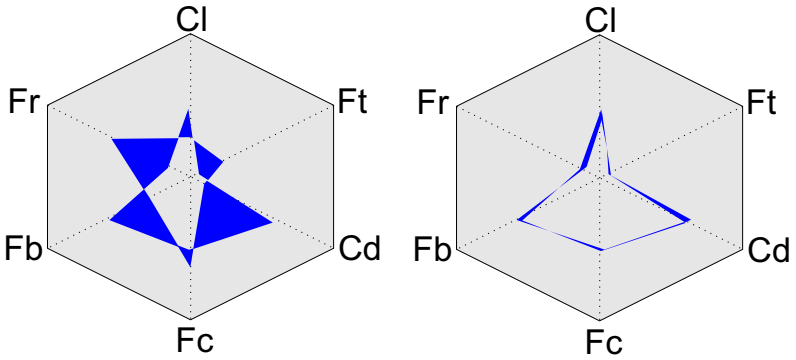
faire d'un seul coup d'œil une impression rapide et efficace sur le *contenu* de la séquence, sans passer du temps à regarder le film ou un résumé du film. Après une étape d'adaptation et d'apprentissage à ce nouvel outil bien sûr, on peut imaginer que les gamuts proposés puissent servir comme *résumés visuels des caractéristiques* de la séquence.

Une autre application possible est l'utilisation dans un *moteur de recherche* de base de données vidéo. On peut envisager de formuler les requêtes de recherche en utilisant les gamuts sémantiques. Ainsi, si l'utilisateur recherche un film d'action avec une certaine distribution de couleurs, dans un premier temps ses requêtes seront converties à l'aide d'un outil graphique en gamuts sémantiques. La recherche des films demandés sera ensuite effectuée en comparant les formes des gamuts obtenus (les requêtes) avec celles des gamuts des films de la base de données. Les films ayant des caractérisations sémantiques similaires comporteront des gamuts de formes similaires.

En utilisant ce type de représentation, comme nous l'avons déjà mentionné, on simplifie *la tâche de comparaison des caractéristiques de films*. Par exemple, la surface de la différence entre l'union et l'intersection des gamuts sémantiques peut être vue comme une mesure de distance entre films, mesure prenant en compte simultanément toutes les caractéristiques de la séquence :

$$d_{gamut}(G_1, G_2) = Surf(G_1 \cup G_2 - G_1 \cap G_2) \tag{7.8}$$

où G_1 et G_2 sont deux gamuts sémantiques, représentant la même catégorie sémantique, des deux films à comparer et $Surf()$ est l'opérateur retournant la surface. Plus les caractéristiques des deux films sont différentes, plus l'intersection des gamuts associés sera faible et la valeur de d_{gamut} élevée. Il faut noter que la mesure ainsi obtenue ne permet que des comparaisons relatives, et ne présente pas de caractère absolu. Elle reste cependant une distance au sens mathématique du terme.



(a) $d_{soust}(G^c_C, G^c_{FV})$ (b) $d_{soust}(G^c_C, G^c_{MP})$

FIG. 7.21 – Exemple de distance entre gamuts (les abréviations sont présentées au début de la section).

Pour montrer la capacité discriminatoire de la mesure de distance proposée nous présentons un exemple. L'exemple est illustré par la Figure 7.21 où G^c_C, G^c_{FV} et G^c_{MP} sont les *gamuts des propriétés couleurs* des films "Casa", "François le Vaillant" et "Le Moine et le Poisson" (voir les Figures H.1, H.2 et H.3). La distance $d_{gamut}(G^c_C, G^c_{MP})$ entre les caractéristiques couleurs des films "Casa" et "Le Moine et le Poisson" est très faible (voir la Figure 7.21.b), car les deux films sont similaires du point de vue des propriétés des couleurs. Par contre, la distance $d_{gamut}t(G^c_C, G^c_{FV})$ (voir la Figure 7.21.a) est importante car le film "Casa" et le film

"François le Vaillant" sont très différents en ce qui concerne les caractéristiques des couleurs.

7.7 Conclusions générales

Dans ce chapitre nous avons proposé une méthodologie de caractérisation sémantique/ symbolique du contenu de films à partir des paramètres de bas niveaux définis dans la première partie de la thèse. Les caractérisations sémantiques proposées portent sur la *structure temporelle* et la *distribution des couleurs* de la séquence. De plus, nous avons présenté nos perspectives sur la caractérisation du *mouvement*. La méthode utilisée est la *formalisation par des ensembles flous* des paramètres de bas niveau acquis sur la séquence. Les valeurs numériques sont converties en concepts linguistiques proches de la perception humaine, en utilisant des connaissances a priori sur les films d'animation. Des caractérisations sémantiques de plus haut niveau sont obtenues en utilisant des règles floues. La méthodologie proposée a été testée dans le domaine particulier des films d'animation sur une partie de la base de données numériques du CICA [CICA 06].

Dans ce genre d'approche, *l'évaluation des résultats* est une tâche difficile. Comme l'a été la validation des résumés, l'évaluation des caractérisations sémantiques proposées est subjective car elle est liée à la perception subjective de chaque individu. Pour valider les caractérisations obtenues nous avons opté pour une analyse manuelle du contenu des films testés. Ainsi, une vérité terrain a été construite à partir des différentes sources d'informations que nous avions à notre disposition, comme *le synopsis* des films, résumé textuel du contenu de la séquence fourni par les auteurs. Nous avons également utilisé les *fiches techniques* des films disponibles sur les sites du CICA (Animaquid) [CICA 06] et d'UniFrance [uniFrance 06]. Les résultats de la Section 7.5 montrent que la plupart des caractérisations obtenues sont en conformité avec la perception du contenu et avec les informations additionnelles disponibles (fiches techniques, synopsis, etc.) qui ont joué le rôle de vérité terrain.

Les caractérisations sémantiques/symboliques proposées sont liées au choix d'un certain vocabulaire de termes qui a été défini en mélangeant du bon sens et des connaissances à priori sur la construction des films, particulièrement dans le domaine des films d'animation. Les termes linguistiques utilisés sont en général assez naturels et ils sont inspirés de la perception que nous avons des couleurs ou du rythme. Certain des termes utilisés sont moins consensuels et sont encore le sujet de discussions (par exemple le terme "mystère").

En ce qui concerne l'utilisation des descriptions symboliques/sémantiques proposées, nous envisageons trois applications directes à notre système d'analyse de films :

- **une aide à la navigation** : nous avons proposé différentes méthodes d'extraction de résumés (en images et en mouvement, voir le Chapitre 6), pouvant servir à la navigation dans une base de films. Les descriptions symboliques/sémantiques proposées, associées aux résumés, apporteront plus d'informations sur le contenu de la séquence. Par exemple, le résumé d'un film peut être accompagné par des caractéristiques portant sur le rythme du film ou sur les couleurs dominantes.

- **une aide à la recherche** : les caractérisations symboliques proposées peuvent également servir comme index sémantiques lors de la recherche dans une base de films. Il sera plus efficace de chercher un film "triste" ou un film "d'action" en s'appuyant sur les symboles "couleurs froides" ou "rythme élevé" que d'exploiter des critères de bas niveau comme par exemple le type du mouvement de caméra ou le pourcentage de

transitions vidéo.

– **une aide pour les spécialistes** : les méthodes proposées, au travers des outils logiciels développés, peuvent être un outil d'analyse du contenu des films pour les spécialistes du domaine des films d'animation. Les méthodes d'analyse envisagées font en quelque sorte l'opération inverse du processus de montage de la séquence. Notre démarche peut s'assimiler à un processus de "reverse engineering". De plus, les méthodologies employées étant totalement originales dans le monde de l'animation, cela permet aux spécialistes de porter un autre regard sur les films d'animation. A titre d'exemple, nous proposons ci-dessous (Figure 7.22) une fiche compacte, qui pourrait être générée automatiquement, résumant l'ensemble des principaux éléments descriptifs d'un film.

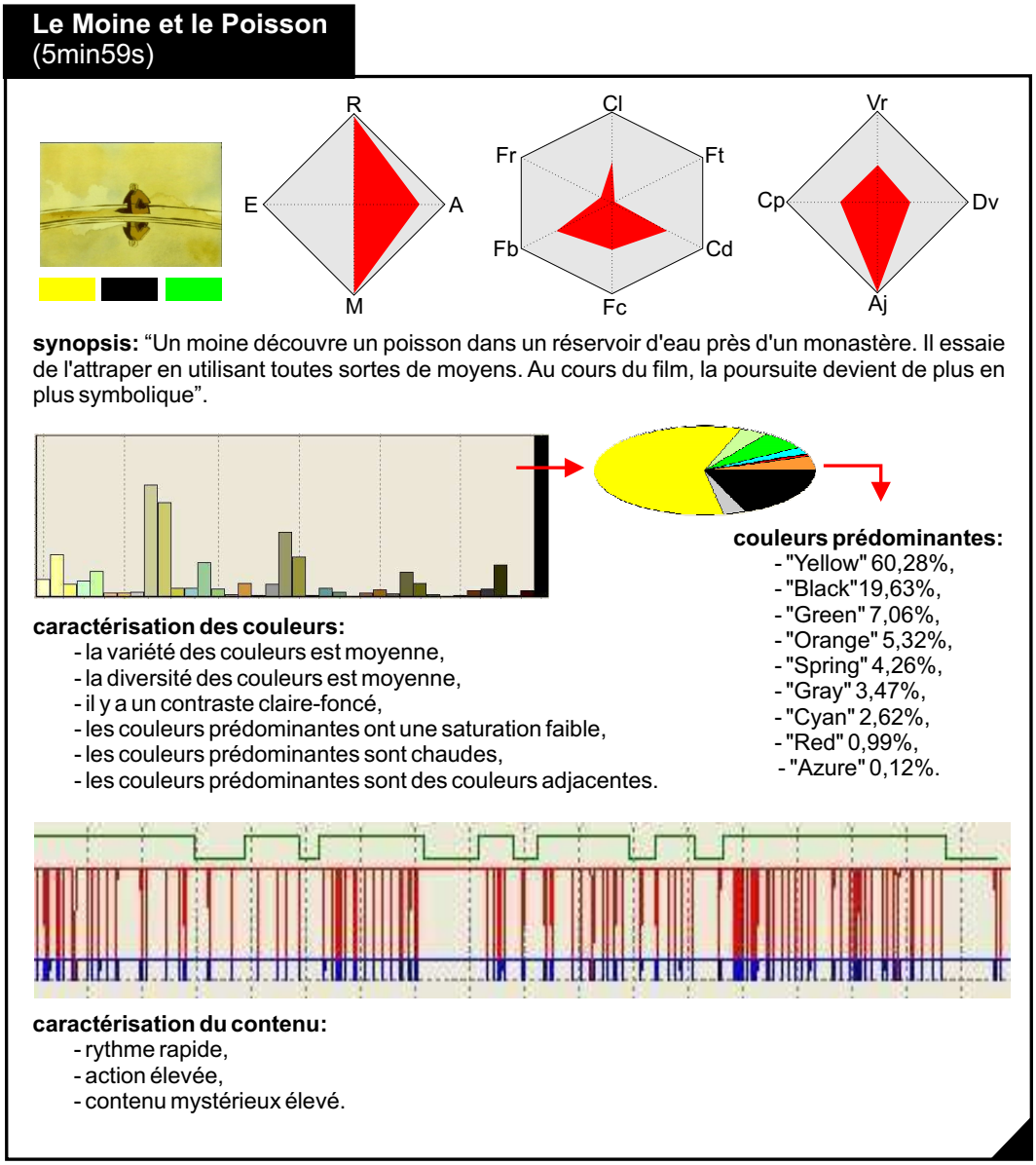


FIG. 7.22 – Exemple de fiche compacte pour le film "Le Moine et le Poisson" [Folimage 06b].

Enfin, nous envisageons de poursuivre cette approche de description symbolique en définissant d'autres attributs sémantiques. En particulier, les descripteurs couleurs que nous avons construits sont des caractérisations globales de la séquence qui ne prennent pas en compte la disposition spatiale des couleurs dans les images. Des approches mélangeant segmentation en régions et suivi de mouvement permettront d'envisager de nouvelles caractérisations.

Chapitre 8

Classification des films selon le contenu

Résumé : *Pour qu'un système d'indexation de séquences d'images soit facile à utiliser, il est indispensable que celui-ci soit basé sur des critères de haut niveau, permettant une description proche de la perception humaine. L'objectif de ce chapitre est d'étudier la pertinence des descriptions sémantiques/symboliques proposées dans les chapitres précédents vis à vis d'un système d'indexation d'une base de séquences d'images. Par ce système, nous cherchons à regrouper les séquences similaires du point de vue de leur contenu. Des tests expérimentaux ont été menés sur un échantillon de la base vidéo du Festival International du Film d'Animation. Nous avons effectué une classification selon les couleurs prédominantes (particularité de films d'animation), une classification selon l'action contenue dans la séquence et une classification selon les techniques de couleurs utilisées. Les résultats obtenus montrent le pouvoir discriminant de nos descripteurs, pouvant servir à des tâches d'indexation sémantique.*

Actuellement, les systèmes d'indexation de séquences d'images cherchent plutôt à s'appuyer sur des *critères sémantiques* proches de la perception humaine, que sur des approches syntaxiques classiques (pour un état de l'art, voir la Section 1.2).

Dans les chapitres précédents nous avons proposé différentes méthodologies pour extraire des caractéristiques sémantiques du contenu d'une séquence d'images. Nous avons envisagé trois directions différentes :

- l'analyse de *l'action contenue dans la séquence*,
- l'analyse des *techniques de couleurs employées*,
- l'analyse *du mouvement*.

L'objectif de ce chapitre est de *tester la pertinence* des descripteurs proposés pour une utilisation future comme index sémantiques dans un système de recherche de séquences d'images basé sur le contenu ("content-based retrieval"). Le système souhaité à terme est un système de recherche parmi les films d'animation du CICA [CICA 06], accessible en ligne sur Internet via le moteur de recherche Animaquid.

Les outils que possède actuellement le CICA permettent une recherche à partir d'informations textuelles fournies par les auteurs des films ou obtenues dans d'autres médias. Les

informations textuelles sont en général de courts résumés (synopsis) ou des fiches techniques décrivant le film. Ils ne permettent pas de décrire de façon détaillée le contenu artistique des films qui est typiquement *présent dans l'image* (voir la Section 1.5). Il semble donc nécessaire de disposer d'un système de recherche permettant aux artistes, ou aux personnes intéressées, d'accéder aux films d'une manière sémantique.

En complément des informations telles que le titre de la séquence, la technique d'animation utilisée, le réalisateur etc., paramètres qui ne sont pas toujours explicites pour l'utilisateur non-avisé, nous proposons d'utiliser des informations sémantiques sur la perception, plus facilement appréhendables. On peut ainsi envisager de rechercher les séquences *tristes* (caractérisées par des couleurs foncées, froides et un rythme lent), *joyeuses* (forte diversité des couleurs, rythme élevé, utilisation de couleurs complémentaires) ou une séquence similaire du point de vue de la distribution des couleurs à une autre sélectionnée par l'utilisateur, etc.

Afin de tester le pouvoir discriminant des paramètres que nous proposons, nous testons dans ce chapitre une *méthode de classification* ayant pour objectif de regrouper d'une manière automatique les différentes séquences en fonction de leur ressemblance selon différents critères basés sur le contenu des films (voir [Ionescu 06b][Ionescu 07a]).

8.1 La méthode de classification utilisée

Il existe de nombreuses méthodes de classification [Jain 99] : supervisées, non supervisées, paramétriques, hiérarchiques, stochastiques, etc. Notre propos ici n'est pas de faire un panorama des approches existantes ou de proposer une nouvelle technique. Nous désirons seulement tester la capacité des caractérisations proposées pour définir des classes sémantiquement pertinentes.

Nous avons décidé d'utiliser une méthode de classification non supervisée car nous ne disposons d'aucune vérité terrain sur la base vidéo (pas d'ensemble d'apprentissage). La méthode non supervisée choisie est la méthode des *nuées dynamiques*. C'est une méthode qui est un très bon compromis entre la rapidité et la qualité des résultats [Seber 84]. La méthode des nuées dynamiques consiste, en se fixant a priori un nombre de classes et une position initiale des centres de ces classes, à modifier itérativement les classes et les centres des classes selon un critère de minimisation de la dispersion de chaque classe. L'algorithme tend à construire un ensemble de classes compactes et bien séparées les unes des autres. Pour ce qui est du choix des paramètres associés à la méthode nous avons utilisé les considérations suivantes.

Comme pour la plupart des méthodes de minimisation numérique, la solution dépend souvent *de la position de départ des centres des classes*. Il se peut que les nuées dynamiques se positionnent sur un minimum local. La solution alors obtenue n'est pas globalement optimale. Pour remédier à cela, *le processus de classification est répété plusieurs fois* (10 itérations dans le cas de nos tests) en utilisant à chaque fois avec un nouvel ensemble initial des centres de classes. La solution finale retenue est la solution de l'itération qui correspond à la somme des distances entre les objets et les centres des classes la plus faible.

La méthode des nuées dynamiques est basée sur la notion de proximité entre objets. La méthode dépend donc *de la mesure de distance* utilisée. Après avoir réalisé plusieurs tests expérimentaux de classification en utilisant les distances Euclidienne, cityblock, cosinus et corrélation, nous avons pu observer que *la distance Euclidienne* a fourni les meilleurs résultats

avec nos données. Nous avons donc retenu cette distance pour nos tests de classification.

Un autre problème important est la *visualisation et l'interprétation des résultats*. Nous avons décidé d'évaluer la qualité de nos résultats de deux manières :

– par l'analyse *des silhouettes des classes*,
– par *l'analyse visuelle de la répartition des données* en classes.

Une silhouette est une représentation graphique de la distribution des objets des classes par rapport aux classes voisines [Kaufman 90]. La valeur de la silhouette pour un certain objet i de la classe C, notée $S(i)$, est donnée par l'équation suivante :

$$S(i) = \frac{DMM(i) - DM(i)}{max(DM(i), DMM(i))} \tag{8.1}$$

où $DM(i)$ est la distance moyenne de l'objet i à tous les autres objets de la même classe C et $DMM(i)$ est la valeur minimale des distances moyennes de l'objet i à tous les autres objets des classes différentes de C. Les valeurs de S varient de 1 (écart maximal) à -1 indiquant des objets qui ont été probablement associés par erreur à cette classe (valeur négative). Une valeur de 0 est donnée aux objets qui ne sont pas bien séparés des classes voisines. La silhouette est présentée en ordonnant les valeurs par ordre décroissant à l'intérieur de chaque classe. Un exemple est donné dans la Figure 8.1.

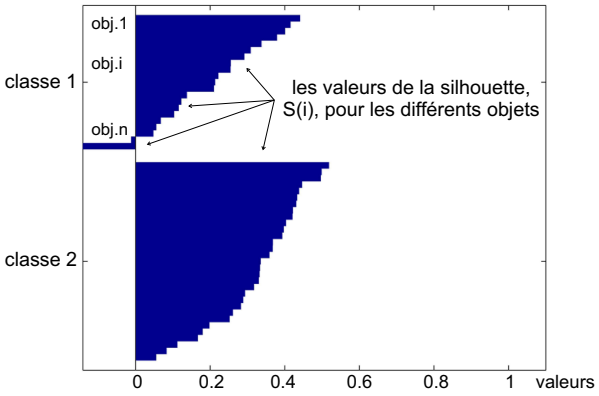


FIG. 8.1 – Exemple de silhouette dans le cas de deux classes (l'axe oX correspond à la valeur de la silhouette et l'axe oY aux objets des classes).

La silhouette nous fournit des informations sur le bon choix du nombre de classes, N. Si N est bien choisi, les valeurs de la silhouette sont alors élevées et donc les classes sont bien délimitées. Par exemple, dans la Figure 8.1 les deux classes obtenues doivent être très proches car il y a des valeurs négatives de S, indiquant que certains objets ont été mal classés, et de plus il y a un nombre important d'objets pour lesquels nous avons obtenu une valeur faible de S (inférieure à 0.2).

Pour l'interprétation des résultats nous avons utiliser l'analyse de l'image de la répartition des données en classes. Pour résoudre le problème de la visualisation de données à n-dimensions (avec $n > 3$) nous avons décorrélé les données par *analyse en composantes principales* (ACP). Nous avons ensuite choisi de visualiser ces données en les représentant dans un espace 2D ou 3D correspondant respectivement au deux ou trois premières composantes

principales, celles qui sont en général porteuses de la majorité de l'information. Les classes sont alors illustrées dans ces espaces de dimension réduite.

Enfin, le dernier paramètre à régler est *le nombre de classes*, N. Dans notre cas, N est complètement dépendant de la base de films que nous allons classer. La diversité des films fait que le choix de N est très difficile. Pour trouver la bonne valeur de N nous avons réalisé des tests en faisant varier le nombre de classes et chaque résultat a été évalué avec les méthodes décrites ci-dessus. Dans la suite nous allons détailler les différents tests que nous avons effectués.

8.2 Résultats expérimentaux

L'algorithme des nuées dynamiques est utilisé pour la classification d'un extrait de la base des films d'animation du CICA [CICA 06] dans le but de regrouper les films selon la similarité de leur contenu (voir [Ionescu 07c]). La base vidéo utilisée comporte 52 courts métrages d'une durée totale de 6 heures et 7 minutes et comporte une bonne variété de genres et de techniques d'animation (voir l'Annexe F). Le contenu de chaque film a été analysé manuellement et répertorié en utilisant toutes les informations textuelles dont disposait le CICA (le moteur de recherche Animaquid), ainsi que des informations recueillies sur les sites Internet des producteurs. Ce travail nous a permis de construire une sorte de vérité terrain nous aidant à l'interprétation et la validation des résultats.

8.2.1 Classification en fonction de l'action et des couleurs

Nous avons réalisé une classification en utilisant les informations issues de l'analyse conjointe de l'action et des couleurs de chaque film. Nous avons utilisé comme données d'entrée les degrés d'appartenance de chaque concept sémantique/symbolique proposé dans la Section 7.2 (pour les couleurs) et Section 7.3 (pour l'action).

La classification a été réalisée sur la base des 52 films, chaque vecteur de caractéristiques contenant 18+7 paramètres (couleur + action). Dans la sélection de ces paramètres, nous avons réduit la redondance des données en utilisant le fait que les descriptions floues ont toujours été construites de manière à ce que la somme des différents degrés fasse 1. Par exemple, pour le concept "présence de couleurs claires" nous n'utilisons que les degrés d'appartenance aux symboles *"présence faible de couleurs claires"* et *"présence moyenne de couleurs claires"*, car le degré d'appartenance au symbole *"présence élevée de couleurs claires"* se déduit des deux autres degrés.

Puisque nous ne connaissons pas le nombre exact de classes dans la base de films, nous avons fait varier le nombre de classes, N, de 2 à 7.

En analysant les résultats (les silhouettes et la répartition en classes - Figure 8.2) nous avons obtenu deux catégories intéressantes :

- d'une part les films caractérisés par une action réduite et une diversité des couleurs également réduite, par exemple les films : "Amerlock", "Sculptures", "The Wall", "At the Ends of the World" (voir l'Annexe F),
- d'autre part les films comportant une diversité des couleurs moyenne, des couleurs foncées et une action élevée, par exemple les films "La Bouche Cousue", "La Grande Migration", "Nos Adieux au Music Hall", "Petite Escapade" (voir l'Annexe F).

Dans une classification en deux classes ($N = 2$), ces films sont réunis en une seule classe (voir classe C dans la Figure 8.2). Pour N allant de 3 à 7, on constate que ces deux catégories sont bien isolées et restent pratiquement inchangées quel que soit la valeur de N (classes $C1$ et $C2$ dans la Figure 8.2). On peut noter que la classe $C1$ contient en particulier les films utilisant comme technique d'animation la pâte à modeler[1].

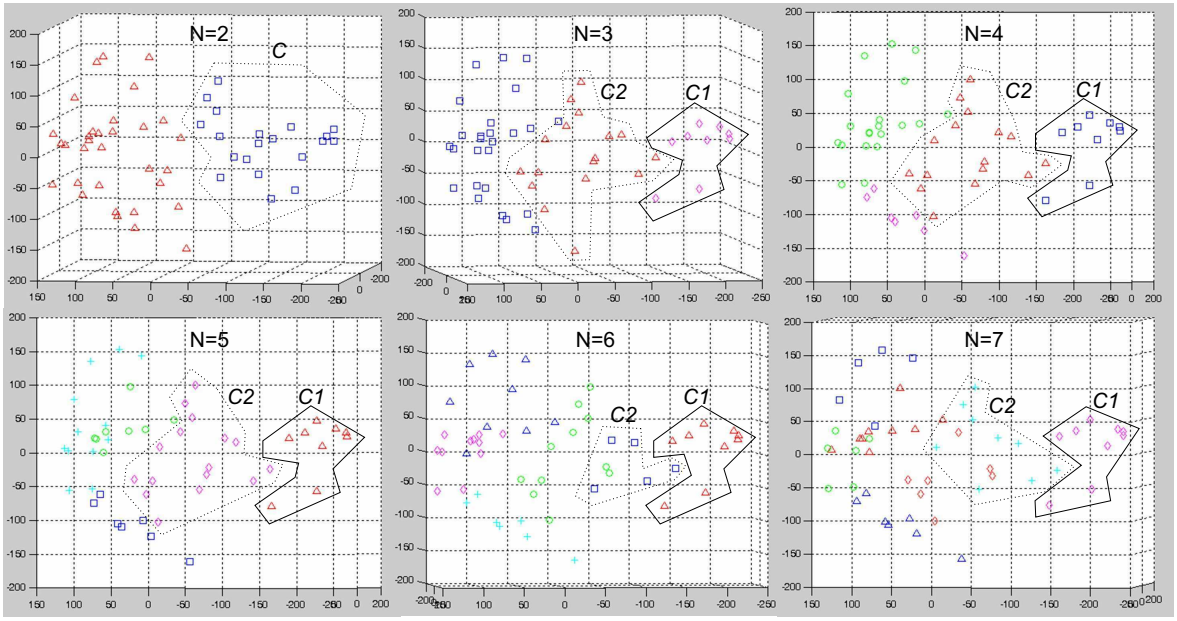


FIG. 8.2 – La répartition des films pour un nombre de classes, N, variant de 2 à 7. Chaque film est représenté par les trois premières composantes principales du vecteur de caractéristiques. Le point de vue de la visualisation a été choisi pour représenter au mieux les classes.

La collaboration entre ces deux sources d'informations nous a permis de séparer les films ayant : *une action, un rythme et une diversité des couleurs réduite* - classe $C1$, et les films ayant *une action élevée/moyenne et une diversité des couleurs moyenne* - classe $C2$.

A l'exception de ces deux catégories que l'on a réussi à retrouver, les autres classes présentent une forte disparité des caractéristiques des films qui les composent. Ceci est dû à la diversité importante de la base et au fait que le contenu des couleurs n'est pas vraiment corrélé à l'action contenue dans la séquence, sauf pour la technique d'animation particulière qu'est la pâte à modeler. Par exemple, un film avec une distribution de couleurs froides et foncées n'est pas forcement lié à une action réduite (voir comme exemple le film "François le Vaillant" dans la Section 7.5). L'utilisation conjointe de ces deux informations n'est donc pas efficace. Nous allons donc effectuer des tests de classification indépendamment sur les descriptions des couleurs et de l'action contenue dans la séquence.

8.2.2 Classification selon les couleurs prédominantes

Le second test de classification a été motivé par le fait que la plupart des films d'animation utilisent des palettes de couleurs particulières (voir la Section 1.5 ou la Section 4.3). Les

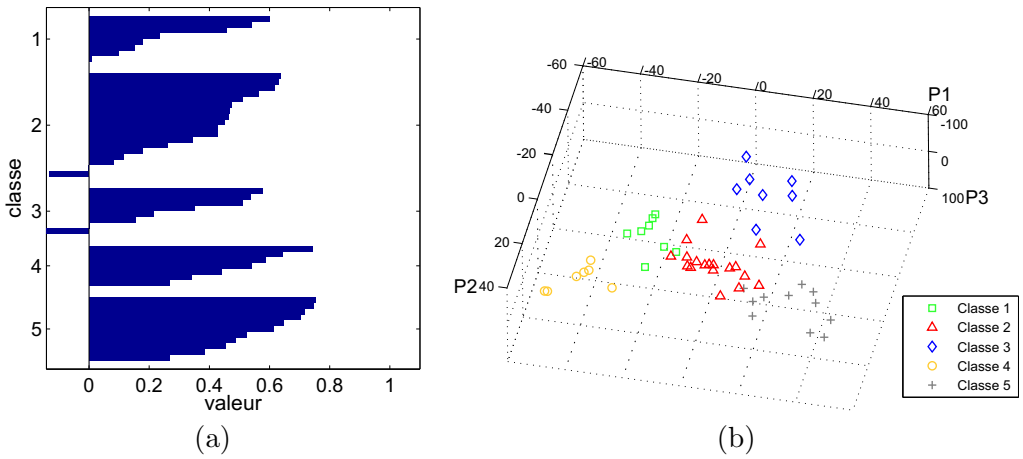
[1]cette technique est caractérisée par une palette de couleurs réduite et un rythme réduit, l'action se déroulant généralement sur peu de scènes, voir le film "Amerlock" dans l'Annexe G

couleurs prédominantes sont une caractéristique importante des films d'animation. Ce test consiste à regrouper les films selon la similarité de la distribution globale de leurs couleurs. Le paramètre utilisé pour effectuer ce regroupement est l'histogramme des couleurs élémentaires, $h_{élém}()$, défini par l'équation 7.1 de la Section 7.2.1. $h_{élém}()$ est une mesure de la répartition globale des couleurs élémentaires de la séquence.

Pour déterminer le bon nombre de classes, N, que nous allons utiliser pour la classification, nous avons effectué d'abord une classification manuelle. Plusieurs personnes ont regroupé les films en fonction de leur contenu en terme de couleurs. Après avoir réalisé l'intersection des résultats obtenus (les différentes personnes ont regroupé les films de manières légèrement différentes), nous avons pu observer que parmi les 52 films, il y a 5 classes homogènes du point de vue de la distribution des couleurs :

- $classe_1$ - couleur prédominante Verte,
- $classe_2$ - couleur prédominante Rouge/Marron,
- $classe_3$ - Bleu,
- $classe_4$ - Jaune/Orange,
- $classe_5$ - Gris/Noir.

Nous avons donc appliqué la méthode de classification par nuées dynamiques en utilisant $N = 5$ classes et avons choisi de prendre comme vecteurs de caractéristiques pour chaque film, les 15 valeurs (12 couleurs élémentaires, ainsi que le gris, le noir et le blanc) de l'histogramme élémentaire $h_{élém}()$. Les 52 films ont été regroupés en 5 classes en fonction des couleurs prédominantes. La répartition en 5 classes est illustrée par la Figure 8.3.



(a) (b)

FIG. 8.3 – La répartition en classes : (a) les silhouettes des classes, (b) la répartition des films en classes (chaque film est représenté par les trois premières composantes principales de $h_{élém}()$, notées P_1, P_2 et P_3).

Les silhouettes des classes nous permettent de dire que nous avons obtenu une bonne délimitation et une bonne homogénéité des classes car la plupart des valeurs des silhouettes sont supérieures à 0.4.

Pour interpréter les résultats nous avons utilisé une projection 2D de l'espace 3D de la répartition des films en classes, répartition présentée par la Figure 8.3.b. La projection a été choisie de façon à avoir la meilleure visualisation des frontières entre les classes. De plus, chaque film est représenté par une image caractéristique de la séquence (l'image de

chaque film est centrée sur la valeur initiale du graphe 3D). Le nouveau graphe ainsi obtenu est illustré par la Figure 8.4 (les classes ont été manuellement délimitées par une ligne en couleur facilitant la visualisation).

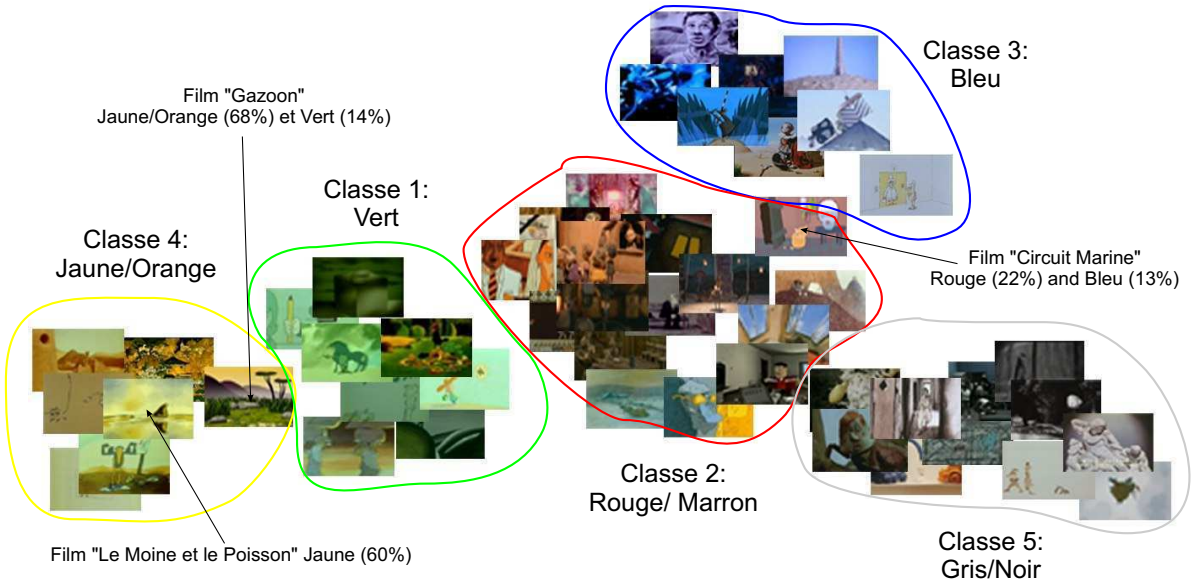


FIG. 8.4 – Une projection 2D de l'espace 3D de la répartition en 5 classes issues de la Figure 8.3.

Tout comme la construction de la vérité terrain, l'évaluation de la pertinence des résultats est quelque chose de subjectif, puisque cela dépend de la perception humaine. Comme chaque personne perçoit de manière différente le contenu des films, il est très difficile de porter un jugement de valeur sur cette classification. Mais, nous pouvons tout de même noter que l'intersection entre la classification réalisée avec la méthode des *nuées dynamiques* et celle que nous avons réalisée "à la main", notre vérité terrain, est de plus de 90%.

L'histogramme des couleurs élémentaires nous a permis de regrouper les films ayant des couleurs similaires. Chaque film présentant une ou deux couleurs prédominantes a été associé à une classe différente. Par exemple, le film "Le Moine et le Poisson" (couleur prédominante : le Jaune à 60%) est le centre de la $classe_4$ (voir la Figure 8.4). De plus, les films présentant plusieurs couleurs prédominantes se retrouvent proches des classes contenant ces couleurs, comme le film "Gazoon" (couleurs prédominantes : Jaune/Orange 68% et Vert 14%) qui appartient à la $classe_4$, qui est le cluster Jaune/Orange, tout en étant proche également de la frontière de la $classe_1$ (couleur prédominante Verte). Un autre exemple est le film " Circuit Marine " (couleurs prédominantes : Rouge 22% et Bleu 13%) qui appartient à la $classe_2$ (couleur prédominante Rouge/Marron) mais reste proche de la frontière de la $classe_3$ (couleur prédominante Bleue).

8.2.3 Classification en fonction des techniques de couleurs utilisées

Le troisième test concerne la classification des films en fonction des techniques de couleurs utilisées. La classification par nuées dynamiques a été effectuée en utilisant, comme pour les tests précédents, les degrés d'appartenance des descriptions sémantiques/symboliques des

couleurs proposées dans la Section 7.2, soit 18 paramètres pour chaque film de la base. Nous avons fait varier le nombre de classes, N, de 2 à 4. La répartition en classes et les silhouettes sont présentées dans la Figure 8.5. Nous avons pu observer une délimitation correcte des classes car les valeurs des silhouettes sont en général supérieures à 0.4.

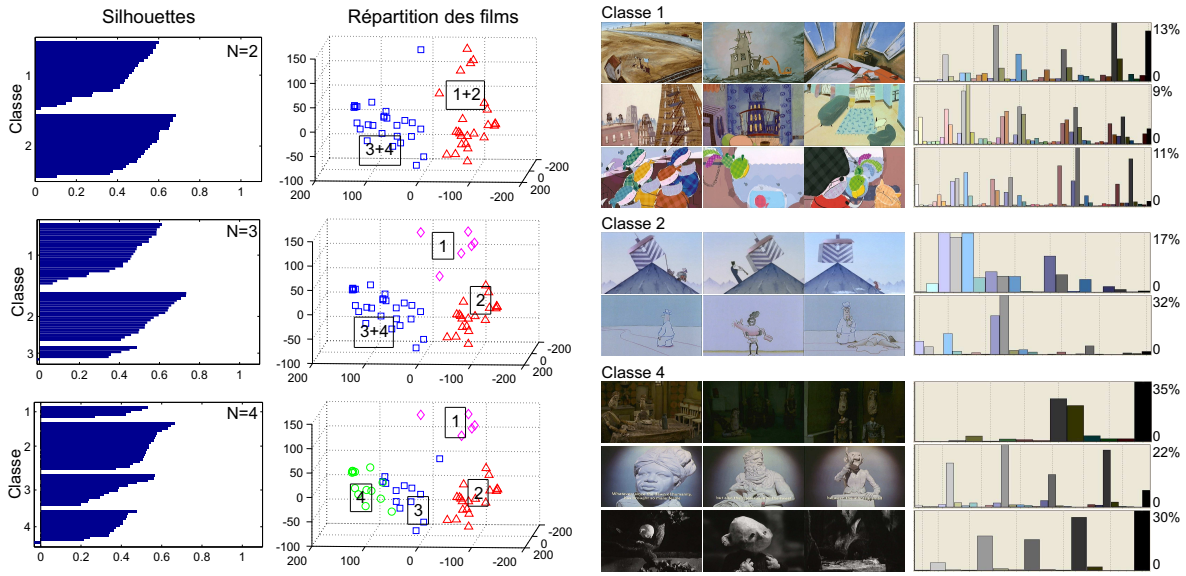


FIG. 8.5 – Les résultats de la classification en fonction des techniques de couleurs utilisées. Pour la répartition en classes, chaque film est représenté par les trois premières composantes principales du vecteur de caractéristiques (le numéro de la classe est encadré). Certains films représentatifs des classes sont résumés par quelques images et l'histogramme $h_{seq}()$.

Comme pour les tests précédents, nous avons validé la pertinence des résultats en analysant les regroupements obtenus. Nous avons trouvé que les films se retrouvent bien groupés selon la similarité des techniques de couleurs utilisées (voir la Figure 8.5) :

- pour $N = 2$ les films sont divisés en deux groupes : les films ayant un contenu riche en couleurs ($classe_{1+2}$) et les films utilisant une palette réduite de couleurs ($classe_{3+4}$).
- en augmentant le nombre de classes à $N = 3$, seule la $classe_{1+2}$ se subdivise : la $classe_2$ contient les films ayant une diversité moyenne et des couleurs adjacentes, $classe_1$ contient les films ayant une variété/diversité élevée des couleurs,
- pour $N = 4$ les $classe_1$ et $classe_2$ restent inchangées. Seule la $classe_{3+4}$ se divise en une $classe_3$ contenant quelques films à palette couleur réduite mais sans autres caractéristiques communes et une $classe_4$ contenant la plupart des films ayant une palette très réduite et ayant comme couleurs prédominantes des couleurs sombres.

De plus, pour N variant de 2 à 4 la classification a tendance à préserver l'homogénéité de certaines classes comme le montre la Figure 8.5.

En conclusion ce test nous a permis de retrouver les *films très colorés* ($classe_1$), les *films sombres* ($classe_4$) et enfin les films ayant plutôt des *couleurs adjacentes* ($classe_2$).

8.2.4 Classification selon l'action contenu dans la séquence

Le dernier test est une classification en utilisant les degrés d'appartenance des descriptions de l'action contenue dans la séquence, paramètres proposés dans la Section 7.3 (7 paramètres). Nous avons fait varier le nombre de classes N de 2 à 5. Certains résultats obtenus sont présentés dans la Figure 8.6.

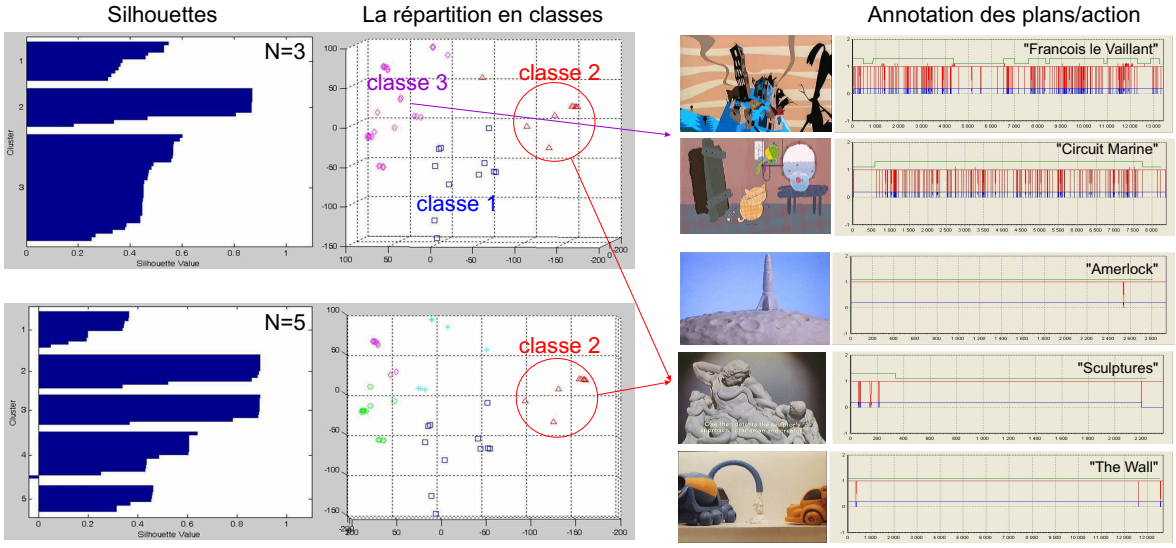


FIG. 8.6 – La répartition en classes en fonction de l'action (chaque film est représenté par les trois premières composantes principales du vecteur de caractéristiques).

Après le dépouillement manuel des résultats nous avons pu remarquer que pour $N = 3$ les films sont divisés en trois classes particulières :

- $classe_1$ contenant les films ayant un rythme élevé et une action moyenne,
- $classe_2$ contenant les films ayant un rythme faible et une action faible,
- $classe_3$ contenant les films ayant une action élevée, un rythme élevé et un contenu mystérieux élevé.

En faisant varier le nombre des classes de 3 à 5, la $classe_2$ est toujours correctement identifiée. Comme nous l'avons déjà dit, cette classe regroupe principalement les films ayant un rythme réduit, c'est-à-dire les films qui ont un nombre restreint de plans. C'est le cas de certaines techniques d'animation, comme la pâte à modeler, l'utilisation du sable, ou l'animation de figurines en papier, où, pour des raisons de difficulté de réalisation (chaque scène est construite réellement, le mouvement des personnages est fait manuellement et est enregistré image par image, etc.), les auteurs utilisent très peu de plans. Cette classification selon l'action permet donc indirectement une classification, encore grossière, selon les techniques.

8.3 Conclusions générales

Dans ce chapitre nous avons présenté nos premiers travaux sur la classification des films en fonction des couleurs et de l'action contenue dans la séquence. L'intérêt de nos tests, est de valider *le pouvoir discriminant* des descriptions sémantiques/symboliques proposées.

Ces descripteurs pourront nous servir d'index de recherche dans un système d'indexation sémantique CBR ("content-based retrieval").

Différents tests de classification ont été effectués en utilisant la méthode de classification non supervisée par *nuées dynamiques* appliquée sur un extrait de la base des films d'animation du CICA [CICA 06] (52 films). Pour remédier au problème de l'interprétation et de la validation des résultats nous avons analysé manuellement les distributions des films en utilisant comme "vérité terrain" les différentes informations disponibles sur le moteur de recherche Animaquid et sur Internet. Nous pouvons alors dire que :

- **la classification selon l'action et la distribution des couleurs** : nous n'avons pas trouvé de relation particulière entre l'action contenue dans la séquence et la distribution couleur car ces deux informations ne semblent pas corrélées. Notons toutefois, que la technique d'animation qui utilise la pâte à modeler est le seul groupe qui ressort clairement de cette classification. Ceci s'explique par un jeu de couleurs restreint, une palette et une action réduite, puisque très peu de scènes sont réalisées à cause de la difficulté de réalisation de cette technique, comme le montre la Section 8.2.1.

- **la classification selon les couleurs prédominantes** : l'histogramme des couleurs élémentaires, $h_{élém}()$, permet de regrouper les films selon la similarité de la palette des couleurs utilisées comme le montre le test présenté dans la Section 8.2.2. Ce résultat est important dans le contexte particulier des films d'animation car la plupart des films d'animation utilisent des palettes particulières, ce qui donne à la couleur un pouvoir discriminant, permettant ainsi de classer les films d'animation.

- **classification selon la technique de couleur utilisée** : la description des couleurs nous a permis de retrouver d'abord les films colorés (contenu riche en couleurs différentes), ensuite les films sombres (contenu pauvre en couleurs, prédominance de couleurs foncées) et enfin les films utilisant des couleurs adjacentes (un nombre réduit de teintes élémentaires adjacentes, voir la Section 8.2.3).

- **la classification selon l'action** : les descripteurs de l'action contenue dans la séquence (rythme, action, mystère et explosivité) nous ont permis de retrouver les films ayant un rythme et une action faible (c'est le cas d'un certain nombre de techniques d'animation, comme la pâte à modeler, l'utilisation du sable ou l'animation de figurines en papier) et les films ayant une action importante, un rythme élevé (voir la Section 8.2.4).

Néanmoins, les résultats que nous avons obtenus sont fortement dépendants du choix des attributs sur lesquels nous avons effectué la classification. Ces attributs sont en fait une traduction de la connaissance du domaine étudié. En ce sens, cette approche ne présente pas un aspect générique.

Pour conclure, notons que les résultats de la classification représentent pour nous, le point de départ d'une tâche plus importante qu'est la recherche de films à partir du contenu. Nous avons montré que les descriptions floues du contenu ont un bon potentiel et peuvent donc servir à l'indexation de films. Cependant, l'ensemble des paramètres utilisés n'est pas complet. D'autres paramètres sont à prendre en compte dans notre analyse, en particulier le *mouvement*.

Quatrième partie

Conclusions et perspectives

Chapitre 9

Conclusions et perspectives

9.1 Conclusions

Les travaux proposés par cette thèse ont comme but la constitution d'un système d'analyse et d'indexation du contenu de séquences d'images. L'apport original du notre travail se situe dans la nature sémantique des informations extraites, les aspects sémantiques constituant un point dur des systèmes d'indexation actuels (voir le Chapitre 1).

D'une manière générale, les systèmes d'indexation sémantique que l'on est capable de construire aujourd'hui ne présentent pas de caractère général. Ils sont le plus souvent adaptés à un domaine d'application particulier. Et, plus on augmente le niveau de description sémantique, plus le système perd de sa généralité. Le développement d'un système général reste actuellement une barrière infranchissable.

Ainsi, même si notre démarche a été construite en lui donnant un caractère générique, l'application de cette démarche a été faite dans un domaine spécifique, celui du film d'animation. Le contexte local a joué un rôle important dans le choix de ce domaine. En effet, Annecy, avec son Festival International du Film d'Animation, est devenu depuis plus de quarante ans une référence mondiale dans le monde de l'animation. De plus, l'industrie de l'animation a connu ces dernières années un essor important, en particulier grâce à l'évolution des techniques de synthèse d'images 3D. Dans ce contexte, nos travaux constituent une des premières démarches s'intéressant à l'indexation sémantique des films d'animation.

La plupart des études menées dans cette thèse se focalisent sur l'annotation du contenu qui est en général le "cœur" d'un système d'indexation. La problématique a été abordée en utilisant une analyse à deux niveaux :

- **l'analyse de bas niveau** où les séquences sont décrites par des paramètres statistiques liés à la structure du film, la couleur et le mouvement,
- **l'analyse sémantique/symbolique** où les mesures de bas niveau sont transformées

en concepts linguistiques à l'aide d'une connaissance experte. Cette étape nous permet d'atteindre un niveau de description plus proche de la perception humaine.

Notons que lorsque nous avons débuté nos travaux, il n'y avait pas d'expérience dans le domaine de l'indexation vidéo au sein de l'équipe Traitement de l'Information du LISTIC.

9.1.1 L'analyse de bas niveau

Pour atteindre un niveau sémantique il faut d'abord extraire un certain nombre de paramètres de bas niveau décrivant les propriétés que l'on cherche à caractériser. La qualité de la description finale est bien sûr liée au bon choix de ces paramètres. La variété élevée des informations contenues dans une séquence rend ce choix difficile.

Parmi tous ces paramètres, on peut cependant trouver une certaine hiérarchie. Nous avons ainsi choisi ceux qui nous ont semblé les plus importants pour le contenu d'une séquence d'images : la couleur, la structure temporelle et le mouvement. La démarche que nous avons adoptée alors est constituée de deux étapes : une première étape dont le degré de granularité est l'image, et une seconde étape d'agrégation permettant d'extraire des caractéristiques globales à toute la séquence.

Au niveau structurel de la séquence nous avons étudié la problématique du découpage en plans, étape incontournable pour l'analyse du contenu vidéo. Cette étape, assez classique, a nécessité des développements spécifiques pour s'adapter aux caractéristiques particulières des films d'animation, en ayant deux objectifs principaux : la robustesse et le caractère automatique, réduisant ainsi le plus possible l'intervention humaine dans le système. Ce niveau a permis de construire des caractéristiques liées au rythme et de l'action.

Pour l'analyse du mouvement nous avons utilisé une approche qui mélange l'étude de la continuité/discontinuité du mouvement avec la caractérisation de la nature du mouvement, permettant ainsi de détecter les transitions.

Pour l'analyse des couleurs nous avons proposé une signature globale de la séquence qui prend en compte l'aspect temporel de la séquence. Cette approche est basée sur l'utilisation d'une palette de couleurs particulière associée à un dictionnaire des noms des couleurs, préparant ainsi l'analyse sémantique.

9.1.2 L'analyse de plus haut niveau

La détection de scènes et la construction de résumés constituent une sorte d'étape intermédiaire entre la caractérisation bas niveau et la description sémantique.

Le développement de certaines mesures de similarité entre le contenu des plans a permis une analyse de la décomposition en scènes de la séquence. Cette analyse a l'avantage de fournir une meilleure compréhension des relations existant entre les différents passages de la séquence.

D'autre part, le découpage en plans et l'analyse du rythme de déroulement de l'action ont été utilisés pour résumer le contenu de la séquence, étape nécessaire pour la tâche de navigation. A ce niveau, notre apport consiste dans la proposition d'un résumé intelligent, similaire à la "bande-annonce" d'un film, et dans le développement d'une méthodologie permettant la construction de résumés compacts constitués seulement de quelques images représentatives de la séquence.

9.1.3 L'analyse sémantique/symbolique

D'une manière générale l'analyse sémantique du contenu est une étape difficile car elle est dépendante du domaine d'application. De plus, la définition de symboles est souvent subjective car fortement liée à la façon de percevoir de chacun. Enfin, son évaluation demande l'intervention humaine et les vérités terrain ne sont pas toujours faciles à constituer.

Notre démarche s'appuie sur la représentation des paramètres de bas niveau par des ensembles flous et la modélisation par des règles floues. Ce choix a été motivé par deux facteurs. D'une part la représentation floue permet la conversion des valeurs numériques en concepts linguistiques. D'autre part elle utilise l'introduction "naturelle" de l'expertise humaine.

La caractérisation sémantique demande le choix de termes linguistiques. Ce choix est parfois immédiat, mais dans quelques situations, il s'est avéré délicat, mettant en évidence l'absence de consensus entre spécialistes. Certains des termes proposés pourront encore évoluer.

Notre contribution principale consiste dans la prise en compte de la connaissance pour définir un certain nombre de symboles et concepts pertinents pour la description du contenu. La validation des résultats a été effectuée sur plusieurs niveaux. Pour les résumés nous avons organisé une campagne d'évaluation sur la pertinence de nos résumés impliquant le jugement humain. Pour la description sémantique, ne disposant pas d'une réelle vérité terrain, nous nous sommes limités à la confrontation des résultats avec différentes informations périphériques (synopsis, fiches techniques, commentaires, etc.).

Enfin, pour valider la possibilité d'utilisation de nos descripteurs en tant qu'index sémantiques de recherche dans un système d'indexation, nous les avons exploités à travers une classification de données.

Notons cependant que les caractérisations sémantiques/symboliques que nous avons proposées ne sont pas encore complètes. D'autres informations sont à prendre en compte, comme par exemple le mouvement, la texture, etc. La difficulté de choisir la bonne terminologie fait que le choix de certains termes est encore ouvert. Enfin, certaines descriptions, comme le rythme, ne constituent pas une description absolue, mais restent en partie relatives à la séquence étudiée.

9.2 Nos perspectives

Dans la Figure 9.1 nous avons illustré la structure du système proposé dans cette thèse (voir la Section 1.5). Une échelle relative a été utilisée pour désigner l'avancement de nos travaux pour chacune des étapes de traitement proposées.

D'une manière générale, les perspectives envisagées peuvent être effectuées à deux niveaux différents. D'abord il y a différentes améliorations que l'on peut apporter sur les méthodes proposées. Ensuite, la direction la plus importante et la plus prometteuse est la fusion de différentes sources d'information (analyse *multimodale*) pour atteindre une meilleure caractérisation sémantique.

9.2.1 Amélioration des méthodes proposées

Tout d'abord, maintenant que les outils sont en place, il est nécessaire de les tester sur une plus grande variété de films. Des problèmes de droits d'auteur nous ont contraints à

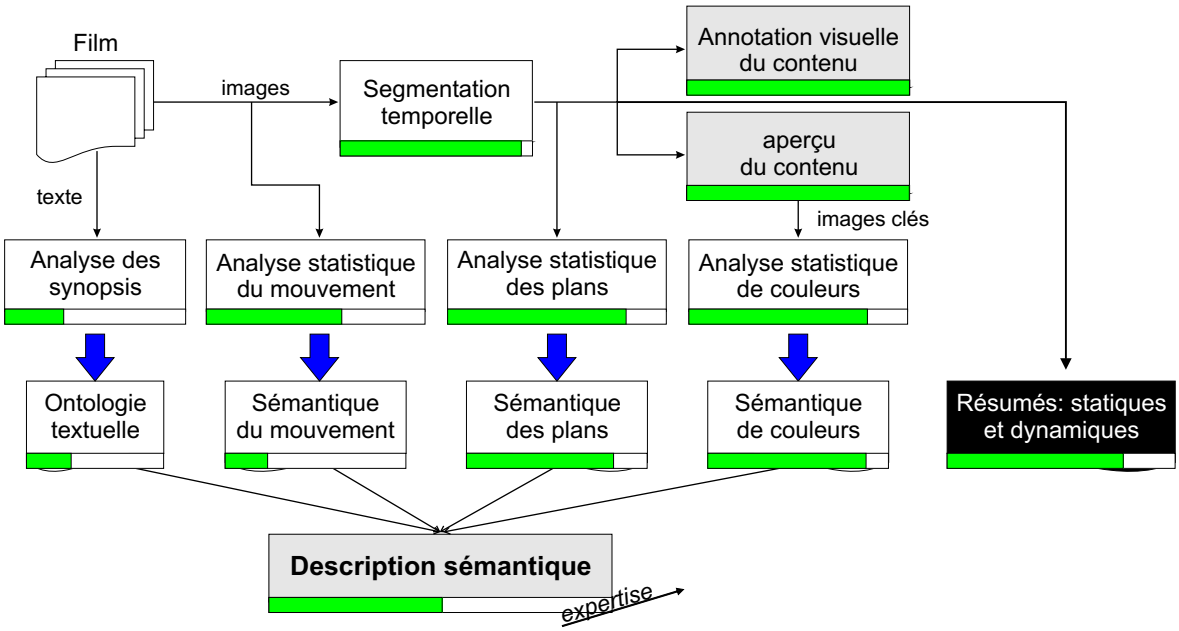


FIG. 9.1 – L'avancement des nos travaux (la ligne verte indique la quantité de travail accomplie).

l'utilisation d'un nombre restreint de films (52). Des accords récents avec CITIA vont nous permettre d'élargir la base de test, ce qui demandera probablement d'adapter les outils à une plus grande variété de sujets. Notons cependant qu'un accord a été déjà signé avec la société Folimage, société française de production de films d'animation, pour les droits d'utilisation de leurs films.

Pour la segmentation temporelle une amélioration du découpage en plan peut être obtenue par la constitution d'une approche couleur-mouvement basée sur une mesure de pertinence de ces deux approches. Nous pouvons également envisager une amélioration de la méthode de calcul du seuil de détection par une approche localement adaptative.

En ce qui concerne l'analyse du mouvement une amélioration peut être obtenue par une analyse plus fine du mouvement, permettant en particulier la détection et le suivi de personnages. Cette analyse plus fine permettra également une estimation 3D du mouvement, ce qui lèvera un certain nombre d'ambiguïtés des estimations par approximation 2D.

Pour l'analyse des couleurs la principale contrainte de notre approche est qu'elle ne prend pas en compte les relations spatiales entre les couleurs, l'outil de base étant l'histogramme. Pour une caractérisation plus exacte une segmentation en régions homogènes de l'image devrait être considérée.

Les techniques d'extraction de résumés que nous avons proposées utilisent principalement comme source d'information la structure temporelle et la couleur. Une amélioration consisterait dans l'apport d'une analyse de l'activité à l'intérieur des plans (intra-plan) basée sur le mouvement.

Pour la méthodologie de caractérisation sémantique/symbolique que nous avons proposée, la première amélioration consisterait à enrichir le lexique utilisé : en introduisant des termes spécifiques au domaine, en ajoutant de nouveaux concepts, etc. Pour augmenter la généralité

de nos concepts linguistiques (un utilisateur non avisé n'est pas nécessairement familiarisé avec nos termes) nous envisageons d'associer un dictionnaire de synonymes. Par exemple le concept "mystérieux" est proche des termes "énigmatique", "étrange", "singulier", etc.

A partir des descriptions symboliques/sémantiques proposées, on peut aussi penser à franchir un pas supplémentaire dans le niveau sémantique de ces descriptions en les fusionnant pour composer des termes encore plus proches de notre perception. Par exemple, la joie peut être vue comme l'association de couleurs variées, chaudes et d'un rythme élevé. La tristesse peut correspondre à peu de couleurs, des couleurs sombres et un rythme lent.

D'autres pistes concernent l'exploitation et la validation des termes proposés. En particulier, pour la classification, on peut envisager l'utilisation de méthodes plus adaptées à notre application (intégration de données d'apprentissage par exemple) et l'exploitation d'une base de données plus vaste, impliquant une plus grande diversité de genres. Nous envisageons aussi d'améliorer le processus de partition des films selon leur contenu par une classification effectuée en deux temps : d'abord sur la couleur et ensuite en terme d'action. Une autre perspective est la mise en place d'une étude plus élaborée sur la validation de la mesure de distance entre les contenus de films que nous avons proposée (le gamut sémantique).

9.2.2 Vers l'analyse multimodale

Un enrichissement des descriptions sémantiques pourrait être obtenu en associant d'autres modalités. En particulier, le son des films et les textes ou péritextes associés sont des sources d'information importantes et complémentaires aux images.

En ce qui concerne le texte nous avons déjà commencé d'étudier une approche conjointe structure-couleur-texte, les travaux étant réalisés en collaboration avec l'équipe d'Ingénierie des Connaissances du LISTIC [Condillac 05] (voir [Beauchêne 05a] [Beauchêne 05b]).

L'information textuelle est extraite de certaines sources périphériques d'informations, comme par exemple les informations fournies par la base de films (titres, noms des auteurs, etc.) ou les fiches descriptives des films disponibles sur le moteur de recherche Animaquid (synopsis, données techniques, etc.). Une connaissance terminologique est apportée ensuite par la construction d'un modèle ontologique OK ("Ontological Knowledge") sur les différentes données textuelles (voir un exemple dans la Figure 9.2).

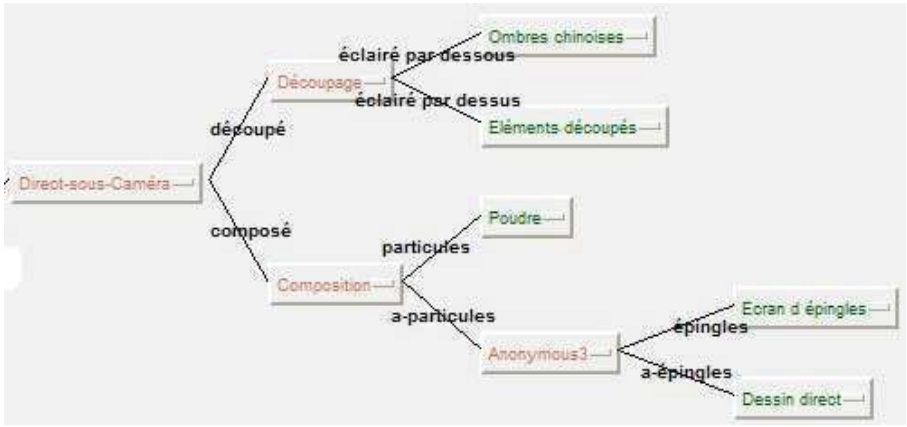


FIG. 9.2 – Extrait de l'ontologie des techniques.

Nous attendons de la fusion de ces deux sources de connaissances, image et texte, l'émergence de nouveaux concepts permettant de caractériser, par exemple, l'atmosphère (ou ambiance) qui se dégage d'un film. Nous envisageons deux approches de fusion :

- **le texte aidant les images** : la procédure, qui s'apparente alors à une classification supervisée des images (voir [Beauchêne 05a]). Selon ce schéma, la terminologie utilisée est celle apportée par le texte, l'image venant renforcer ou suppléer le texte. Bien sûr, l'application de cette démarche demande un choix de descripteurs image en lien avec la terminologie issue du texte. Par exemple, il semble que les caractéristiques attachées au rythme et à la couleur doivent pouvoir se fusionner à une terminologie sur les genres.

- **l'analyse conjointe des descripteurs** : dans cette deuxième approche, le texte et l'image sont d'abord exploités indépendamment de manière à fournir leurs propres caractéristiques. Ceci suppose que l'analyse des images parvienne à fournir des caractéristiques dont la nature sémantique soit suffisante. On peut pour cela s'aider des transcriptions numérique / symbolique proposées par la logique floue. Ensuite, en se plaçant dans l'espace de dimension N des caractéristiques (N est le nombre total de caractéristiques de type texte et image), on procède à une analyse factorielle dans le but de faire émerger des axes principaux caractéristiques. L'intervention des experts est alors indispensable pour donner du "sens" à ces axes factoriels.

La seconde piste de la description multimodale est la prise en compte du son. Le son joue un rôle très important dans la caractérisation sémantique du contenu des films d'animation (présence de dialogue, intensité et rythme de la musique, silences, etc.). Cette analyse conjointe son-image sera abordée dans le cadre d'un projet régional de la région Rhone-Alpes sur l'analyse des séquences d'image.

Cinquième partie

Bibliographie

Bibliographie

[4i2i 06] 4i2i. *H.263 Video Coding Tutorial.* http ://www.4i2i.com/h263_video_codec.htm, 2006.

[Acosta 02] E. Acosta, L. Torres, A. Albiol & E. Delp. *An Automatic Face Detection and Recognition System for Video Indexind Applications.* IEEE, vol. 4, pages 3644–3647, 2002.

[Adames 02] B. Adames, C. Dorai & S. Venkatesh. *Towards Automatic Extraction of Expressive Elements of Motion Pictures : Tempo.* IEEE Transactions on Multimedia, vol. 4, no. 4, pages 472–481, decembre 2002.

[Adjeroh 01] D.A. Adjeroh & M.C. Lee. *On Ratio-Based Color Indexing.* IEEE Transactions on Image Processing, vol. 10, no. 1, pages 36–48, janvier 2001.

[Aigrain 95] P. Aigrain, P. Jolly & V. Longueville. *Medium Knowledge-Based Macro-Segmentation into Sequences.* Working notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval, pages 5–14, Montreal, Canada 1995.

[Akutsu 92] A. Akutsu, Y. Tonomura, H. Hashimoto & Y. Ohba. *Video Indexing Using Motion Vectors.* SPIE Visual Communications Image Processing, vol. 1818, 1992.

[Alatan 01] A.A. Alatan, A.N. Akasu & W. Wolf. *Multimodal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing.* Multimedia Tools and Applications, vol. 14, no. 2, pages 137–151, 2001.

[Alattar 97] A.M. Alattar. *Detecting Fade Regions in Uncompressed Video Sequences.* IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, pages 3025–3028, 1997.

[Allen 83] J.F. Allen. *Maintaining Knowledge About Temporal Intervals.* Communications of the ACM, vol. 26, no. 11, novembre 1983.

[Aner 01] A. Aner & J.R. Kender. *Mosaic-Based Clustering of Scene Locations in Videos.* IEEE Workshop on Content-based Access of Image and Video Libraries, decembre, Hawaii, USA 2001.

[ARGOS 06] ARGOS. *Campagne d'Evaluation d'Outils de Surveillance de Contenus Vidéo.* http ://www.irit.fr/argos, 2006.

[Ariki 03] Y. Ariki, M. Kumano & K. Tsukada. *Highlight Scene Extraction in Real Time From Baseball Live Video.* ACM 5th International Workshop on Multimedia Information Retrieval, pages 209–214, 2003.

[Arman 93a] F. Arman, A. Hsu & M.Y. Chiu. *Feature Management for Large Video Databases.* SPIE Storage and Retrieval for Image and Video Databases, vol. 1908, pages 2–12, février 1993.

[Arman 93b] F. Arman, A. Hsu & M.Y. Chiu. *Image Processing on Compressed Data for Large Video Database.* ACM International Conference on Multimedia, pages 267–272, août, Anaheim, USA ? 1993.

[Babaguchi 02] N. Babaguchi, Y. Kawai & T. Kitahashi. *Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration.* IEEE Trans. Multimedia, vol. 4, no. 1, 2002.

[Barnard 03] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D.M. Blei & M.I. Jordan. *Matching Words and Pictures.* J. Mach. Learn. Res., vol. 3, pages 1107–1135, 2003.

[Beauchêne 05a] D. Beauchêne, F. Deloule, B. Ionescu & P. Lambert. *Base de Connaissances Multimédia pour le Cinéma d'Animation.* CIDE.8 Conférence Internationale sur le Document Electronique, vol. CD-Rom, mai Beyrouth, Liban 2005.

[Beauchêne 05b] D. Beauchêne, F. Deloule, B. Ionescu & P. Lambert. *Ontologies et Indexations Vidéo pour les Films d'Animation.* 23ème Congrès INFORSID, MetSI - Atelier Métadonnées et Systèmes d'Information, vol. CD-Rom, mai Grenoble, France 2005.

[Beaver 94] F. Beaver. *Dictionary of Film Terms.* New York : Twayne, 1994.

[Ben-Yacoub 99] S. Ben-Yacoub, B. Fasel & J. Luettin. *Fast Face Detection Using MLP and FFT.* Second International Conference on Audio and Video-Based Biometric Person Authentication, pages 31–36, 1999.

[Benavente 04] Robert Benavente & Maria Vanrell. *Fuzzy Colour Naming Based on Sigmoid Membership Functions.* The Second European Conference on Colour Graphics, Imaging and Vision, pages 135–139, april 2004.

[Benitez 01] A. Benitez, S.-F. Chang & J.R. Smith. *IMKA : A Multimedia Organization System Combining Perceptual and Semantic Knowledge.* ACM Multimedia, vol. 18, pages 121–129, 2001.

[Berlin 91] B. Berlin & P. Kay. *Basic Color Terms : Their Universality and Evolution.* University of California Press, Berkeley 1991.

[Bimbo 99] A. Del Bimbo. *Visual Information Retrieval.* Morgan Kaufmann Publishers, San Francisco, USA 1999.

[Birren 69] F. Birren. *Principles of Color - A Review of Past Traditions and Modern Theories of Color Harmony.* New York : Reinhold, 1969.

[Boreczky 98] J.S. Boreczky & L.D. Wilcox. *A Hidden Markov Model Framework for Video Segmentation Using Audion and Image Features.* IEEE International Conference on Acoustics, Speech, and Signal Processing, mai, Seattle, USA 1998.

[Bouthemy 98] P. Bouthemy & R. Fablet. *Motion Characterization from Temporal Cooccurrences of Local Motion-Based Measures for Video Indexing.* Pattern Recognition, vol. 1, pages 905–908, août 1998.

[Bouthemy 99] P. Bouthemy, M. Gelgon & F. Ganansia. *A Unified Approach to Shot Change Detection and Camera Motion Characterization.* IEEE Transactions on Circuits ans Systems for Video Technology, vol. 9, no. 7, pages 1030–1044, octobre 1999.

[Bruns 00] M.W. Bruns & J.T. Whittlesey. *4 :4 :4 Compression of Moving Pictures for Digital Cinema Using the MPEG-2 Toolkit*. Whitepaper, Grass Valley Group, http ://www.thomsongrassvalley.com /wp/Bruns/D-Cinema_Compression/SMPTE2000-Bruns_wp.pdf, 2000.

[Calic 02] J. Calic & E. Izquierdo. *A Multiresolution Technique for Video Indexing and Retrieval*. IEEE International Conference on Image Processing, vol. 1, pages 952–955, 2002.

[Chang 98] S.-F. Chang, W. Chen, H. Meng, H. Sundara & D. Zhong. *A Fully Automatic Content-Based Video Search Engine Supporting Multi-Object Spatio-Temporal Queries*. IEEE Tran. Circuits Systems Video Technol., vol. 8, pages 602–615, 1998.

[Chang 99] S.-F. Chang, W. Chen, H.J. Meng, H. Sundaram & D. Zhong. *Evaluation of Texture Segmentation Algorithms*. Proc. Conference on Computer Vision and Pattern Recognition, vol. 1, pages 294–299, 1999.

[Chanussot 98] J. Chanussot. *Approches Vectorielles ou Marginales pour le Traitement d'Images Multi-composantes*. Thèse de l'Université de Savoie, Annecy, France, 1998.

[Chen 99] Y. Chen, E.K. Wong, M.M. Yeunh, Y. Boon-Lock & A.C. Charles. *Augmented Image Histogram for Image and Video Similarity Search*. SPIE Conf. Storage and Retrieval for Image and Video Database VII, vol. 3656, pages 523–532, 1999.

[Chen 05] S.-C. Chen, M.-L. Shyu & N. Zhao. *An Enhanced Query Model for Soccer Video Retrieval Using Temporal Relationships*. IEEE Int. Conference on Data Engineering, avril 2005.

[Chen 06] J.-Y. Chen, C. Taskiran, E.J. Delp & C.A. Bouman. *ViBE : A New Paradigme for Video Database Browsing and Search*. http ://stargate.ecn.purdue.edu/ ips/ViBE/, 2006.

[Choi 98] H. Choi & R. Baraniuk. *Multiscale Texture Segmentation using Wavelet-Domain Hidden Markov Models*. Proc. 32nd Asilomar Conf. Signals, Systems and Computers, vol. 2, pages 1692–1697, 1998.

[CICA 06] CICA. *Centre International du Cinema D'Animation*. http :// www.annecy.org, 2006.

[CITIA 06] CITIA. *Cité de l'Image en Mouvement d'Annecy*. http ://www.citia. info, 2006.

[Coldefy 04] F. Coldefy & P. Bouthemy. *Unsupervised Soccer Video Abstraction Based on Pitch, Dominant Color and Camera Motion Analysis*. ACM Multimedia, pages 268–271, Ney York, USA 2004.

[Colombo 99] C. Colombo, A. Del Bimbo & P. Pala. *Semantics in Visual Information Retrieval*. IEEE Multimedia, vol. 6, no. 3, pages 38–53, 1999.

[Condillac 05] Condillac. *Equipe Ingénierie des Connaissances - LISTIC, Université de Savoie*. http ://ontology.univ-savoie.fr/condillac/, 2005.

[Cooper 02] M. Cooper & J. Foote. *Summarizing Video Using Non-Negative Similarity Matrix Factorization*. IEEE Workshop on Multimedia Signal Processing, pages 25–28, St. Thomas, US Virgin Islands 2002.

[Corridoni 99] J.M. Corridoni, A. Del Bimbo & P. Pala. *Retrieval in Paintings Using Effects Induced by Color Features*. IEEE Multimedia, vol. 6, no. 3, pages 38–53, 1999.

[Cox 00] I.J. Cox, M. Miller, T.P. Minka, T.V. Papathomas & P.N. Yianilos. *The Bayesian Image Retrieval System, PicHunter : Theory, Implementation and Psychophysical Experiments*. IEEE Trans. Image Processing, vol. 9, pages 20–37, 2000.

[Dagtas 00] S. Dagtas, W. Al-Khatib, A. Ghafoor & R.L. Kashyap. *Models for Motion-Based Video Indexing and Retrieval*. IEEE Transactions on Image Processing, vol. 9, no. 1, pages 88–101, janvier 2000.

[Detyniecki 03] M. Detyniecki & C. Marsala. *Discovering Knowledge for Better Video Indexing Based on Colors*. IEEE International Conference on Fuzzy Systems, vol. 2, pages 1177–1181, Paris, France 2003.

[Dictionaries 06] Color-Name Dictionaries. *http ://swiss.csail.mit.edu/ jaffer/Color/Dictionaries.html*. 2006.

[Donderler 04] M.E. Donderler, O. Ulusoy & U. Gudukbay. *Rule-Based Spatio-Temporal Query Processing for Video Databases*. VLDB Journal, vol. 13, no. 1, janvier 2004.

[Doulamis 00a] A.D. Doulamis, N. Doulamis & S. Kollias. *Non-Sequential Video Content Representation Using Temporal Variation of Feature Vectors*. IEEE Transactions on Consumer Electronics, vol. 46, no. 3, pages 758–768, 2000.

[Doulamis 00b] A.D. Doulamis, N.D. Doulamis & S.D. Kollias. *A Fuzzy Video Content Representation for Video Summarization and Content-Based Retrieval*. Signal Processing, vol. 80, no. 6, pages 1049–1067, juin 2000.

[Drew 00] M.S. Drew, Z.N. Li & X. Zhong. *Video Dissolve and Wipe Detection Via Spatio-Temporal Images of Chromatic Histograms Differences*. IEEE International Conference on Image Processing, vol. 3, pages 929–932, 2000.

[Duan 06] L.-Y. Duan, J.S. Jin & C.-S. Xu Q. Tian. *Nonparametric Motion Characterization for Robust Classification of Camera Motion Patterns*. IEEE Transactions on Multimedia, vol. 8, no. 2, pages 323–340, avril 2006.

[Dufaux 00] F. Dufaux. *Key Frame Selection to Represent a Video*. IEEE International Conference on Multimedia and Expo, vol. 2, pages 275–278, 2000.

[Eidenberger 04] H. Eidenberger. *A Video Browsing Application Based on Visual MPEG-7 Descriptors and Self-Organising Maps*. TFSA International Journal of Fuzzy Systems, vol. 6, no. 3, pages 122–135, septembre 2004.

[Erol 00] B. Erol & F. Kossentini. *Automatic Key Video Object Plane Selection Using the Shape Information in the MPEG-4 Compressed Domain*. IEEE Transactions on Multimedia, vol. 2, pages 129–138, 2000.

[Erol 03] B. Erol & D.-S.H.J. Lee. *Multimodal Summarization of Meeting Recordings*. IEEE International Conference on Multimedia and Expo, vol. 3, pages 25–28, 2003.

[Evans 03] B.L. Evans, V. Monga & N. Damera-Venkata. *Variations on Error Diffusion : Retrospectives and Future Trends*. SPIE Color Imaging VIII : Processing, Hardcopy, and Applications, vol. 5008, pages 371–389, janvier 2003.

[Fablet 02] R. Fablet, P. Bouthemy & P. Pérez. *Nonparametric Motion Characterization Using Causal Probabilistic Models for Video Indexing and Retrieval*. IEEE Transactions on Image Processing, vol. 11, no. 4, pages 393–407, 2002.

[Fan 01] J. Fan, W.G. Aref, A.K. Elmagamid, M.-S. Hacid, M.S. Marzouk & X. Zhu. *Multi-Level Video Content Representation and Retrieval*. J. Electron. Imaging Special Issue on Multimedia Database, vol. 10, no. 4, pages 895–908, 2001.

[Fan 04] J. Fan, H. Luo & A.K. Elmagarmid. *Concept-Oriented Indexing of Video Databases : Toward Semantic Sensitive Retrieval and Browsing*. IEEE Trans. on Image Processing, vol. 13, no. 7, pages 974–991, 2004.

[Ferman 99] A.M. Ferman & A.M. Tekalp. *Probabilistic Analysis and Extraction of Video Content*. IEEE International Conference on Image Processing, vol. 2, pages 91–95, octobre, Kobe, Japan 1999.

[Fisher 93] N.I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1993.

[Flickner 95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Patkovic, D. Steele & P. Yanker. *Query by Image and Video Content : The QBIC System*. IEEE Computer, vol. 28, no. 9, pages 23–32, september 1995.

[Folimage 06a] Folimage. *l'Equipée de Folimage*. Musée de Valence/Musées d'Annecy, Imprimerie Jalin, 2006.

[Folimage 06b] Studio Folimage. *Présentation du studio et de ses productions*. http ://www.folimage.com, 2006.

[Furht 95] B. Furht, S.W. Smoliar & H. Zhang. *Video and Image Processing in Multimedia Systems*. Norwell, Kluwer 1995.

[Gargi 00] U. Gargi, R. Kasturi & S.H. Strayer. *Performance Characterization of Video-Shot-Change Detections Methods*. IEEE Transactions on Circuits, Systems for Video Technology, vol. 10, no. 1, pages 1–13, février 2000.

[Ge 02] J. Ge & G. Mirchandani. *A New Hybrid Block-Matching Motion Estimation Algorithm*. IEEE ICASSP, vol. 4, page 4190, 2002.

[Gilvarry 99] J. Gilvarry. *Extraction of Motion Vectors from an MPEG Stream*. Rapport technique Dublin City University, http ://www.cdvp.dcu.ie/Papers/MVector.pdf, 1999.

[Gimel'farb 96] G.L. Gimel'farb & A.K. Jain. *On Retrieving Textured Images from an Image Database*. Pattern Recognition, vol. 29, no. 9, pages 1461–1483, 1996.

[Gong 03] Y. Gong & X. Liu. *Video Summarization and Retrieval Using Singular Value Decomposition.* ACM Multimedia Systems Journal, vol. 9, pages 157–168, 2003.

[Guillaume 01] S. Guillaume. *Induction de Règles Floues Interprétables.* Thèse de doctorat INSA Toulouse, France, novembre 2001.

[Guimaraes 03] S.J.F. Guimaraes, M. Couprie, A. de A. Araujo & N.J. Leite. *Video Segmentation Based on 2D Image Analysis.* Pattern Recognition Letters, no. 24, pages 947–957, 2003.

[H. Sundaram 00] S. Chang H. Sundaram. *Determining Computable Scenes in Films and Their Structures using Audio-Visual Memory Models.* ACM Multimedia, pages 95 – 104, Marina del Rey, California, US 2000.

[Haering 00] N. Haering, R. Qian & I. Sezan. *A Semantic Event-Detection Approach and Its Application to Detecting Hunts in Wildlife Video.* IEEE Trans. on Circuits and Systems for Video Technology, vol. 10, no. 6, pages 857–868, 2000.

[Hafner 95] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner & W. Niblack. *Efficient color histogram indexing for quadratic form distance functions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 7, pages 729–736, 1995.

[Han 02] M. Han, W. Hua, W. Xu & Y. Gong. *An Integrated Baseball Digest System Using Maximum Entropy Method.* ACM Multimedia, Juan-les-Pins, France 2002.

[Hanjalic 97] A. Hanjalic, M. Ceccarelli, R.L. Lagendijk & J. Biemond. *Automation of Systems Enabling Search on Stored Video Data.* SPIE Storage and Retrieval for Image and Video Databases V, vol. 3022, pages 427–438, février 1997.

[Hanjalic 02] A. Hanjalic. *Shot-Boundary Detection : Unraveled and Resolved ?* IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 2, pages 90–105, février 2002.

[Hauptmann 98] A.G. Hauptmann & M.J. Witbrock. *Story Segmentation and Detection of Commercials in Broadcast News Video.* Advances in Digital Libraries, pages 168–179, avril, Santa Barbara, USA 1998.

[He 99] L. He, E. Sanocki, A. Gupta & J. Grudin. *Auto-Summmarization of Audio-Video Presentations.* ACM Multimedia, pages 489–498, Orlando, USA 1999.

[Heng 99] W.J. Heng & K.N. Ngan. *Post Shot Boundary Detection Technique : Flashlight Scene Determination.* Signal Processing and Its Applications, vol. 1, pages 447–450, août 1999.

[Heng 01] W.J. Heng & K.N. Ngan. *Enhanced Shot Boundary Refinement for Post-Shot Boundary Detection.* IEEE, pages 259–263, 2001.

[Hsu 02] C.-T. Hsu & S.-J. Teng. *Motion Trajectory Based Video Indexing and Retrieval.* IEEE International Conference on Image Processing, vol. 1, pages 605–608, septembre, New York, USA 2002.

[Huang 98] J. Huang & Y. Wang Z. Liu. *Integration of Audio and Visual Information for Content-Based Video Segmentation.* IEEE International

Conference on Image Processing, vol. 3, pages 526–530, octobre, Chicago, USA 1998.

[Ionescu 03] B. Ionescu. *Contribution à l'Etude d'un Système d'Indexation Vidéo.* Rapport de Stage DEA, LISTIC, ESIA, Université de Savoie, juin, Annecy, France 2003.

[Ionescu 04a] B. Ionescu. *Caractérisation symbolique de Séquence d'Images : Application au Résumé, à la Navigation et à la Recherche.* 2ème et 3ème rapports de doctorat, LAPI, UPB - Université « Politehnica » Bucarest, octobre, Bucarest, Roumanie 2004.

[Ionescu 04b] B. Ionescu. *Caractérisation Symbolique de Séquences d'Images : Application au Résumé, à la Navigation et à la Recherche.* Rapport de stage de doctorat, LISTIC, ESIA, Université de Savoie, juillet, Annecy, France 2004.

[Ionescu 05a] B. Ionescu. *Caractérisation Symbolique des Couleurs dans les Films d'Animation.* Rapport de stage de doctorat, LISTIC, ESIA, Université de Savoie, juillet, Annecy, France 2005.

[Ionescu 05b] B. Ionescu. *Caractérisation Symbolique des Plans Vidéo dans les Films d'Animation.* Rapport de stage de doctorat, LISTIC, ESIA, Université de Savoie, juillet, Annecy, France 2005.

[Ionescu 05c] B. Ionescu. *Caractérisation Symbolique du Mouvement dans les Films d'Animation.* Rapport de stage de doctorat, LISTIC, ESIA, Université de Savoie, juin, Annecy, France 2005.

[Ionescu 05d] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *Analysis and Characterization of Animation Movies.* ORASIS journées francophones des jeunes chercheurs en vision par ordinateur, vol. CD-Rom, mai, Fournois, Puy-de-Dôme, France 2005.

[Ionescu 05e] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *An Approach to Scene Detection in Animation Movies and Its Applications.* Sci. Bull. Revue de l'Université «Politehnica» de Bucarest, vol. Series C, 67, no. 2, pages 45–57, 2005.

[Ionescu 05f] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *The Influence of the Color Reduction on Cut Detection in Animation Movies.* Actes du 20ème Colloque GRETSI sur le Traitement et l'Analyse du Signal et d'Image, vol. CD-Rom, septembre, Louvain-la-Neuve, Belgique 2005.

[Ionescu 05g] B. Ionescu, P. Lambert, D. Coquin & V. Buzuloiu. *Analysis of Animation Movies : Segmentation, Abstraction and Annotation.* Sci. Bull. Revue de l'Université «Politehnica» de Bucarest, vol. Series C, 67, no. 4, pages 3–12, 2005.

[Ionescu 05h] B. Ionescu, P. Lambert, D. Coquin & L. Dîlea. *Color-Based Semantic Characterization of Cartoons.* IEEE International Symposium on Signals, Circuits and Systems, session invitée : Statistical Models in Image Processing, vol. 1, pages 223–226, juillet, Iaśi, Romania 2005.

[Ionescu 06a] B. Ionescu, V. Buzuloiu, P. Lambert & D. Coquin. *Improved Cut Detection for the Segmentation of Animation Movies.* IEEE Internatio-

nal Conference on Acoustic, Speech and Signal Processing, vol. CD-Rom, mai, Toulouse, France 2006.

[Ionescu 06b] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *Semantic Characterization of Animation Movies Based on Fuzzy Action and Color Information*. AMR 4th International Workshop on Adaptive Multimedia Retrieval, vol. CD-Rom, juillet, Genève, Switzerland 2006.

[Ionescu 06c] B. Ionescu, P. Lambert, D. Coquin & V. Buzuloiu. *Fuzzy Color-Based Semantic Characterization of Animation Movies*. IS&T CGIV - 3th European Conference on Colour in Graphics, Imaging, and Vision, juin, Leeds, United Kingdom 2006.

[Ionescu 06d] B. Ionescu, P. Lambert, D. Coquin, L. Ott & V. Buzuloiu. *Animation Movies Trailer Computation*. ACM Multimedia, vol. CD-Rom, octobre, Santa Barbara, CA, USA 2006.

[Ionescu 07a] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *Fuzzy Semantic Action and Color Characterization of Animation Movies in the Video Indexing Task Context*. Springer-Verlag LNCS - Lecture Notes in Computer Science, Eds. S. Marchand-Maillet et al., vol. 4398, pages 119–135, sous presse 2007.

[Ionescu 07b] B. Ionescu, P. Lambert, D. Coquin & V. Buzuloiu. *Caractérisation du Mouvement dans les Films d'Animation*. Sci. Bull. Revue de l'Université «Politehnica» de Bucarest, vol. Series C, 69, no. 2, 2007.

[Ionescu 07c] B. Ionescu, P. Lambert, D. Coquin & V. Buzuloiu. *Color-Based Content Retrieval of Animation Movies A Study*. IEEE Fifth International Workshop on Content-Based Multimedia Indexing, juin, Bordeaux, France, sous presse 2007.

[Irani 95] Michal Irani, P.Anandan & Steve Hsu. *Mosaic Based Representations of Video Sequences and Their Applications*. Computer Vision, pages 605–611, juin 1995.

[Itten 61] J. Itten. *The Art of Color : The Subjective Experience and Objective Rationale of Color*. New York : Reinhold, 1961.

[Jain 91] J.R. Jain & A.K. Jain. *Displacement measurement and its application in interframe image coding*. IEEE Transactions on Communications, vol. 29, no. 12, pages 1799–1806, décembre 1991.

[Jain 99] A.K. Jain, M.N. Murty & P.J. Flynn. *Data Clustering : A Review*. ACM Computing Surveys, vol. 31, no. 3, pages 264–323, septembre 1999.

[Jeannin 01] S. Jeannin & A. Divakaran. *MPEG-7 Visual Motion Descriptors*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pages 720–724, juin 2001.

[Jin 02] R. Jin, Y. Qi & A. Hauptmann. *A Probabilistic Model for Camera Zoom Detection*. The Sixteenth Conference of the International Association for Pattern Recognition, no. 3, pages 859– 862, août, Quebec, Canada 2002.

[J.M.Corridoni 95] J.M.Corridoni & A. Del Bimbo. *Film Semantic Analysis*. Proceedings of Computer Architectures for Machine Perception, pages 202–209, septembre, Como, Italy 1995.

[Kang 01] H.-B. Kang. *A Hierarchical Approach to Scene Segmentation*. IEEE, 2001.

[Kanjanawanishkul 05] K. Kanjanawanishkul & B. Uyyanonvara. *Novel Fast Color Reduction Algorithm for Time-Constrained Applications*. Journal of Visual Communication and Image Representation, vol. 16, no. 3, pages 311–332, juin 2005.

[Katsavounidis 97] I. Katsavounidis & C.-C.J. Kuo. *A Multiscale Error Diffusion Technique for Digital Halftoning*. IEEE Transactions on Image Processing, vol. 6, no. 3, pages 483–490, 1997.

[Kaufman 90] L. Kaufman & P.J. Rousseeuw. *Finding Groups in Data : An Introduction to Cluster Analysis*. Wiley, 1990.

[Kay 03] P. Kay & T. Regier. *Resolving the Question of Color Naming Universals*. National Academy of Sciences, vol. 100, no. 15, 2003.

[Kelly 76] K.L. Kelly & D.B. Judd. *Color : Universal Language and Dictionary of Names*. National Bureau of Standards, 1976.

[Kim 00a] C. Kim & J.N. Hwang. *An Integrated Scheme for Object-Based Video Abstraction*. ACM Multimedia, pages 303 – 311, 2000.

[Kim 00b] E.Y. Kim, K.I. Kim, K. Jung & H.J. Kim. *A Video Indexing System Using Character Recognition*. IEEE, pages 358–359, Los Angles, CA, USA 2000.

[Kim 02] S.H. Kim & R.H. Park. *Robust Video Indexing for Video Sequences with Complex Brightness Variation*. IASTED International Conference on Signal and Image Processing, pages 410–414, Kauai, Hawaii 2002.

[Kim 04] J.-G. Kim, H.S. Chang, J. Kim & H.-M. Kim. *Threshold-Based Camera Motion Characterization of MPEG Video*. ETRI Journal, vol. 26, no. 3, pages 269–272, juin 2004.

[Klir 95] G. J. Klir & B. Yuan. *Fuzzy Sets and Fuzzy Logic : Theory and Applications*. Prentice Hall, New Jersey, 1995.

[Kobla 99] V. Kobla, D. DeMenthon & D. Doermann. *Special Effect Edit Detection Using VideoTrails : A Comparison with Existing Techniques*. SPIE Storage and Retrieval for Image and Video Databases VII, vol. 3656, pages 302–313, 1999.

[Kobla 00] V. Kobla, D. DeMenthon & D. Doermann. *Identifying Sports Video Using Replay, Text and Camera Motion Features*. SPIE Storage and Retrieval for Media Database, vol. 3972, pages 332–343, 2000.

[Kosch 01] H. Kosch, L. Boszormányi, A. Bachlechner, B. Dorflingera, C. Hanin, C. Hofbauer, M. Lang, C. Riedler & R. Tusch. *SMOOTH - A Distributed Multimedia Database System*. VLDB 2001, Rome, Italy 2001.

[Kramer 05] P. Kramer & J.B. Pineau. *Camera Motion Detection in the Rough Index Paradigm*. TREC Video Retrieval Evaluation Online Proceedings, TRECVID, novembre 2005.

[Lab. 05] Compaq Corporate Res. Lab. *Audio Search Using Speech Recognition*. http ://speechbot.research.compaq.com, 2005.

[Laboratoires 05] Mitsubishi Electric Research Laboratoires. *MERL - Timetunnel Interface for Video Browsing.* http ://www.merl.com/projects/timetunnel2/, 2005.

[Lambert 07] P. Lambert, B. Ionescu & D. Coquin. *La couleur dans les séquences d'images.* EHINC - Ecole d'Hiver sur l'Image Numérique Couleur, janvier, Poitiers, France 2007.

[Lay 04] J.A. Lay & L. Guan. *Retrieval for Color Artistry Concepts.* IEEE Transactions on Image Processing, vol. 13, no. 3, pages 125–129, mars 2004.

[Lee 01] M.-S. Lee, Y.-M. Yang & S.-W. Lee. *Automatic Video Parsing Using Shot Boundary Detection and Camera Operation Analysis.* Pattern Recognition, vol. 34, pages 711–719, 2001.

[Lee 02] S. Lee & M.H. Hayes. *Real-Time Camera Motion Classification for Content-Based Indexing and Retrieval using Templates.* IEEE, vol. 4, pages 3664–3667, 2002.

[Lescieux 06] M. Lescieux. *Introduction à la Logique Floue : Plan du Cours.* http ://auto.polytech.univ-tours.fr/automatique/AUA/ressources/Introduction_logique_floue.ppt, 2006.

[Li 01] Y. Li, T. Zhang & D. Tretter. *An Overview of Video Abstraction Techniques.* HP Laboratories, HPL-2001-191, 2001.

[Li 03] Y. Li, S. Narayanan & C.-C.J. Kuo. *Movie Content Analysis, Indexing and Skimming via Multimodal Information.* Video Mining, Chapter 5, Eds. Kluwer Academic Publishers, 2003.

[Lienhart 97] R. Lienhart, C. Kuhmunch & W. Effelsberg. *On the Detection and Recognition of Television Commercials.* IEEE Conf. on Multimedia Computing and Systems, pages 509–516, Ottawa, Ont., Canada 1997.

[Lienhart 99a] R. Lienhart. *Comparison of Automatic Shot Boundary Detection Algorithms.* SPIE Storage and Retrieval for Still Image and Video Databases VII, vol. 3656, pages 290–301, 1999.

[Lienhart 99b] R. Lienhart, S. Pfeiffer & W. Effelsberg. *Scene Determination Based on Video and Audio Features.* IEEE International Conference on Multimedia, Computing and Systems, vol. 1, pages 685–690, juin, Florence, Italy 1999.

[Lienhart 00] R. Lienhart. *Dynamic Video Summarization of Home Video.* SPIE Storage and Retrieval for Media Databases, vol. 3972, pages 378–389, janvier 2000.

[Lienhart 01a] R. Lienhart. *Reliable Dissolve Detection.* SPIE Storage and Retrieval for Media Databases, vol. 4315, pages 219–230, janvier 2001.

[Lienhart 01b] R. Lienhart. *Reliable Transition Detection in Videos : A Survey and Practitiner's Guide.* MRL, Intel Corporation, http ://www.lienhart.de/Publications/IJIG_AUG2001.pdf, août, Santa Clara, USA 2001.

[Lin 98] C.-W. Lin, Y.-J. Chang & Y.-C. Chen. *Hierarchical Motion Estimation Algorithm Based on Pyramidal Successive Elimination.* International Computer Symposium, octobre 1998.

[Lin 02] W.-H. Lin & A.G. Hauptmann. *News Video Classification Using SVM-Based Multimodal Classifiers and Combination Strategies*. ACM Multimedia, pages 323–326, Juan-les-Pins, France 2002.

[Liu 02a] C.-C. Liu & A.L.P. Chen. *3D-List : A Data Structure for Efficient Video Query Processing*. IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 1, pages 106–122, janvier-février 2002.

[Liu 02b] T. Liu & J.R. Kender. *Optimization Algorithms for the Selection of Key Frames Sequences of Variable Length*. European Conference on Computer Vision, vol. 2353, pages 403–417, London, UK 2002.

[Liu 03] T. Liu, H.-J. Zhang & F. Qi. *A Novel Video Key-Frame Extraction Algorithm Based on Perceived Motion Energy Model*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 10, pages 1006–1013, octobre 2003.

[Liu 04] T. Liu, X. Zhang, J. Freg & K. Lo. *Shot Reconstruction Degree : A Novel Criterion for Keyframe Selection*. Pattern Recognition Letter, vol. 25, no. 12, pages 1451–1457, septembre 2004.

[Lu 03] S. Lu, I. King & M. Lyu. *Video Summarization Using Greedy Method in a Constraint Satisfaction Framework*. 9th International Conference on Distributed Multimedia Systems, pages 456–461, Miami, Florida, USA 2003.

[Lupatini 98] G. Lupatini, C. Saraceno & R. Leonardi. *Scene Break Detection : A Comparison*. Research Issues in Data Engineering, Workshop on Continuous Media Databases and Applications, pages 34–41, Orlando, FL, USA 1998.

[Ma 01] Y.F. Ma, J. Sheng, Y. Chen & H.J. Zhang. *MSR-Asia at TREC-10 Video Track : Shot Boundary Detection Task*. 10th Text Retrieval Conference, page 371, 2001.

[Maillet 03] S.M. Maillet. *Content-Based Video Retrieval : An Overview*. http ://viper.unige.ch/ marchand/CBVR/, 2003.

[Marichal 98] X. Marichal. *Motion Estimation and Compensation for Very Low Bitrate Video Coding*. These, UCL - Université Catholique de Louvain, Laboratoire de Telecommunications et Teledetection, Louvain-la-Neuve, Belgique 1998.

[Mazière 00] M. Mazière, F. Chassaing, L. Garrido & P. Salembier. *Segmentation and Tracking of Video Objects for a Content-Based Video Indexing Context*. IEEE International Conference on Multimedia Computing and Systems, vol. 2, pages 1191–1194, New York, USA 2000.

[Mehtre 97] B.M. Mehtre, M.S. Kankanhalli & W.F. Lee. *Shape Measures for Content Based Image Retrieval : A Comparison*. Information Processing and Management, vol. 33, no. 3, pages 319–337, mai, 1997.

[Meng 95] J. Meng, Y. Juan & S.F. Chang. *Scene Change Detection in a MPEG Compressed Video Sequence*. SPIE, vol. 2419, pages 14–25, février 1995.

[Miene 01] A. Miene, A. Dammeyer, T. Hermes & O. Herzog. *Advanced and Adaptive Shot Boundary Detection*. ECDL WS Generalized Documents, pages 39–43, 2001.

[Miura 03] K. Miura, R. Hamada, I. Ide, S. Sakai & H. Tanaka. *Motion Based Automatic Abstraction of Cooking Videos.* IPSJ Transactions on Computer Vision and Image Media, vol. 44, 2003.

[Mojsilovic 00] A. Mojsilovic, J. Kovacevic, R.J. Safranek J. Hu & S.K. Ganapathy. *Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns.* IEEE Transactions on Image Processing, vol. 9, no. 1, pages 38–54, 2000.

[Nagasaka 92] A. Nagasaka & Y. Tanaka. *Automatic Video Indexing and Full-Video Search for Object Appearances.* Visual Database Systems II, pages 113–127, Amsterdam, Netherlands 1992.

[Nam 97] J. Nam, A.E. Cetin & A.H. Tewfik. *Speaker Identification and Video Analysis for Hierarchical Video Shot Classification.* IEEE Int. Conference Image Processing, octobre, Santa Barbara, USA 1997.

[Nam 98] J. Nam, M. Alghoniemy & A.H. Tewfik. *Audio-Visual Content-Based Violent Scene Characterization.* IEEE Int. Conference on Image Processing, vol. 1, pages 353–357, Chicago, USA 1998.

[Naphade 01a] M.R. Naphade. *A Probabilistic Framework for Mapping Audio-Visual Features to High-Level Semantics in Terms of Concepts and Context.* Ph.D. dissertation, Dept.Elect.Comput.Eng. Univ. Illinois, 2001.

[Naphade 01b] M.R. Naphade & T.S. Huang. *A Probabilistic Framework for Semantic Video Indexing, Filtering and Retrieval.* IEEE Trans. Multimedia, vol. 3, no. 1, pages 141–151, 2001.

[Naphade 02] M.R. Naphade & T.S. Huang. *Extractiong Semantics from Audio-visual Content : The Final Frontier in Multimedia Retrieval.* IEEE Tran. on Neural Networks, vol. 13, no. 2, pages 793–810, 2002.

[Ngo 00] C.-W. Ngo, T.-C. Pong, H.-J. Zhang & R.T. Chin. *Motion Characterization by Temporal Slices Analysis.* IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, pages 768–773, 2000.

[Ngo 03] C.-W. Ngo, Y.-F. Ma & H.-J. Zhang. *Automatic Video Summarization by Graph Modeling.* IEEE International Conference on Computer Vision, vol. 1, page 104, Nice, France 2003.

[Otsuji 91] K. Otsuji, Y. Tonomura & Y. Ohba. *Video Browsing Using Brightness Data.* SPIE VCIP, vol. 1606, pages 980–989, 1991.

[Ott 05] L. Ott. *Résumé Automatique de Films d'Animation.* Rapport de fin d'étude, LISTIC, ESIA, juin, Annecy, France 2005.

[Pan 01] H. Pan, P. Beek & M. Sezan. *Detection of Slow-Motion Replay Segments in Sports Video for Highlights Generation.* IEEE ICASSP, vol. 3, pages 1649–1652, Salt Lake City, Utah, SUA 2001.

[Pfeiffer 96] S. Pfeiffer, R. Lienhart, S. Fisher & W. Effelsberg. *Abstracting Digital Movies Automatically.* Journal of Visual Communication and Image Representation, vol. 7, no. 4, decembre 1996.

[Pilu 97] M. Pilu. *On Using Raw MPEG Motion Vectors To Determine Global Camera Motion.* HP - Hewlett Packard,

http ://www.hpl.hp.com/techreports/97/HPL-97-102.pdf, août 1997.

[Pineau 05] J.B. Pineau. *Extraction des Objets Couleur en Mouvement des Séquences Vidéo.* LABRI UMR CNRS 5800, www.labri.fr/ImageetSon/AIV, 2005.

[Porter 00] S.V. Porter, M. Mirmehdi & B.T. Thomas. *Video Cut Detection Using Frequency Domain Correlation.* 15th International Conference on Pattern Recognition, vol. 3, pages 413–416, Barcelona, Spain 2000.

[Porter 01] S.V. Porter, M. Mirmehdi & B.T. Thomas. *Detection and Classification of Shot Transitions.* 12th British Machine Vision Conference, pages 73–82, septembre 2001.

[Production 06] AAA Production. *Studio AAA (Animation Art Graphique Audiovisuel).* http ://www.aaaproduction.fr/index.php, 2006.

[QBIC] IBM QBIC. *Hermitage Museum - Query by Image Content.* http ://www.hermitage-museum.org.

[Qian 99] R. Qian, N. Haering & I. Sezan. *A Computational Approach to Semantic Event Detection.* Computer Vision and Pattern Recognition, vol. 1, page 206, juin 1999.

[Radhakrishnan 04] R. Radhakrishnan, A. Divakaran & Z. Xiong. *A Time Series Clustering Based Framework for Multimedia Mining and Summarization Using Audio Features.* ACM International Workshop on Multimedia Information Retrieval, pages 157–164, New York, USA 2004.

[Ren 03] W. Ren & S. Singh. *Video Transition : Modeling and Prediction.* Pattern Analysis and Neural Networks, PANN, http ://www.dcs.ex.ac.uk/research/pann/pdf/pann_SS_089.PDF, 2003.

[Reoxiang 94] L. Reoxiang, Z. Bing & M.L. Liou. *A new three-step search algorithm for block motion estimation.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 4, no. 4, pages 438–442, août 1994.

[Rivlin 95] E. Rivlin & I. Weiss. *Local Invariants for Recognition.* IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 3, pages 226–238, 1995.

[Robert 88] Dictionaire Le Petit Robert. vol. 1, 1988.

[Rowe 01] L.A. Rowe, D. Harley, P. Pletcher & S. Lawrence. *BIBS : A Lecture Webcasting System.* Berkley Multimedia Research Center, TR 2001-160, juin 2001.

[Rubner 97] Y. Rubner, L.J. Guibas & C. Tomasi. *The Earth's Mover Distance Multi-Dimensional Scaling and Color-Based Image Retrieval.* ARPA Image Understanding Workshop, 1997.

[S. Vogl 99] M. Muhlhauser S. Vogl K.Manske. *A VRML Approach to Web Video Browsing.* Multimedia Computing and Networking, pages 276–285, San Jose, USA 1999.

[Saraceno 98] C. Saraceno & R. Leonardi. *Identification of Story Units in AV Sequencies by Joint Audio and Video Processing.* IEEE International Conference on Image Processing, vol. 1, pages 363–367, octobre, Chicago, USA 1998.

[Saur 97] D.D. Saur, Y.P. Tan, S.R. Kulkarni & P.J. Ramadge. *Automated Analysis and Annotation of Basketball Video.* SPIE Symp., vol. 3022, pages 176–187, 1997.

[Scaringella 06] N. Scaringella, G. Zoia & D. Mlynek. *Automatic Genre Classification of Music Content.* IEEE Signal Processing Magazine, vol. 23, no. 2, pages 133–141, mars 2006.

[Schmid 97] C. Schmid & R. Mohr. *Local Grayvalue Invariants for Image Retrieval.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 5, pages 530–535, 1997.

[Schneiderman 00] H. Schneiderman & T. Kanade. *A Statistical Method for 3D Object Detection Applied to Faces and Cars.* IEEE Computer Vision and Pattern Recognition, vol. 1, pages 746–751, Hilton Head Island, SC, USA 2000.

[Seber 84] G.A.F. Seber. *Multivariate Observations.* Wiley, New York, 1984.

[Shahraray 95] B. Shahraray. *Scene Change Detection and Content-Based Sampling of Video Sequences.* SPIE, vol. 2419, pages 2–13, février 1995.

[Shen 97] B. Shen. *HDH Based Compressed Video Cut Detection.* Visual 97, pages 149–156, décembre, San Diego, USA 1997.

[Smeulders 00] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta & R. Jain. *Content-Based Image Retrieval at the End of the Early Years.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pages 1349–1380, décembre 2000.

[Smith 98] M.A. Smith & T. Kanade. *Video Skimming and Characterization Through the Combination of Image and Language Understanding.* International Workshop on Content-Based Access of Image and Video Databases, Bombay, India 1998.

[Smith 04] P. Smith, T. Drummond & R. Cipolla. *Layered Motion Segmentation and Depth Ordering by Tracking Edges.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 4, pages 479–492, avril 2004.

[Smith 99] J.R. Smith. *VideoZoom Spatial-Temporal Video Browsing.* IEEE Trans. Multimedia, vol. 1, juin 99.

[Snoek 05a] C.G.M. Snoek & M. Worring. *Multimedia Event Based Video Indexing Using Time Intervals.* IEEE Transactions on Multimedia, vol. 7, no. 4, août 2005.

[Snoek 05b] C.G.M. Snoek & M. Worring. *Multimodal Video Indexing : A Review of the State-of-the-art.* Multimedia Tool and Applications, vol. 25, no. 1, pages 5–35, 2005.

[Song 02] H.S. Song, I.K. Kim & N.I. Cho. *Scene Change Detection by Feature Extraction from Strong Edge Blocks.* SPIE Visual Communications and Image Processing, vol. 4671, pages 784–792, 2002.

[Sun 00] X.D. Sun & M.S. Kankanhalli. *Video Summarization Using R-Sequences.* Real-Time Imaging, vol. 6, pages 449–459, décembre, 2000.

[Sundaram 02] H. Sundaram & S.-F. Chang. *Video Skims : Taxonomies and an Optimal Generation Framework.* IEEE International Conference on Image Processing, vol. 2, pages 21–24, Rochester, USA 2002.

[Taniguchi 95] Y. Taniguchi, A. Akutsu, Y. Tonomura & H. Hamada. *An Intuitive and Efficient Access Interface to Real-Time Incoming Video Based on Automatic Indexing.* ACM Multimedia, pages 25–33, San Francisco, California, United States 1995.

[Tardini 05] G. Tardini, C. Grana, R. Marchi & R. Cucchiara. *Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos.* 13th International Conference on Image Analysis and Processing, pages 653–660, septembre 2005.

[Tong 01] S. Tong & E. Chang. *Support Vector Machine Active Learning for Image Retrieval.* ACM Multimedia Conf., vol. 9, Ottawa, Canada 2001.

[Trémeau 04] A. Trémeau, C. Fernandez-Maloigne & P. Bonton. *Image Numérique Couleur : De l'Acquisition au Traitement.* DUNOD ISBN 2 10 006843 1, 2004.

[Truong 00a] B.T. Truong & C. Dorai. *Automatic Genre Identification for Content-Based Video Categorization.* IEEE International Conference on Pattern Recognition, vol. 4, pages 230–233, Barcelona, Spain 2000.

[Truong 00b] B.T. Truong, C. Dorai & S. Venkatesh. *New Enhancements to Cut, Fade, and Dissolve Detection Processes in Video Segmentation.* ACM Multimedia, pages 219–227, novembre 2000.

[Truong 01] B.T. Truong & S. Venkatesh. *Determining Dramatic Intensification via Flashing Lights in Movies.* IEEE Int. Conference on Multimedia and Expo, pages 61–64, Tokyo, Japan 2001.

[Truong 06] B.T. Truong & S. Venkatesh. *Video Abstraction : A Systematic Review and Classification.* accepted for ACM Transactions on Multimedia Computing, Communications and Applications, vol. 3, no. 1, 2006.

[Turaga 98] D. Turaga & M. Alkanhal. *Search Algorithms for Block-Matching in Motion Estimation.* http ://www.ece.cmu.edu/∼ee899/project/deepak_mid.htm, 1998.

[uniFrance 06] uniFrance. *uniFrance - Films.* http ://www.unifrance.org/films/, 2006.

[University 06] Kyungpook National University. *Artificial Intelligence Laboratory.* http ://ailab.kyungpook.ac.kr/vindex/video-view.html, 2006.

[van Houten 03] Y. van Houten, M. van Setten & J.-G. Schuurman. *Patch-Based Video Browsing.* Human-Computer Interaction INTERACT, 2003.

[Vasconcelos 00] N. Vasconcelos & A. Lippman. *Statistical Models of Video Structure for Content Analysis and Characterization.* IEEE Transactions on Image Processing, vol. 9, pages 3–19, janvier 2000.

[Vendriga 01] J. Vendriga & M. Worring. *Evaluation of Logical Story Unit Segmentation in Video Sequences.* IEEE Int. Conference on Multimedia and Expo, pages 1092–1095, Tokyo, Japan 2001.

[Vermaak 02] J. Vermaak, P. Prez, M. Gangnet & A. Blake. *Rapid Summarization and Browsing of Video Sequences.* British Machine Vision Conference, vol. 1, pages 424–433, Cardiff, UK 2002.

[Visibone 06] Visibone. *Webmaster Palette.* http ://www.visibone.com/colorlab, 2006.

[W.A.C.Fernando 99] W.A.C.Fernando, C.N.Canagarajah & D.R.Bull. *Fade and Dissolve Detection in Uncompressed and Compressed Video Sequence.* IEEE International Conference on Image Processing, vol. 3, pages 299–303, octobre, Kobe, Japan 1999.

[W.A.C.Fernando 01] W.A.C.Fernando, C.N.Canagarajah & D.R.Bull. *Scene Change Detection Algorithms for Content-Based Video Indexing and Retrieval.* IEE Electronics and Communication Engineering Journal, pages 117–126, juin 2001.

[Wang 92] L.-X. Wang. *Fuzzy Systems are Universal Approximators.* IEEE Conference on Fuzzy Systems, pages 1163–1170, San Diego, USA 1992.

[Wang 00] Y. Wang, Z. Liu & J.-C. Huang. *Multimedia Content Analysis Using Both Audio and Visual Clues.* IEEE Signal Processing Magazine, vol. 17, no. 6, pages 12–36, novembre 2000.

[Wang 01] J.Z. Wang, J. Li & G. Wiederhold. *SIMPLIcity : Semantic-Sensitive Integrated Matching for Picture Libraries.* IEEE Trans. Pattern Anal. Machine Intell., vol. 23, no. 9, pages 947–963, 2001.

[Wolf 96] W. Wolf. *Key Frame Selection by Motion Analysis.* ICASSP, vol. 2, pages 1228–1231, 1996.

[Xiong 98] W. Xiong & J.C.-M. Lee. *Efficient Scene Change Detection and Camera Motion Annotation for Video Classification.* Computer Vision and Image Understanding, vol. 71, 1998.

[Xiong 03] Z. Xiong, R. Radhakrishnan & A. Divakaran. *Generation of Sports Highlights Using Motion Activity in Combination With a Common Audio Feature Extraction Framework.* IEEE International Conference on Image Processing, vol. 1, Barcelona, Spain 2003.

[Yao 01] A. Yao & J. Jin. *The Developing of a Video Metadata Authoring and Browsing System in XML.* ACM International Conference - Pan-Sydney Workshop on Visual Information Processing, vol. 2, Sydney, Australia 2001.

[Yeo 95] B.-L. Yeo & B. Liu. *Rapid Scene Analysis on Compressed Video.* IEEE Transactions on Circuits, Systems and Video Technology, vol. 5, pages 533–544, decembre 1995.

[Yu 97] H. Yu, G. Bozdagi & S. Harrington. *Feature-Based Hierarchical Video Segmentation.* IEEE International Conference on Image Processing, vol. 2, pages 498–501, 1997.

[Yu 03] B. Yu, W.-Y. Ma, K. Nahrstedt & H.-J. Zhang. *Video Summarization Based on User Log Enhanced Link Analysis.* ACM Multimedia, pages 382–391, Berkeley, USA 2003.

[Yu 04] X.-D. Yu, L. Wang, Q. Tian & P. Xue. *Multi-level Video Representation With Application to Keyframe Extraction.* International Conference on Multimedia Modeling, pages 117–121, Brisban, Australia 2004.

[Zabih 95] R. Zabih, J. Miller & K. Mai. *A Feature-Based Algorithm for Detecting and Classifying Scene Breaks.* ACM Multimedia, pages 189–200, novembre, San Francesco, USA 1995.

[Zabih 99] R. Zabih, J. Miller & K. Mai. *A Feature-Based Algorithm for Detecting and Classification Production Effects.* Multimedia Systems, vol. 7, pages 119–128, 1999.

[Zadeh 65] L.A. Zadeh. *Fuzzy Sets.* Information and Control, vol. 8, no. 3, pages 338–353, 1965.

[Zahariadis 96] T. Zahariadis & D. Kalivas. *A Spiral Search Algorithm For Fast Estimation Of Block Motion Vectors.* Signal Processing VIII, theories and applications. Proceedings of the EUSIPCO 96. Eighth European Signal Processing Conference, vol. 2, pages 1079–1082, 1996.

[Zeng 02] W. Zeng, W. Gao & D. Zhao. *Video Indexing by Motion Activity Maps.* IEEE Int. Conference Image Processing, vol. 1, pages 912–915, 2002.

[Zhang 93] H. Zhang, A. Kankanhalli & S.W. Smoliar. *Automatic Partitioning of Full-Motion Video.* Multimedia Systems, vol. 1, no. 1, pages 10–28, 1993.

[Zhang 94] H. Zhang, C.Y. Low, Y. Gong & S.W. Smoliar. *Video Parsing Using Compressed Data.* SPIE Image and Video Processing II, vol. 2182, pages 142–149, février 1994.

[Zhang 97] H.J. Zhang, J. Wu, D. Zhong & S.W. Smoliar. *An Integrated System for Content-Based Video Retrieval and Browsing.* Pattern Recognition, vol. 30, no. 4, pages 643–658, 1997.

[Zhao 03] M. Zhao, J. Bu & C. Chen. *Audio and Video Combined for Home Video Abstraction.* IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 5, pages 620–623, Hong Kong, China 2003.

[Zhong 96] D. Zhong, H. Zhang & C. Chang. *Clustering Methods for Video Browsing and Annotation.* SPIE Storage and Retrieval for Still Image and Video Databases IV, vol. 2670, pages 239–246, 1996.

[Zhong 97] D. Zhong & S.-F. Chang. *Spatio-temporal Video Search using the Object-Based Video Representation.* IEEE Int. Conf. Image Processing, vol. 1, pages 1–12, 1997.

[Zhou 02] W. Zhou, S. Dao & C.-C.J. Kuo. *On-Line Knowledge and Rule-Based Video Classification System for Video Indexing and Dissemination.* Information Systems, vol. 27, no. 8, 2002.

[Zhu 05] C.-Z. Zhu, T. Mei & X.-S. Hua. *Video Booklet - Natural Video Browsing.* ACM Multimedia, pages 265 – 266, novembre, Singapore 2005.

[Zhuang 98] Y. Zhuang, Y. Rui, T.S. Huang & S. Mehrota. *Adaptive Key Frame Extraction Using Unsupervized Clustering.* IEEE International Conference on Image Processing, vol. 1, pages 866–870, 1998.

Sixième partie

Annexes

Annexe A

La diffusion d'erreur

L'algorithme de réduction des couleurs utilisant la diffusion d'erreur dans l'espace XYZ fonctionne de la manière suivante.

Les pixels sont analysés par un balayage de l'image de haut en bas et de droite à gauche. La couleur du pixel courant, c, est changée en la couleur c_{min} de la palette "Webmaster" (voir la Section 4.2.3) la plus proche au sens de la distance Euclidienne.

Ensuite, l'erreur d'approximation, c'est à dire l'écart entre c et c_{min}, est calculée dans l'espace XYZ par :

$$E^i = |c^i - c^i_{min}| \tag{A.1}$$

où c^i est la valeur de la composante i de la couleur c, $i \in \{X, Y, Z\}$. Les erreurs obtenues, E^X, E^Y et E^Z, sont diffusées en utilisant le masque de filtrage de Floyd et Stenberg, défini ci-dessous, et appliqué sur chaque composante XYZ [Katsavounidis 97].

$$\frac{1}{16} \cdot \begin{bmatrix} 0 & 0 & 0 \\ 0 & -16 & 7 \\ 3 & 5 & 1 \end{bmatrix} \tag{A.2}$$

Les couleurs des pixels voisins du pixel courant sont changées en fonction des coefficients du masque.

Par exemple, si c_S est la couleur du voisin Sud, les trois nouvelles valeurs des composantes X, Y et Z sont données par la relation :

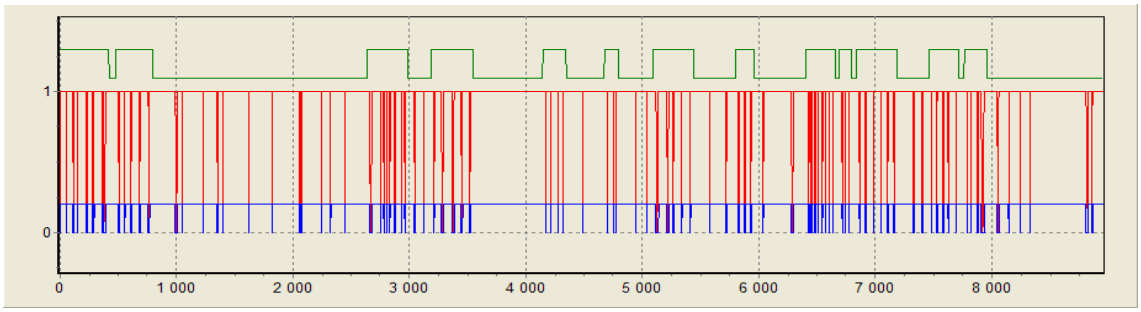
$$c^i_S = c^i_S + \frac{5}{16} \cdot E^i, \quad i \in \{X, Y, Z\} \tag{A.3}$$

Il faut noter que le sens dans lequel le traitement est effectué assure, en fin de traitement, de n'obtenir que des couleurs de la palette choisie, même si provisoirement d'autres couleurs sont utilisées dans l'algorithme.

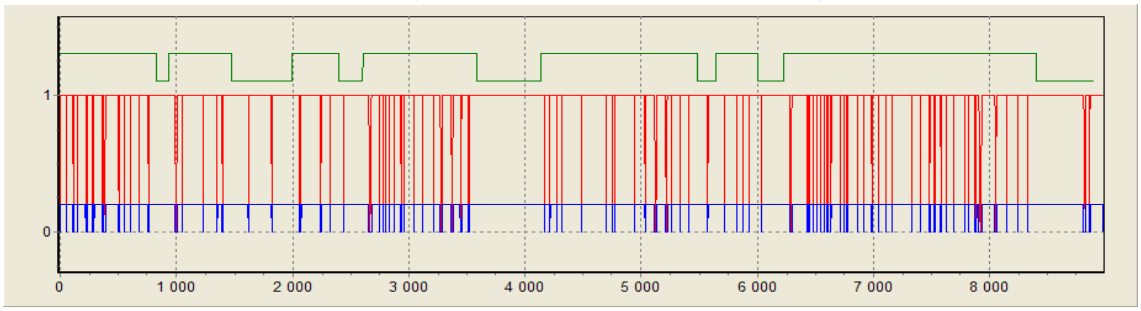
Annexe B

Les segments d'action : choix de T

On peut remarquer que pour $T = 1s$ le nombre de segments d'action (marqués par la ligne verte) est trop important, tandis que pour $T = 10s$ il est trop faible. Pour ces deux valeurs de T, le nombre et la durée des segments d'action ne correspondent pas à la réalité. Un bon compromis est de choisir $T = 5s$ (voir la Figure B.1 et la Figure B.2).

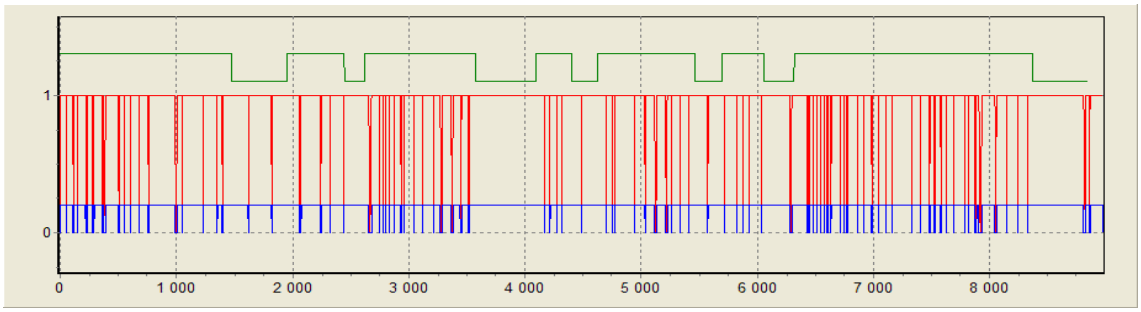


$T = 1s$ (de nombreux segments d'action)

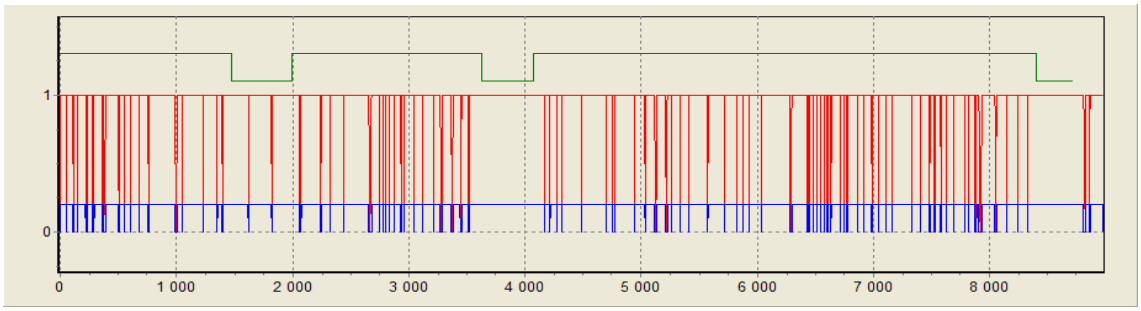


$T = 3s$

Fig. B.1 – Les segments d'action obtenus pour différentes valeurs de T pour le film "Le Moine et le Poisson" [Folimage 06b]. Les segments d'action sont marqués par la ligne verte.



$T = 5s$



$T = 10s$ (segments d'action trop longs)

FIG. B.2 – Les segments d'action obtenus pour différentes valeurs de T pour le film "Le Moine et le Poisson" [Folimage 06b] (suite).

Annexe C

L'estimation du mouvement par blocs de pixels

Le principe de l'estimation du mouvement par blocs s'appuie sur la recherche de la nouvelle position d'un bloc de pixels de l'image courante dans l'image suivante, en utilisant comme critère la minimisation d'une fonction de coût, $F_c()$.

Dans un premier temps l'image courante analysée à l'instant t, donc $I(t)$, est divisée en blocs de $B \times B$ pixels, sans recouvrement. Pour chaque bloc $I(\vec{r}, t)$, positionné dans l'image $I(t)$ aux coordonnées $\vec{r} = (x', y')$, on cherche sa nouvelle position dans l'image suivante, à l'instant $t+l$, donc $I(t+l)$, avec l une valeur entière représentant le pas d'analyse (habituellement $l = 1$).

La recherche se fait dans une fenêtre d'exploration limitée, S, typiquement de taille de $(2B+1) \times (2B+1)$ pixels. La fenêtre S est centrée au milieu du bloc $I(\vec{r}, t+l)$ dans l'image $I(t+l)$, donc aux coordonnées $(x'+\frac{B}{2}, y'+\frac{B}{2})$. La nouvelle position du bloc courant est calculée par la minimisation d'une fonction de coût, $F_c()$, qui exprime l'erreur de l'approximation du bloc courant $I(\vec{r}, t)$ par un des blocs de la fenêtre de recherche S de l'image suivante $I(t+l)$.

Le vecteur de mouvement pour le bloc $I(\vec{r}, t)$ est donc donné par la relation :

$$\vec{d} = argmin_{\vec{d} \in S} F_c(I(\vec{r} - \vec{d}, t+l), I(\vec{r}, t)) \tag{C.1}$$

où $I(\vec{r}, t)$ est le bloc courant analysé, de taille $B \times B$ et à la position $\vec{r} = (x', y')$ dans l'image I, S est la fenêtre de recherche et $\vec{d}$ est le déplacement du bloc courant compris dans S pour lequel la fonction de coût est minimale. Les valeurs de $\vec{d}$ sont toutes les valeurs de déplacements pour lesquels le bloc de comparaison est à l'intérieur de S.

Les fonctions de coût, $F_c()$, les plus fréquemment utilisées sont : *l'erreur quadratique moyenne* (MSE - Mean Square Error), *la valeur absolue de l'erreur moyenne* (MAE - Mean Absolute Error), *l'inter-corrélation*, *la projection intégrale* ou la classification des différences entre les pixels (PDC - *pixel difference classification*) [Turaga 98]. La valeur absolue de l'erreur moyenne, que nous avons utilisée, est définie par l'équation suivante :

$$MAE(I(\vec{r} - \vec{d}, t+l), I(\vec{r}, t)) = \frac{1}{B \cdot B} \cdot \sum_{x=1}^{B} \sum_{y=1}^{B} |I((x' - x_d + x, y' - y_d + y), t+l) - I((x' + x, y' + y), t)| \tag{C.2}$$

où $\vec{r} = (x', y')$, $\vec{d} = (x_d, y_d)$ et $B \times B$ est la taille d'un bloc de pixels.

La complexité de calcul d'une méthode par blocs dépend des facteurs suivants : *l'algorithme de recherche*, *la fonction de coût* et *la dimension de la fenêtre de recherche*. Différentes solutions ont été proposées pour réduire la complexité du calcul, tout en gardant une bonne qualité du champ vectoriel de mouvement obtenu. Les différentes méthodes d'estimation du mouvement par blocs se différencient essentiellement par l'algorithme de recherche utilisé.

On peut ainsi distinguer (voir [Turaga 98]) :

- **la recherche complète** : la minimisation de la fonction de coût, $F_c()$, est effectuée pour tous les blocs de pixels à l'intérieur de la fenêtre de recherche S. La recherche complète est optimale mais au détriment de la vitesse de calcul.
- **les algorithmes en trois étapes** : la recherche de la nouvelle position du bloc courant est d'abord effectuée avec une précision grossière (déplacement important du bloc courant), puis la précision est améliorée d'une manière itérative (décroissance de la valeur du déplacement). A chaque itération, l'algorithme retient les trois dernières valeurs de la fonction de coût, $F_c()$. La condition d'arrêt de l'algorithme est fonction de ces trois valeurs (voir [Reoxiang 94]). Ce principe est utilisé dans le standard de codage vidéo H.263+ [4i2i 06].
- **la recherche logarithmique** : la fenêtre de recherche S est explorée dans la direction du minimum local de la fonction de coût
- **la recherche hiérarchique** : le principe est d'utiliser une recherche complète mais qui est appliquée sur une image avec un niveau de détail réduit. En fonction de la précision désirée les résultats sont corrigés d'une manière adaptative (voir [Lin 98]).
- **la recherche hybride** : au début de l'estimation on détermine le type de mouvement présent dans la séquence (lent, rapide ou stationnaire). Ensuite, une méthode adaptée à la situation est choisie pour calculer les vecteurs de mouvement (voir [Ge 02]).
- **la recherche binaire** : le principe est de diviser la fenêtre de recherche S en un certain nombre de régions et d'effectuer la recherche complète seulement dans une de ces régions (voir [Zahariadis 96]).
- **autres méthodes** : comme autres méthodes on peut mentionner : la recherche en 4 étapes, la recherche orthogonale, l'algorithme "one at a time", l'algorithme "cross-search", la recherche en spirale (voir [Turaga 98]).

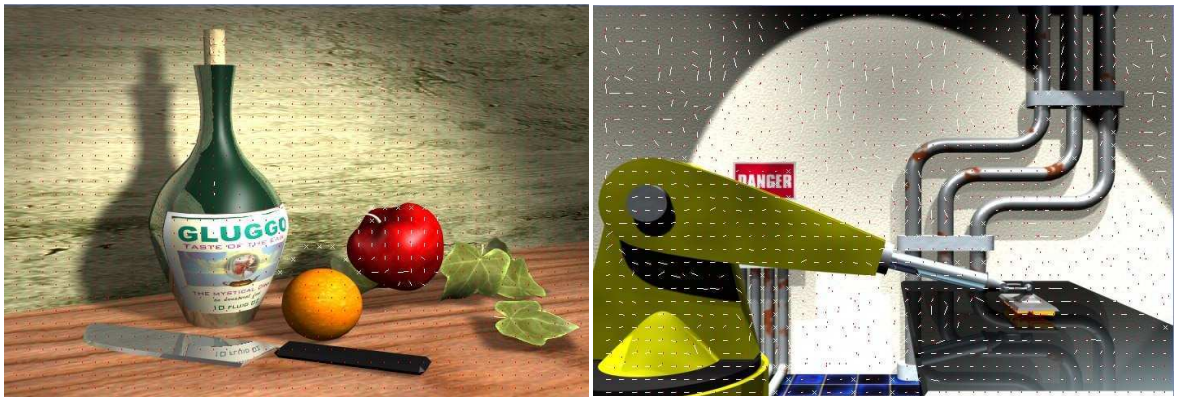
Pour la caractérisation du mouvement de caméra proposée dans le Chapitre 3, parmi toutes ces approches, nous avons sélectionné trois algorithmes de recherche : *la recherche complète*, *la recherche logarithmique* et *la recherche en trois étapes*. Nous proposons ici une étude comparative de ces trois méthodes.

Les méthodes de recherche ont été testées, sans optimisation, sur plusieurs films d'animation en utilisant une machine AMD 1.4GHz et 512MB de RAM sur des images RVB d'une résolution de 720×546 pixels. Les temps moyens d'exécution obtenus sont les suivants : *recherche complète 4.8s*, *recherche logarithmique 3.82s* et *standard H.263+ 3.89s*.

L'algorithme le plus rapide est *la recherche logarithmique* mais au détriment de la qualité de l'estimation du mouvement obtenue (voir Figure C.1.a). Comme la recherche de la nouvelle position du bloc courant commence à une certaine distance du bloc courant, si la texture de l'image est similaire à celle de la fenêtre de recherche, l'algorithme peut se positionner sur un minimum local de la fonction de coût, et de ce fait l'algorithme ne convergera pas.

L'algorithme en trois étapes du standard H.263+ a un temps de calcul voisin de celui de la recherche logarithmique, mais avec une meilleure précision de l'estimation du mouvement (voir Figure C.1.b). Ceci vient du fait que cette méthode utilise un déplacement initial fixé

à un pixel, déplacement qui ne change pas pendant l'estimation.



(a) Recherche logaritmique

(b) H.263+

(c) Recherche complète

FIG. C.1 – Exemple d'estimation du mouvement en utilisant les 3 méthodes de recherche : déplacement d'un objet (pomme) vers le haut (images de gauche), déplacement de caméra vers la gauche (images de droite). La direction de déplacement est indiquée par le point rouge et la discontinuité du mouvement par le x ($\tau_{discont} = 10000$, voir Section 3.2.2).

L'algorithme le plus lent est bien sûr *la recherche complète* car il utilise tous les blocs dans la fenêtre de recherche. Cependant, c'est lui qui donne la meilleure précision d'estimation (voir Figure C.1.c).

Annexe D

Exemples de résumés adaptatifs

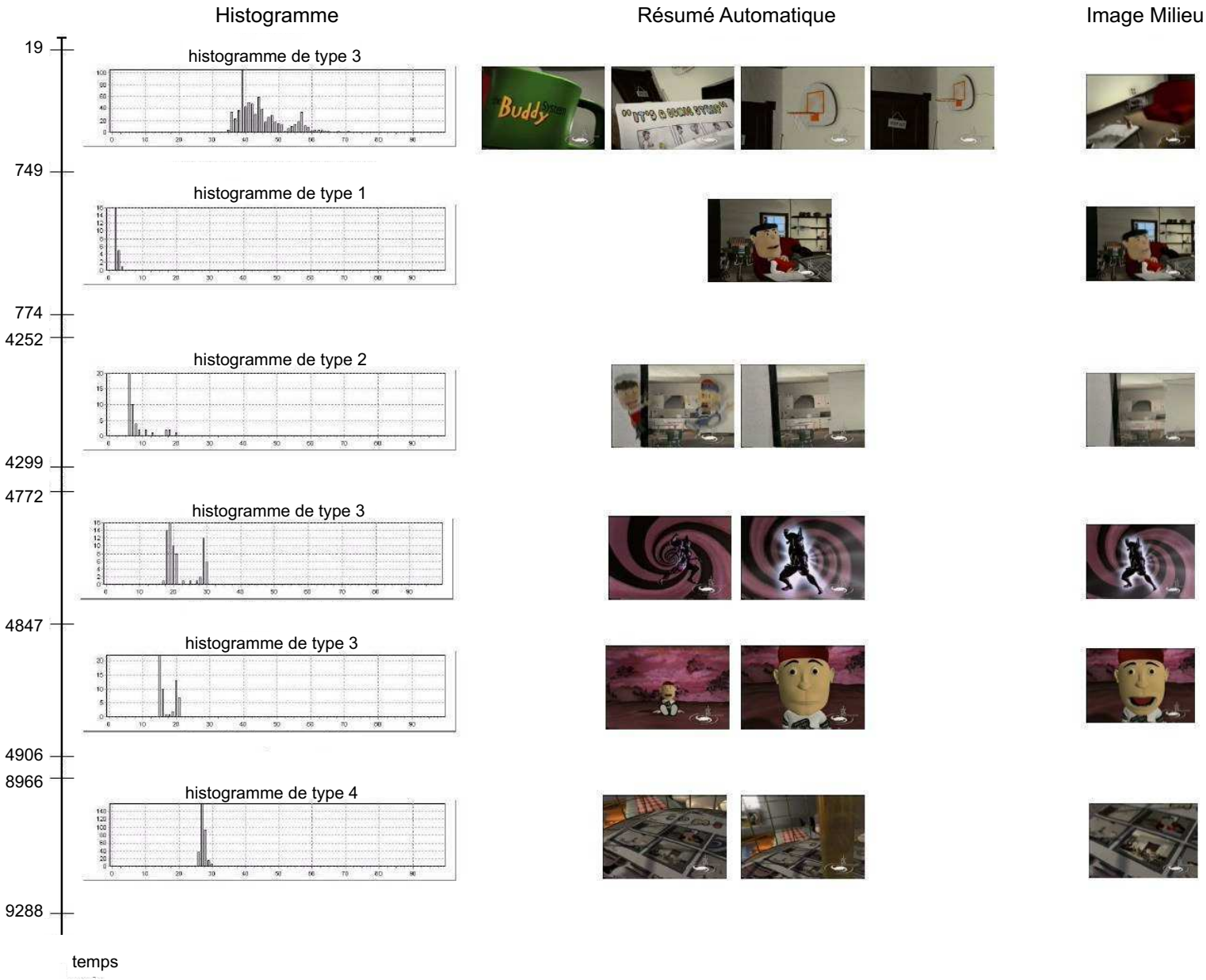


FIG. D.1 – Résultats de l'extraction adaptative des images sur quelques plans du film "The Buddy System" [CICA 06].

FIG. D.2 – Résultats de l'extraction adaptative des images sur quelques plans du film "Gazoon" [CICA 06].

Annexe E

L'évaluation de résumés

« Casa » (6m05s)

Résumé en images = *représentation globale du contenu du film*

- Estimez-vous que le « résumé en images » représente bien le **contenu** du film ?

X	1	2	3	4	5	6	7	8	9	10
Je ne sais pas	Pas du tout		Très peu		Partiellement		En grande partie		Totalement	

- Estimez-vous que le nombre d'images est :

X	1	2	3	4	5	6	7	8	9	10
Je ne sais pas	Trop petit		Petit		Suffisant		Elevé		Trop élevé	

Résumé « bande-annonce » = *résumé des passages où l'action est importante*

- Pensez-vous que le résumé « bande-annonce » **contient les passages les plus importants** du film ?

X	1	2	3	4	5	6	7	8	9	10
Je ne sais pas	Pas du tout		Très peu		Partiellement		En grande partie		Totalement	

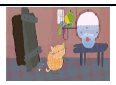
- Comment trouvez-vous **la durée** du résumé proposé ?

Trop courte	Courte	Correcte	Longue	Trop Longue

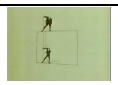
Fig. E.1 – Exemple de questionnaire utilisé pour l'évaluation de résumés.

Annexe F

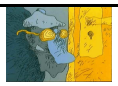
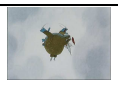
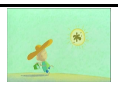
Extrait de la base d'animation

$N°$	Image	Nom	Durée	Technique/Son/Année
1		**A Crushed World**	6min42s	*Animation d'objets, sans dialogue ni commentaire*, (1986)
2		**Amerlock**	1min57s	*Pâte à modeler, sans dialogue ni commentaire*, (1989)
3		**A Viagem**	7min32s	*Vues réelles, sans dialogue ni commentaire*, (1998)
4		**At the End of the Earth**	7min28s	*Dessin sur cellulos, sans dialogue ni commentaire*, (1998)
5		**Casa**	6min05s	*Dessin sur cellulos, dialogue*, (2003)
6		**Cœur de Secours**	9min05s	*Sans commentaire ni dialogue*, (1973)
7		**David**	8min12s	*Dessin sur cellulos, dialogue*, (1977)
8		**Circuit Marine**	5min35s	*Dessin sur papier, éléments découpés, sans dialogue ni commentaire*, (2003)
9		**Ex-Enfant**	4min7s	*Écran d'épingles, sans dialogue ni commentaire*, (1994)
10		**Ferrailles**	6min15s	*Marionnettes, sans dialogue ni commentaire*, (1996)
11		**Firehouse**	5min9s	*Effects spéciaux, sans dialogue ni commentaire*, (1998)
12		**Gallina Vogelbirdae**	12min51s	*Sans commentaire ni dialogue*, (1963)

Tab. F.1 – Extrait de la base des films d'animation utilisée pour les tests (partie 1).

$N°$	Image	Nom	Durée	Technique/Son/Année
13		**Gazoon**	2min47s	*Ordinateur 3D, Sans dialogue ni commentaire*, (1998)
14		**Greek Tragedy**	6min32s	*Dessin sur cellulos, sans dialogue ni commentaire*, (1985)
15		**Le Moine et le Poisson**	5min59s	*Dessin sur cellulos (encre de chine, gouache), sans dialogue ni commentaire*, (1994)
16		**Le Pas**	8min57s	*Sans commentaire ni dialogue*, (1974)
17		**Le Rêve du Diable**	10min9s	*Aucune information disponible*
18		**L'Homme aux Bras Ballants**	3min38s	*Marionnettes, sans dialogue ni commentaire*, (1997)
19		**Mr. Pascal**	6min50s	*Dessin sur cellulos, sans commentaire ni dialogue*, (1979)
20		**Moznosti Dialogu**	11min13s	*Animation d'objets, sans dialogue ni commentaire*, (1982)
21		**Och, och**	5min52s	*Dessin sur cellulos, sans commentaire ni dialogue*, (1973)
22		**Paradise**	14min3s	*Éléments découpés, sans dialogue ni commentaire*, (1984)
23		**Ptica i Crv**	5min29s	*Sans commentaire ni dialogue*, (1977)
24		**Repete**	7min52s	*Dessin sur papier (crayon), sans dialogue ni commentaire*, (1995)
25		**Ropedance**	8min27s	*Sans commentaire ni dialogue*
26		**Tamer of Wild Horses**	7min33s	*Dessin sur cellulos, sans commentaire ni dialogue*, (1964)
27		**Tango**	8min2s	*Pixillation, sans commentaire ni dialogue*, (1981)
28		**The Breath**	3min59s	*Peinture sur verre, sans commentaire ni dialogue*, (1977)
29		**The Buddy System**	6min19s	*Ordinateur 3d, dialogue*, (1999)
30		**The Flying Magpie**	9min31s	*Eléments découpés en papier, sans commentaire ni dialogue*, (1964)

TAB. F.2 – Extrait de la base des films d'animation utilisée pour les tests (partie 2).

$N°$	Image	Nom	Durée	Technique/Son/Année
31		**The Flying Man**	2min21s	*Peinture sur papier, sans commentaire ni dialogue*, (1962)
32		**The Hill Farm**	16min39s	*Dessin sur papier (crayon), sans dialogue ni commentaire*, (1988)
33		**The Lion and the Song**	15min15s	*Marionnettes, sans commentaire ni dialogue*, (1959)
34		**The Sand Castle**	12min12s	*Marionnettes, sans commentaire ni dialogue*, (1977)
35		**The Young Lady and the Cellist**	8min42s	*Eléments découpés en papier, sans commentaire ni dialogue*, (1964)
36		**Fini Zayo**	7min2s	*Sans dialogue ni commentaire*, (2000)
37		**François le Vaillant**	8min56s	*Sans dialogue ni commentaire*, (2002)
38		**Histoire Extraordinaire Demm Keeskemet**	3min41s	*Animation sur celluloïd, sans commentaire ni dialogue*, (1992)
39		**L'Egoiste**	3min9s	*Dessins sur cellos et décors en volume, commentaires*, (1996)
40		**La Bouche Cousue**	2min48s	*Marionnettes, dialogue*, (1998)
41		**La Cancion du Microsillon**	8min29s	*Marionnettes, sans dialogue ni commentaire*, (2002)
42		**La Grande Migration**	7min14s	*Dessin sur papier (pastel), sans dialogue ni commentaire*, (1995)
43		**Le Chat d'Appartement**	6min42s	*Dessins animés sur cello et papier découpé, dialogue*, (1998)
44		**Le Château des Autres**	4min58s	*Pâte à modeler, sans dialogue ni commentaire*
45		**Le Roman de Mon Ame**	5min20s	*Dessin sur cellulos, sans dialogue ni commentaire*, (1997)
46		**Le Trop Petit Prince**	6min26s	*Dessin sur cellulos, sans dialogue ni commentaire*, (2002)
47		**The Wall**	9min7s	*Pâte à modeler, sans dialogue ni commentaire*, (1992)

Tab. F.3 – Extrait de la base des films d'animation utilisée pour les tests (partie 3).

$N°$	**Image**	**Nom**	**Durée**	**Technique/Son/Année**
48		**Nos Adieux au Music Hall**	2min1s	*Animation d'objets, sans dialogue ni commentaire*, (1989)
49		**Paroles en l'Air**	6min50s	*Dessin sur cellulos (fusain), sans dialogue ni commentaire*, (1995)
50		**Petite Escapade**	5min21s	*Marionnettes, dialogue*, (2001)
51		**Sculptures**	1min34s	*Pâte à modeler, dialogue*, (1988)
52		**Une Bonne Journée**	7min12s	*Dessins animés au crayon de couleurs sur papier, dialogue*, (1994)

TAB. F.4 – Extrait de la base des films d'animation utilisée pour les tests (partie 4).

Annexe G

Résultats : description du contenu

Le film "Amerlock"

Pour le film "Amerlock" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.1) :

- **synopsis :** "Un jeu avec quelques grands mythes et personnages mythifiés des États-Unis d'Amérique." [CICA 06].
- **couleurs élémentaires :** "Blue" 43,23%, "Azure" 29,33%, "Gray" 24,09%, "Red" 7,25%, "Black" 6,26%, "Magenta" 2,98%, "White" 0,31%, "Yellow" 0,22%, "Cyan" 0,12%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 0.04$	*"rythme lent"* (**1**)[1]
$100 \cdot R_{action} = 8.44\%$	*"action faible"* (**1**)
$100 \cdot R_{trans} = 0\%$	*"contenu mystérieux réduit"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 66.18\%$	*"présence élevée de couleurs claires"* (**1**)
$100 \cdot P_{foncées} = 33.82\%$	*"présence faible de couleurs foncées"* (**0.95**)
$100 \cdot P_{fortes} = 0\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 55.49\%$	*"présence moyenne de couleurs faiblement saturées"* (**1**)
$100 \cdot P_{chaudes} = 10.48\%$	*"présence faible de couleurs chaudes"* (**1**)
$100 \cdot P_{froides} = 58.86\%$	*"présence moyenne de couleurs froides"* (**1**)
$100 \cdot P_{var} = 30.09\%$	*"variété des couleurs faible"* (**0.98**)
$100 \cdot P_{div} = 30.77\%$	*"diversité des couleurs faible"* (**0.87**)
$100 \cdot P_{adj} = 50\%$	*"couleurs adjacentes : oui"* (**0.51**)
$100 \cdot P_{compl} = 33.33\%$	*"couleurs complémentaires : non"* (**0.98**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0.95**) *"les couleurs prédominantes sont foncées"* (**0**)

[1] le nombre entre parenthèses représente le degré de vérité de la description, valeur 0 = négation totale, valeur 1 = affirmation totale.

	"il y a un contraste claire-foncé" (**0**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une faible saturation"* (**0**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0.51**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0.02**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**0.02**)

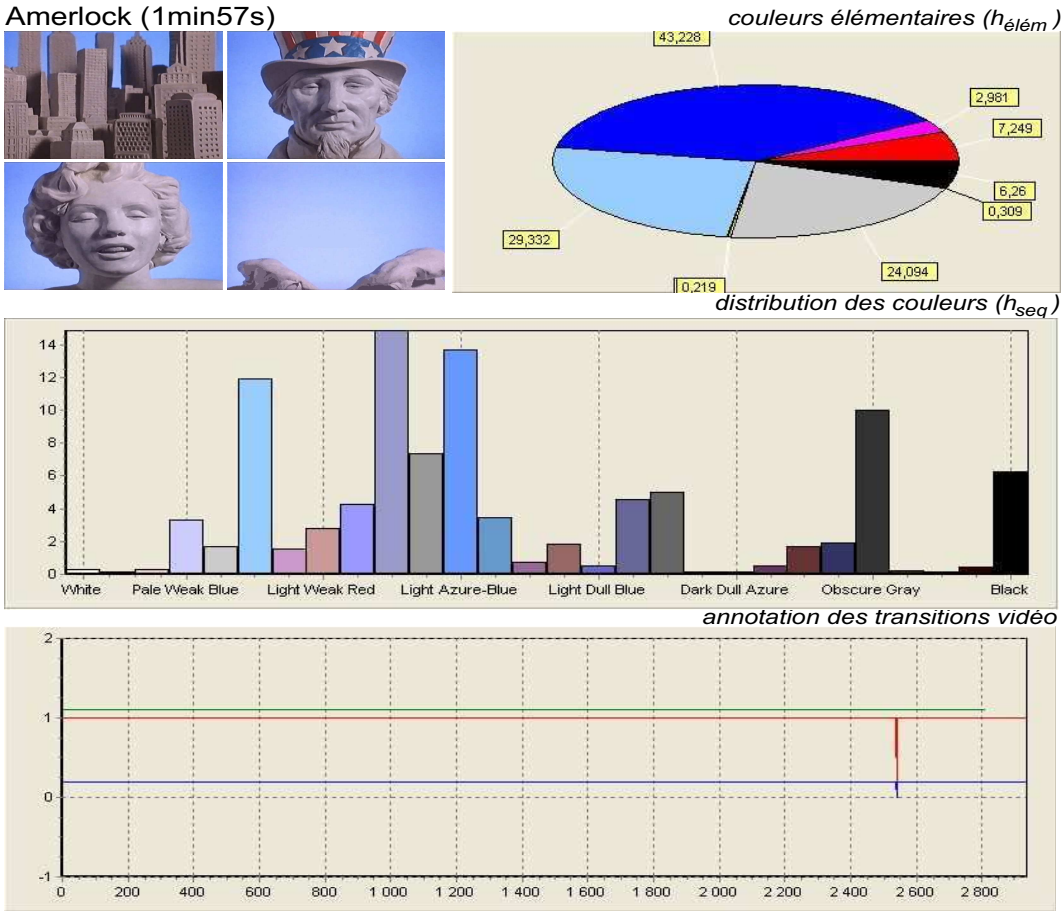


FIG. G.1 – Film "Amerlock" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

La technique d'animation utilisée par le film "Amerlock" est la pâte à modeler. L'action du film se déroule sur une seule scène où un morceau de pâte à modeler prend la forme et la

couleur de différents personnages. Ainsi, le *rythme* et *le contenu en terme d'action* sont-ils faibles (voir l'annotation visuelle des transitions dans la Figure G.1).

L'utilisation de la pâte à modeler fait que la palette couleur utilisée est limitée à quelques couleurs. Les couleurs élémentaires prédominantes sont le *"Blue"*, *"Azure"*, *"Gray"* et *"Red"*. La plupart des couleurs de la séquence sont *claires*, avec une *variété et une diversité des couleurs réduite* (voir les histogrammes couleurs dans la Figure G.1). Il n'y a pas prédominance de couleurs chaudes ou saturées, la séquence utilise en faibles proportions des couleurs froides, chaudes, et faiblement saturées.

Le film "Casa"

Pour le film "Casa" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.2) :

- **synopsis :** "À travers le regard d'une femme, nous vivons le retour d'un jeune homme après plusieurs années de séparation. Quels sont leurs liens ? Pourquoi cette absence ? Une tragédie en trois actes : attente, confrontation et acceptation." [uniFrance 06].
- **couleurs élémentaires :** "Orange" 35,39%, "Red" 28,33%, "Yellow" 15,88%, "Gray" 13,07%, "Cyan" 8,06%, "Black" 6,47%, "Green" 4,85%, "White" 0,17%, "Azure" 0,13%, "Magenta" 0,12%, "Pink" 0,11%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 0.74$	*"rythme moyen"* (**1**)
$100 \cdot R_{action} = 87.86\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 2.38\%$	*"contenu mystérieux moyen"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 49.1\%$	*"présence moyenne de couleurs claires"* (**0.94**)
$100 \cdot P_{foncées} = 50.64\%$	*"présence moyenne de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 6.61\%$	*"contenu faible en couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 61.22\%$	*"présence moyenne de couleurs faiblement saturées"* (**0.79**)
$100 \cdot P_{chaudes} = 67.21\%$	*"présence élevée de couleurs chaudes"* (**1**)
$100 \cdot P_{froides} = 13.1\%$	*"présence faible de couleurs froides"* (**1**)
$100 \cdot P_{var} = 34.72\%$	*"variété des couleurs moyenne"* (**0.78**)
$100 \cdot P_{div} = 46.15\%$	*"diversité des couleurs moyenne"* (**1**)
$100 \cdot P_{adj} = 87.5\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 62.5\%$	*"couleurs complémentaires : oui"* (**0.89**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**0**) *"il y a un contraste claire-foncé"* (**0.94**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une faible saturation"* (**0.21**)

	"il y a un contraste de saturation" (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**1**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0.11**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**0.89**)

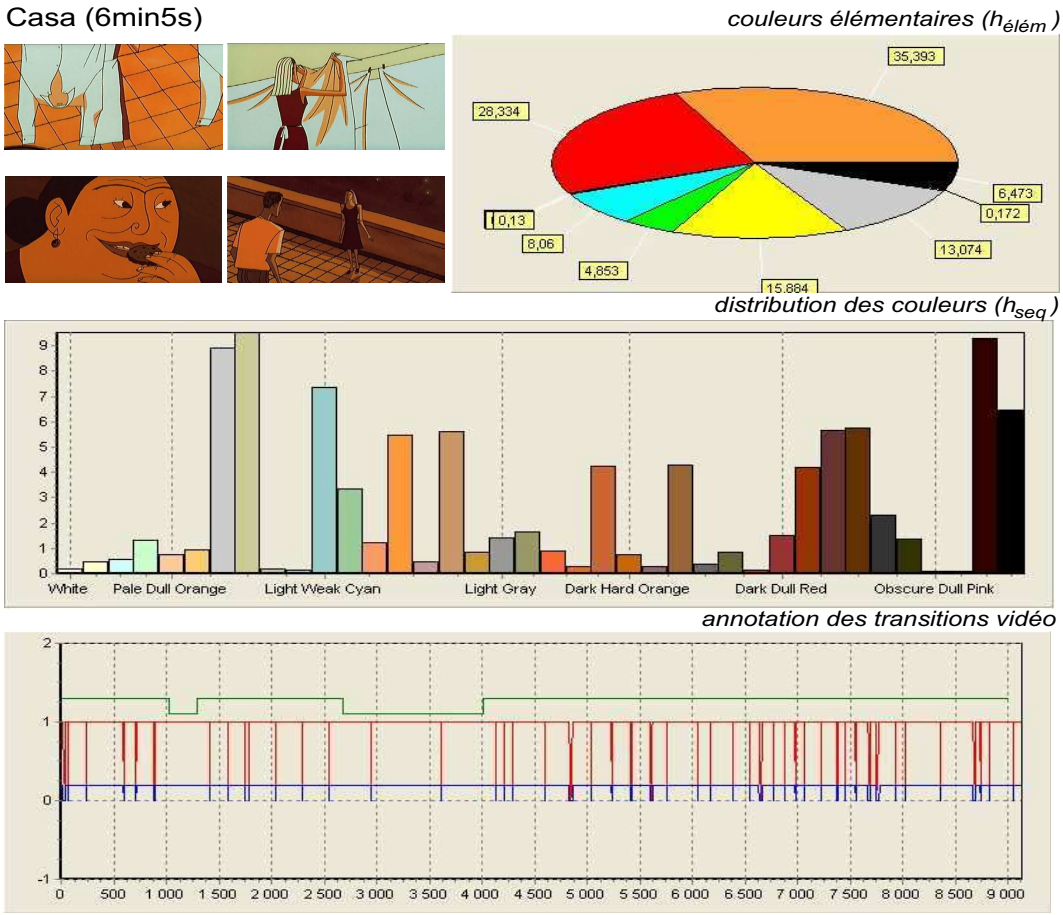


FIG. G.2 – Film "Casa" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Le film "Casa" a un *rythme moyen* à peu près constant dans la première partie de la séquence. Le rythme augmente vers la fin du film. La plupart des passages du film sont importants du point de vue de l'action contenue (*contenu en terme d'action élevé*) (voir l'annotation visuelle des transitions dans la Figure G.2). De plus, le contenu est *énigmatique*, car la confrontation des personnages n'est montrée que vers la fin : "une tragédie en trois

actes : attente, confrontation et acceptation" (voir le synopsis). L'évolution de l'action, de l'attente à la confrontation, est visible dans la structure des plans. La première partie du film, qui correspond à l'attente, a une fréquence de changements de plans plus réduite que la deuxième partie correspondant à la confrontation (voir la Figure G.2).

En ce qui concerne la distribution des couleurs, les couleurs élémentaires prédominantes sont le *"Orange"*, *"Red"*, *"Yellow"* et *"Gray"*, couleurs prédominantes qui sont plutôt des couleurs *chaudes*. Le film utilise dans les mêmes proportions des couleurs claires et foncées (*contraste clair-foncé*). En ce qui concerne la relation entre les couleurs on retrouve un contraste de couleurs *adjacentes-complémentaires*. La diversité et la variété des couleurs sont moyennes (voir les histogrammes couleurs en Figure G.2).

Le film "Circuit Marine"

Pour le film "Circuit Marine" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.3) :

- **synopsis :** "Être mangé ou ne pas être mangé ? Là est la question !" [uniFrance 06].
- **couleurs élémentaires :** "Gray" 31,23%, "Red" 22,20%, "Blue" 13,29%, "Azure" 10,96%, "Orange" 9,32%, "Cyan" 4,57%, "White" 4,21%, "Yellow" 4,15%, "Magenta" 1,16%, "Black" 0,99%, "Pink" 0,66%, "Green" 0,53%, "Spring" 0,43%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 1.82$	*"rythme rapide"* (**1**)
$100 \cdot R_{action} = 87.59\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 0\%$	*"contenu mystérieux réduit"* (**1**)
$100 \cdot R_{SCC} = 0.49\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 49.86\%$	*"présence moyenne de couleurs claires"* (**0.99**)
$100 \cdot P_{foncées} = 50.13\%$	*"présence moyenne de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 1.16\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 58.63\%$	*"présence moyenne de couleurs faiblement saturées"* (**1**)
$100 \cdot P_{chaudes} = 34.37\%$	*"présence faible de couleurs chaudes"* (**0.92**)
$100 \cdot P_{froides} = 29.91\%$	*"présence faible de couleurs froides"* (**1**)
$100 \cdot P_{var} = 66.2\%$	*"variété des couleurs élevée"* (**1**)
$100 \cdot P_{div} = 53.85\%$	*"diversité des couleurs moyenne"* (**1**)
$100 \cdot P_{adj} = 100\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 80\%$	*"couleurs complémentaires : oui"* (**1**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**0**) *"il y a un contraste claire-foncé"* (**0.99**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**0**) *"il y a un contraste de saturation"* (**0**)

Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**1**)

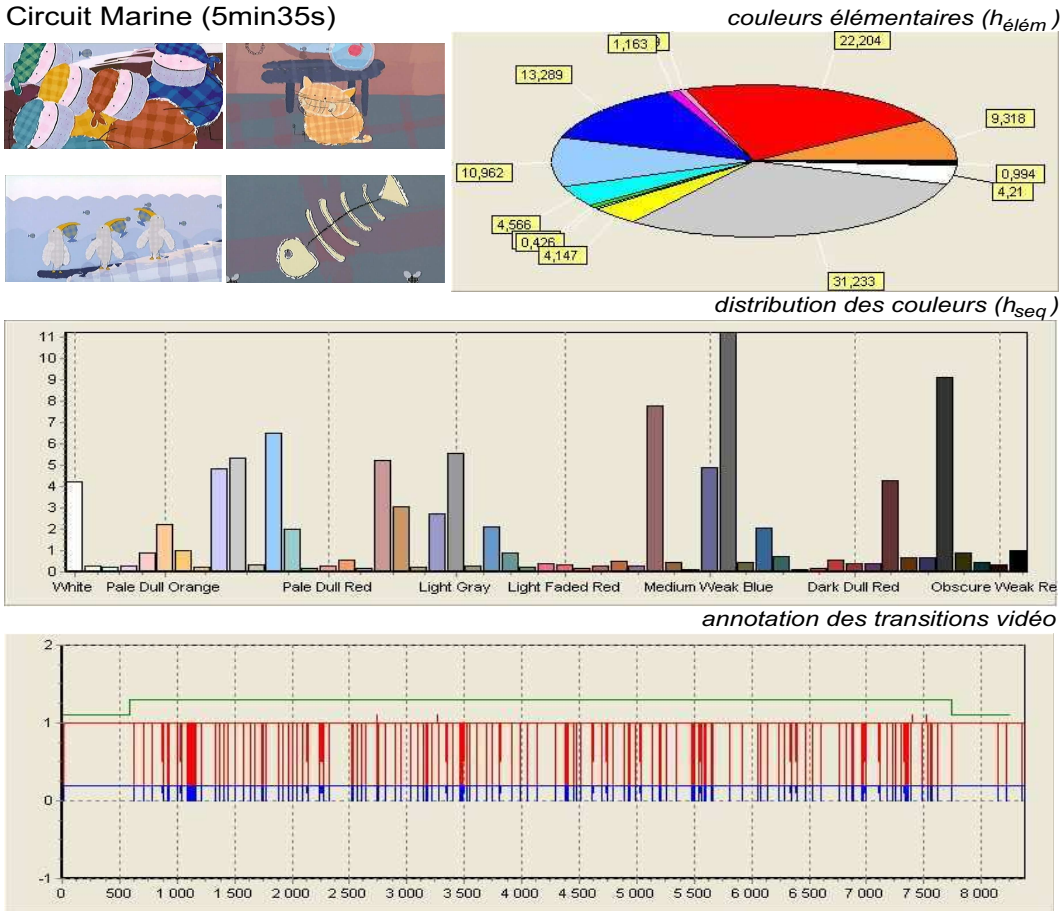


FIG. G.3 – Film "Circuit Marine" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Le film "Circuit Marine" a un rythme de déroulement de l'action constamment *alerte* : un chat fait des essais successifs et de plus en plus désespérés pour attraper et manger un poisson. Tous les passages de la séquence sont des passages d'action - *contenu en terme d'action est élevé* - (voir l'annotation visuelle des transitions dans la Figure G.3). Les couleurs élémentaires prédominantes du films sont le *"Gray"*, *"Red"*, *"Blue"*, *"Azure"* et *"Orange"*. Le film utilise en proportion identique des couleurs claires et foncées (*contraste*

clair-foncé). Il n'y a pas de prédominance d'une quelconque chaleur ou saturation des couleurs, la séquence utilisant en faibles proportions des couleurs froides, chaudes, saturées. Il y a une prédominance des couleurs plutôt faiblement saturées. En ce qui concerne la relation entre les couleurs on retrouve un contraste de couleurs *adjacentes-complémentaires*. La *variété des couleurs est élevée* car plus de 142 couleurs (sur un total de 216) sont présentes dans la séquence. D'autre part, la *diversité des couleurs est moyenne*.

Le film "Le Moine et le Poisson"

Pour le film "Le Moine et le Poisson" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.4) :

- **synopsis :** "Un moine découvre un poisson dans un réservoir d'eau près d'un monastère. Il essaie de l'attraper en utilisant toutes sortes de moyens. Au cours du film, la poursuite devient de plus en plus symbolique." [CICA 06].
- **couleurs élémentaires :** "Yellow" 60,28%, "Black" 19,63%, "Green" 7,06%, "Orange" 5,32%, "Spring" 4,26%, "Gray" 3,47%, "Cyan" 2,62%, "Red" 0,99%, "Azure" 0,12%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 2.37$	*"rythme rapide"* (**1**)
$100 \cdot R_{action} = 74.51\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 4.62\%$	*"contenu mystérieux élevé"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 48.91\%$	*"présence moyenne de couleurs claires"* (**0.93**)
$100 \cdot P_{foncées} = 51.1\%$	*"présence moyenne de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 2.94\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 70.18\%$	*"présence élevée de couleurs faiblement saturées"* (**1**)
$100 \cdot P_{chaudes} = 67.1\%$	*"présence élevée de couleurs chaudes"* (**1**)
$100 \cdot P_{froides} = 9.8\%$	*"présence faible de couleurs froides"* (**1**)
$100 \cdot P_{var} = 41.67\%$	*"variété des couleurs moyenne"* (**1**)
$100 \cdot P_{div} = 38.46\%$	*"diversité des couleurs moyenne"* (**1**)
$100 \cdot P_{adj} = 100\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 42.86\%$	*"couleurs complémentaires : non"* (**0.7**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**0**) *"il y a un contraste claire-foncé"* (**0.93**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**1**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**1**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)

Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0.7**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**0.3**)

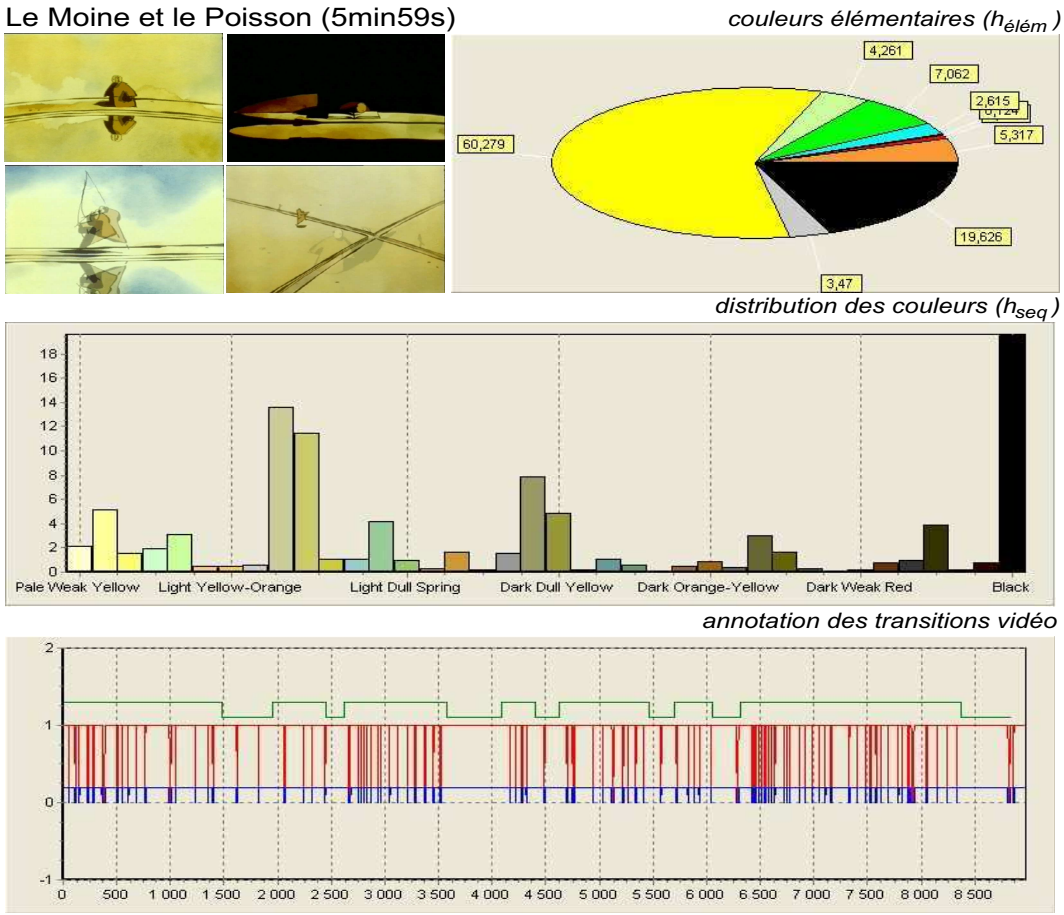


FIG. G.4 – Film "Le Moine et le Poisson" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

En observant l'annotation visuelle des transitions (voir la Figure G.4) on constate tout de suite que le *rythme et le contenu en terme d'action sont élevés*. La répétition des actions est indiquée aussi par le synopsis : la poursuite répétitive du poisson. De plus, le film à un contenu *mystérieux/énigmatique élevé* : la poursuite du poisson devient de plus en plus symbolique (voir aussi la Section 7.3.1).

En ce qui concerne la distribution des couleurs, les couleurs élémentaires prédominantes sont le *"Yellow"*, *"Black"* et *"Green"*. Le film utilise en proportions identiques des couleurs claires et foncées (*contraste clair-foncé*). La technique d'animation du film est le dessin sur cellulos en utilisant de l'encre de chine et des gouaches. L'utilisation de l'encre fait que les

couleurs sont plus diluées, les couleurs prédominantes du film ont donc une *faible saturation*. Du point de vue de la perception, les *couleurs chaudes* et *adjacentes* sont prédominantes. La diversité et la variété des couleurs est *moyenne* car le film utilise une seule couleur en grand proportion : la couleur Jaune représente 60% de la distribution totale (voir les histogrammes couleurs dans la Figure G.4), et toutes les autres couleurs sont en faible proportion.

Le film "Och, och"

Pour le film "Och, och" nous avons obtenu les paramètres et les descriptions linguistiques suivants (voir la Figure G.5) :

- **synopsis :** "Quand la construction du bâtiment se déroule bien, tout va bien ; mais dans le cas contraire..."
- **couleurs élémentaires :** "Gray" 52,80%, "Cyan" 26,96%, "Yellow" 6,15%, "Black" 5,11%, "Green" 4,88%, "Red" 1,78%, "Orange" 1,16%, "Azure" 1,10%, Blue 0,18%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 1.03$	*"rythme moyen"* (**0.82**)
$100 \cdot R_{action} = 67.3\%$	*"action élevée"* (**0.71**)
$100 \cdot R_{trans} = 1.58\%$	*"contenu mystérieux moyen"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 22.48\%$	*"présence faible de couleurs claires"* (**1**)
$100 \cdot P_{foncées} = 77.52\%$	*"présence élevée de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 0\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 41.83\%$	*"présence moyenne de couleurs faiblement saturées"* (**0.52**)
$100 \cdot P_{chaudes} = 9.15\%$	*"présence faible de couleurs chaudes"* (**1**)
$100 \cdot P_{froides} = 32.93\%$	*"présence faible de couleurs froides"* (**1**)
$100 \cdot P_{var} = 34.72\%$	*"variété des couleurs moyenne"* (**0.79**)
$100 \cdot P_{div} = 30.77\%$	*"diversité des couleurs faible"* (**0.87**)
$100 \cdot P_{adj} = 85.71\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 57.14\%$	*"couleurs complémentaires : oui"* (**0.73**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**1**) *"il y a un contraste claire-foncé"* (**0**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**0**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0.27**)

	"les couleurs prédominantes sont des couleurs complémentaires" (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**0.73**)

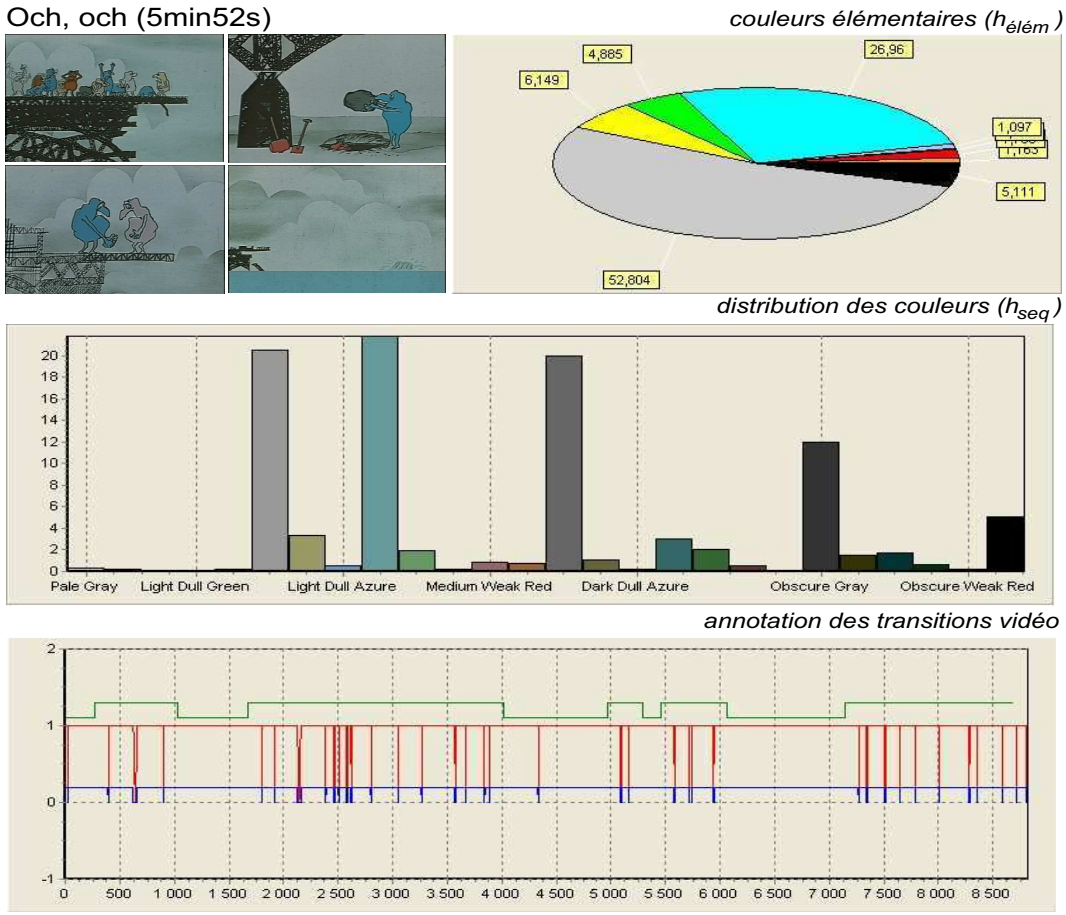


FIG. G.5 – Film "Och, och" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Les actions du film "Och, och" sont liées à un certain nombre d'événements qui se déroulent pendant la construction d'un bâtiment. Le rythme de déroulement de ces actions est *moyen*. La séquence comporte cinq passages avec de nombreuses transitions ce qui dénote une action élevée (voir la Figure G.5), c'est pourquoi la séquence a été classée avec un contenu en terme *d'action élevé* mais avec toutefois une valeur réduite (de 0.71). L'analyse des transitions vidéo a permis de caractériser le contenu mystérieux de la séquence comme *moyen*.

En ce qui concerne la distribution des couleurs, les couleurs élémentaires prédominantes sont le *"Gray"*, *"Cyan"* et *"Yellow"*. Les couleurs prédominantes dans ce cas sont des *couleurs foncées*. La prédominance du gris dans la distribution couleur de la séquence entraîne une distribution *faible* en couleurs chaudes, froides et elles sont faiblement saturées. Les couleurs

utilisées sont plutôt *adjacentes* que *complémentaires*, et le degré de vérité pour le contraste adjacent-complémentaire a une valeur de 0.73. Comme nous avons un nombre réduit de couleurs élémentaires, la diversité couleur de la séquence est *faible*. D'autre part la séquence utilise près de 73 couleurs différentes sur un total de 216 de la palette "Webmaster", ce qui donne une *variété des couleurs moyenne* (de 0.79) (voir les histogrammes couleurs dans la Figure G.5).

Le film "Tamer of Wild Horses"

Pour le film "Tamer of Wild Horses" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.6) :

- **synopsis :** "L'homme arrive-t-il à apprivoiser la bête en métal et feu ? Oui, mais seulement si c'est sans violence. Comprise et aimée, elle ramène l'homme dans l'espace."
- **couleurs élémentaires :** "Green" 39,92%, "Spring" 28,30%, "Teal" 10,60%, "Yellow" 10,51%, "Black" 7,63%, "Cyan" 5,97%, "Gray" 0,41%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 1.22$	*"rythme rapide"* (**1**)
$100 \cdot R_{action} = 79.94\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 1.51\%$	*"contenu mystérieux moyen"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 34.32\%$	*"présence faible de couleurs claires"* (**0.92**)
$100 \cdot P_{foncées} = 65.68\%$	*"présence élevée de couleurs foncées"* (**0.94**)
$100 \cdot P_{fortes} = 1.26\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 87.29\%$	*"présence élevée des couleurs faiblement saturées"* (**1**)
$100 \cdot P_{chaudes} = 37.3\%$	*"présence faible de couleurs chaudes"* (**0.75**)
$100 \cdot P_{froides} = 56.1\%$	*"présence moyenne de couleurs froides"* (**1**)
$100 \cdot P_{var} = 42.59\%$	*"variété des couleurs moyenne"* (**1**)
$100 \cdot P_{div} = 46.15\%$	*"diversité des couleurs moyenne"* (**1**)
$100 \cdot P_{adj} = 100\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 0\%$	*"couleurs complémentaires : non"* (**1**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**0.92**) *"il y a un contraste claire-foncé"* (**0**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**1**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0.25**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**1**)

	"les couleurs prédominantes sont des couleurs complémentaires" (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**0**)

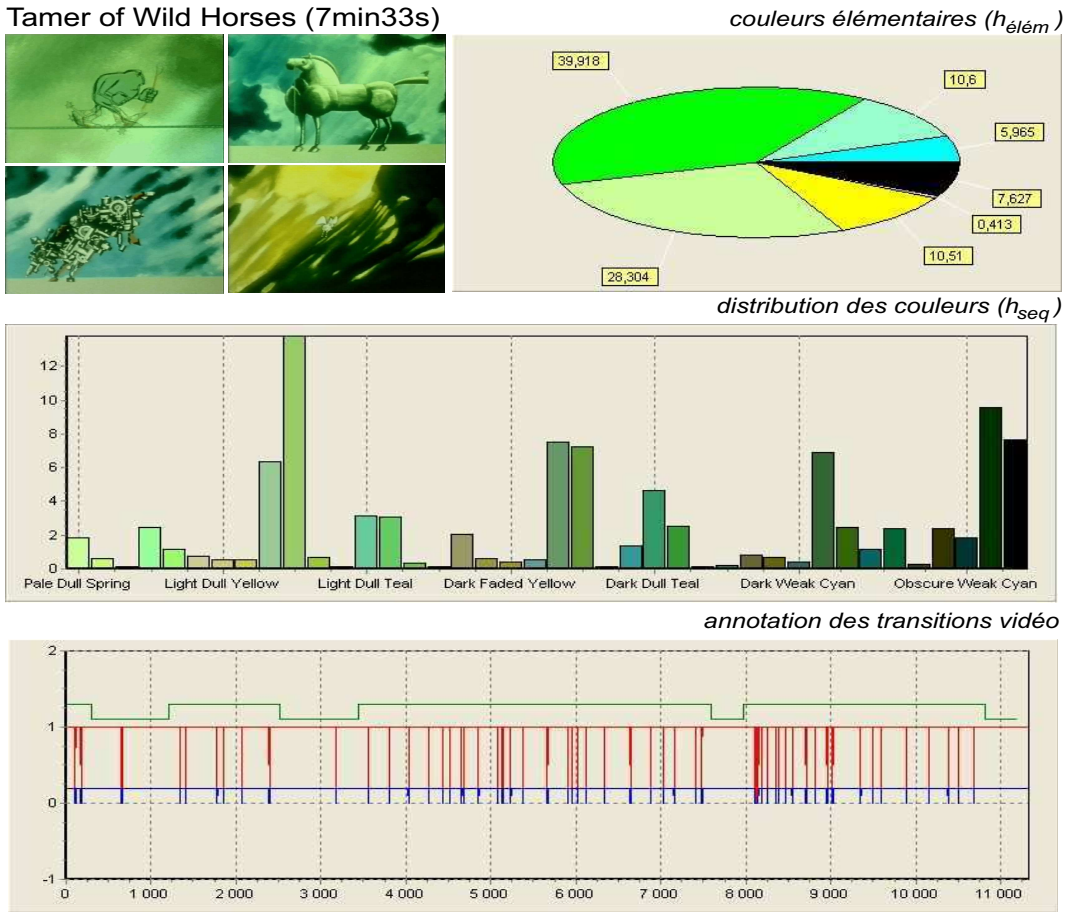


FIG. G.6 – Film "Tamer of Wild Horses" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Le film "Tamer of Wild Horses" a un *rythme élevé* et un *contenu en terme d'action élevé* (voir l'annotation visuelle des transitions dans la Figure G.6). Le *contenu mystérieux moyen* est lié au caractère fantastique du film (voir le synopsis). L'action est difficilement prévisible et du fait des couleurs utilisées les scènes ont une apparence mystérieuse.

Les couleurs élémentaires prédominantes sont le *"Green"*, *"Spring"*, *"Teal"*, *"Yellow"* et *"Black"*. La plupart des couleurs utilisées par le film sont des *couleurs foncées*. La prédominance de couleurs *faiblement saturées* est liée à la technique d'animation utilisée : le dessin sur cellulos. Les couleurs utilisées sont également *adjacentes* car toutes les couleurs prédominantes, "Teal", "Green", "Spring", "Yellow", sont des couleurs voisines sur la roue des couleurs d'Itten (voir la Figure 4.7 dans la Section 4.2.3), d'où une diversité et une variété des couleurs qui peut être qualifiées de *moyennes* (voir les histogrammes couleurs dans la

Figure G.6).

Le film "La Cancion du Microsillon"

Pour le film "La Cancion du Microsillon" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.7) :

- **synopsis :** "Dans un désert, un homme donne chaque soir un spectacle de music-hall devant des gradins totalement vides." [uniFrance 06].
- **couleurs élémentaires :** "Gray" 33,19%, "Orange" 21,31%, "Red" 14,55%, "Yellow" 10,59%, "Azure" 7,59%, "Black" 6,78%, "Blue" 4,77%, "Cyan" 4,00%, "Magenta" 0,16%, "Green" 0,13%, "White" 0,10%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 1.08$	*"rythme moyen"* (**0.6**)
$100 \cdot R_{action} = 55.59\%$	*"action moyenne"* (**1**)
$100 \cdot R_{trans} = 1.36\%$	*"contenu mystérieux moyen"* (**1**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 20.83\%$	*"présence faible de couleurs claires"* (**1**)
$100 \cdot P_{foncées} = 79.17\%$	*"présence élevée de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 0.04\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 56.71\%$	*"présence moyenne de couleurs faiblement saturées"* (**1**)
$100 \cdot P_{chaudes} = 44.13\%$	*"présence moyenne de couleurs chaudes"* (**0.65**)
$100 \cdot P_{froides} = 15.8\%$	*"présence faible de couleurs froides"* (**1**)
$100 \cdot P_{var} = 40.74\%$	*"variété des couleurs moyenne"* (**1**)
$100 \cdot P_{div} = 46.15\%$	*"diversité des couleurs moyenne"* (**1**)
$100 \cdot P_{adj} = 75\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 50\%$	*"couleurs complémentaires : oui"* (**0.51**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**1**) *"il y a un contraste claire-foncé"* (**0**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**0**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0.49**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0**)

	"il y a un contraste des couleurs adjacentes-complémentaires" (**0.51**)

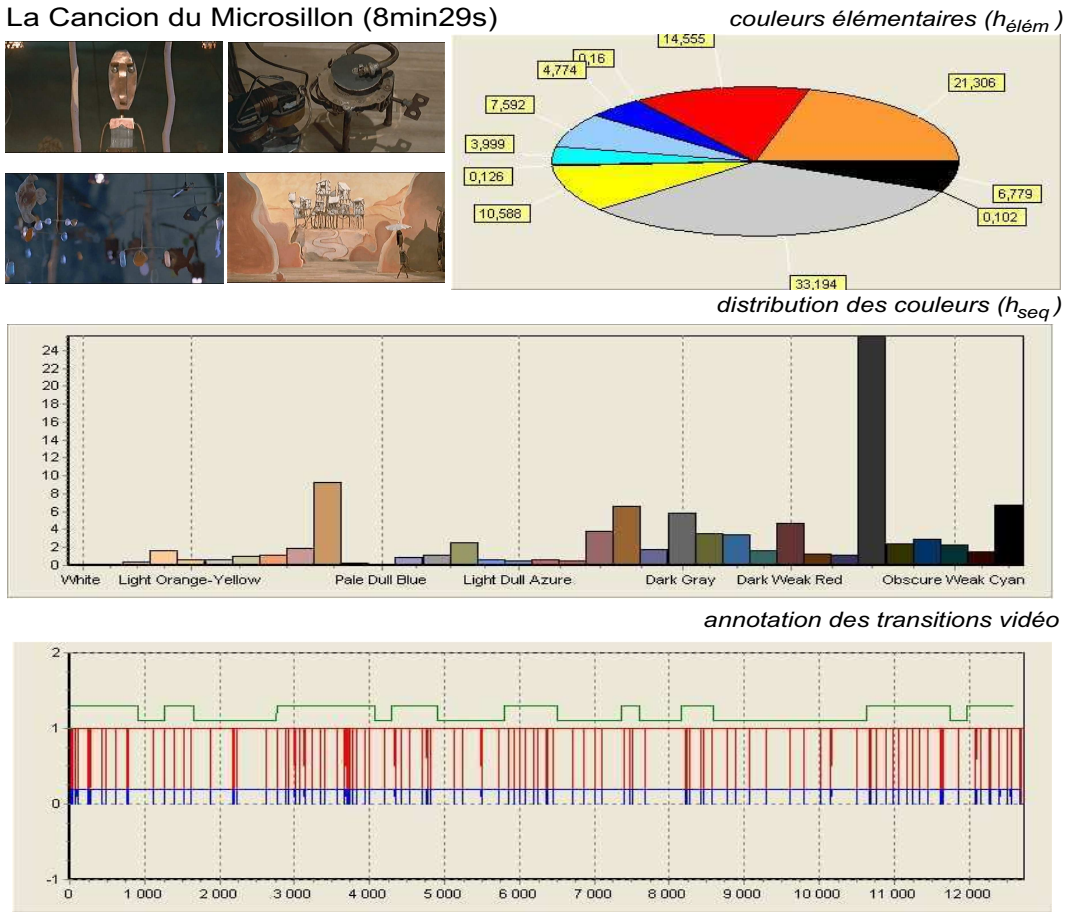


FIG. G.7 – Film "La Cancion du Microsillon" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Le film "La Cancion du Microsillon" a un rythme constant, caractérisé de moyen en terme de déroulement de l'action (*rythme moyen* avec une faible valeur de vérité de 0.6). L'action est ici liée à la répétition des spectacles de music-hall donnés par le personnage principal qui se déroulent dans le calme absolu, car il n'y a pas de spectateur. Le contenu global en terme d'actions du film est *moyen* (malgré de nombreuses transitions visibles dans la Figure G.7). Les scènes, les décors et le contenu lyrique difficilement compréhensible montrent le caractère mystérieux de ce film, c'est pourquoi il a été classé comme ayant un contenu énigmatique/mystérieux *moyen*.

Les couleurs élémentaires prédominantes sont le *"Gray"*, *"Orange"*, *"Red"* et *"Yellow"*. Les couleurs prédominantes sont les *couleurs foncées* et plutôt *adjacentes*. La prédominance du gris dans la distribution des couleurs de la séquence a comme résultat une distribution *faible* en couleurs chaudes, froides et faiblement saturées. La présence des couleurs saturées est négligeable car le pourcentage d'apparition est très faible 0.04%. La variété et la diversité

couleur sont *moyennes* dans ce cas (voir les histogrammes couleurs dans la Figure G.7).

Le film "Le Château des Autres"

Pour le film "Le Château des Autres" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.8) :

- **synopsis :** "Une visite scolaire a lieu dans un château immense. L'un des enfants s'attarde quelques secondes pour contempler une statue et perd son groupe." [uniFrance 06].
- **couleurs élémentaires :** "Orange" 29,78%, "Red" 25,86%, "Gray" 17,49%, "Yellow" 14,69%, "Black" 12,35%, "Cyan" 5,11%, "Green" 1,45%, "Azure" 0,26%, "Pink" 0,15%, "Spring" 0,12%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 1.88$	*"rythme rapide"* (**1**)
$100 \cdot R_{action} = 95.46\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 0.76\%$	*"contenu mystérieux moyen"* (**0.82**)
$100 \cdot R_{SCC} = 0\%$	*"contenu explosif : non"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 14.1\%$	*"présence faible de couleurs claires"* (**1**)
$100 \cdot P_{foncées} = 85.49\%$	*"présence élevée de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 2.09\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 60.98\%$	*"présence moyenne de couleurs faiblement saturées"* (**0.83**)
$100 \cdot P_{chaudes} = 63.28\%$	*"présence élevée de couleurs chaudes"* (**0.54**)
$100 \cdot P_{froides} = 6.84\%$	*"présence faible de couleurs froides"* (**1**)
$100 \cdot P_{var} = 47.22\%$	*"variété des couleurs moyenne"* (**1**)
$100 \cdot P_{div} = 38.46\%$	*"diversité des couleurs moyenne"* (**1**)
$100 \cdot P_{adj} = 100\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 62.5\%$	*"couleurs complémentaires : oui"* (**0.89**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**1**) *"il y a un contraste claire-foncé"* (**0**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**0.17**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0.54**) *"les couleurs prédominantes sont froides"* (**0**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0.11**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0**)

	"il y a un contraste des couleurs adjacentes-complémentaires" (**0.89**)

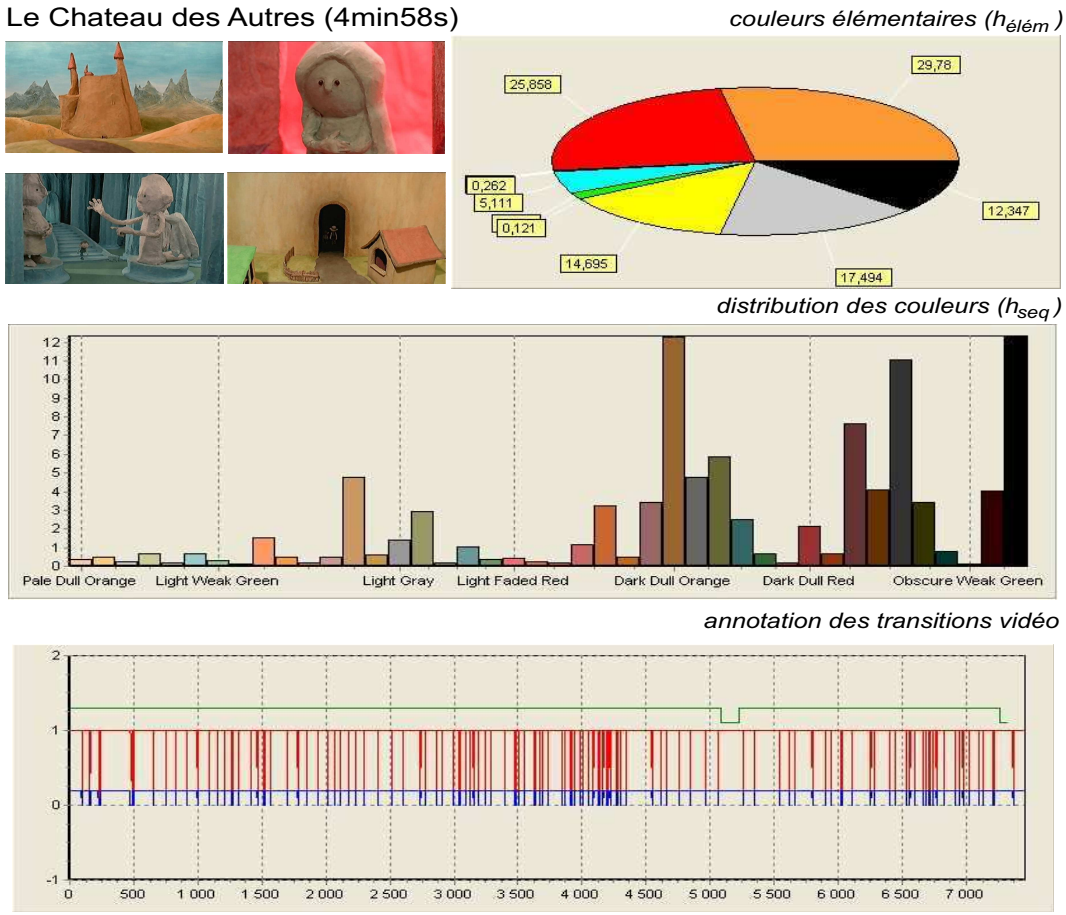


FIG. G.8 – Film "Le Château des Autres" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Le film "Le Château des Autres" a un rythme et un contenu en terme d'action *élevés*. Ceci apparaît dans l'annotation visuelle des transitions (qui présente une densité élevée de changements de plans, voir la Figure G.8). Les passages d'action couvrent le film en totalité. Le contenu du film devient *énigmatique/mystérieux* quand l'un des enfants s'attarde quelques secondes pour contempler une statue et perd son groupe.

Les couleurs élémentaires prédominantes du film sont le *"Orange"*, *"Red"*, *"Gray"*, *"Yellow"* et *"Black"*. Les couleurs prédominantes sont des *couleurs foncées* moyennement *chaudes*. En ce qui concerne la saturation, la plupart des couleurs ont une *faible saturation* (60%). Les couleurs utilisées par le film sont également adjacentes et complémentaires imposant un *contraste adjacent-complementaire* de *0.89*. Nous notons également une variété et une diversité réduites en terme de couleurs qui sont dues à la technique d'animation utilisant de la pâte à modeler (voir les histogrammes couleurs dans la Figure G.8).

Le film "François le Vaillant"

Pour le film "François le Vaillant" nous avons obtenu les paramètres et les descriptions linguistiques suivantes (voir la Figure G.9) :

- **synopsis :** "Une armée médiévale emmenée par un chef cruel et sanguinaire fait régner la terreur. François le Vaillant, la fleur à la lance, traverse, avec un certain détachement, le théâtre des ravages de la guerre." [uniFrance 06].
- ***couleurs élémentaires :*** "Azure" 54,23%, "Cyan" 23,74%, "Black" 13,66%, "Gray" 10,25%, "Red" 4,41%, "Blue" 3,25%, "Orange" 3,05%, "Yellow" 2,45%, "Green" 1,43%, "Teal" 1,11%, "Spring" 0,29%, "Magenta" 0,20%, "Pink" 0,19%, "White" 0,19%.
- **caractérisation des plans :**

Paramètre	Description
$\bar{v}_{T=5s} = 1.57$	*"rythme rapide"* (**1**)
$100 \cdot R_{action} = 70.23\%$	*"action élevée"* (**1**)
$100 \cdot R_{trans} = 0.69\%$	*"contenu mystérieux réduit"* (**0.53**)
$100 \cdot R_{SCC} = 1.78\%$	*"contenu explosif : oui"* (**1**)

- **caractérisation des couleurs :**

Paramètre	Description
$100 \cdot P_{claires} = 29\%$	*"présence faible de couleurs claires"* (**1**)
$100 \cdot P_{foncées} = 66.59\%$	*"présence élevée de couleurs foncées"* (**1**)
$100 \cdot P_{fortes} = 27.29\%$	*"présence faible de couleurs saturées"* (**1**)
$100 \cdot P_{faibles} = 34.5\%$	*"présence faible de couleurs faiblement saturées"* (**0.91**)
$100 \cdot P_{chaudes} = 10.21\%$	*"présence faible de couleurs chaudes"* (**1**)
$100 \cdot P_{froides} = 65.86\%$	*"présence élevée de couleurs froides"* (**0.97**)
$100 \cdot P_{var} = 87\%$	*"variété de couleurs élevée"* (**1**)
$100 \cdot P_{div} = 30.77\%$	*"diversité des couleurs faible"* (**0.87**)
$100 \cdot P_{adj} = 100\%$	*"couleurs adjacentes : oui"* (**1**)
$100 \cdot P_{compl} = 90.91\%$	*"couleurs complémentaires : oui"* (**1**)
Claire/foncé	*"les couleurs prédominantes sont claires"* (**0**) *"les couleurs prédominantes sont foncées"* (**1**) *"il y a un contraste claire-foncé"* (**0**)
Saturé/non saturé	*"les couleurs prédominantes sont saturées"* (**0**) *"les couleurs prédominantes ont une saturation faible"* (**0**) *"il y a un contraste de saturation"* (**0**)
Chaud/Froid	*"les couleurs prédominantes sont chaudes"* (**0**) *"les couleurs prédominantes sont froides"* (**0.97**) *"il y a un contraste chaud-froid"* (**0**)
Adjacent/ Complémentaire	*"les couleurs prédominantes sont des couleurs adjacentes"* (**0**) *"les couleurs prédominantes sont des couleurs complémentaires"* (**0**) *"il y a un contraste des couleurs adjacentes-complémentaires"* (**1**)

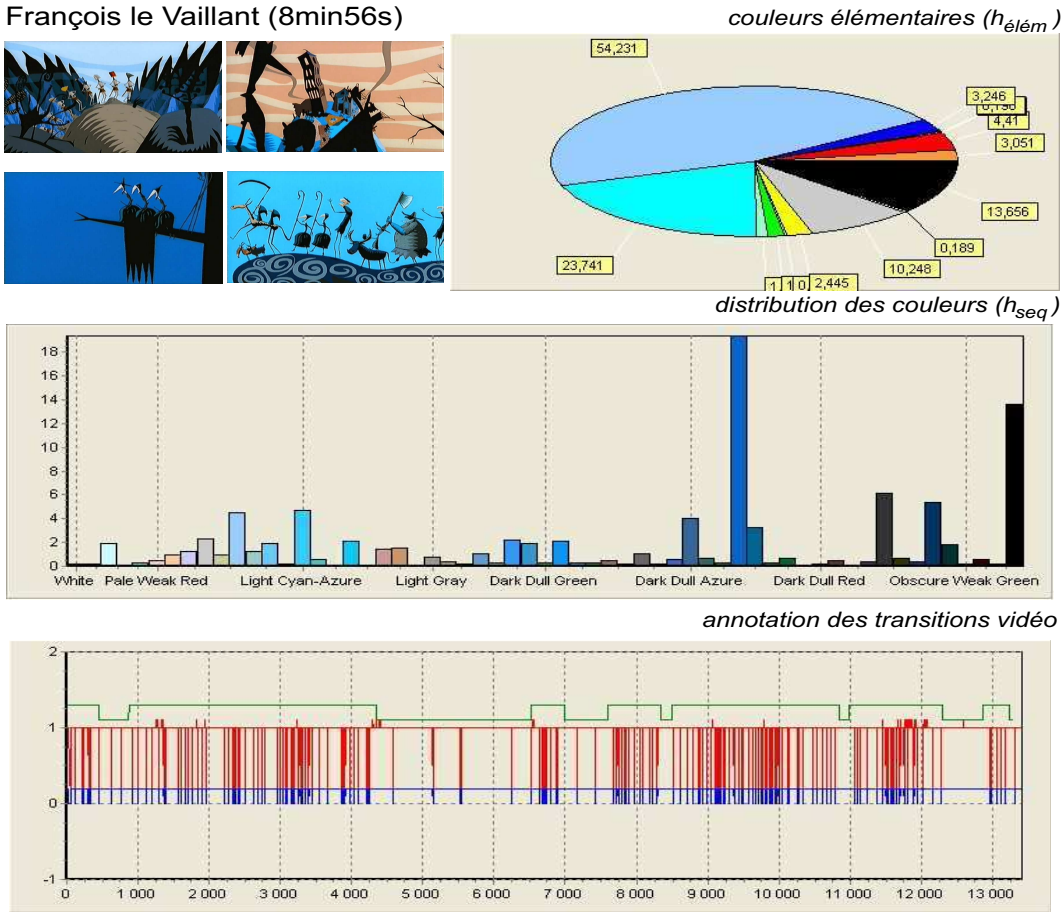


FIG. G.9 – Film "François le Vaillant" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Le film "François le Vaillant" a un rythme *élevé* au niveau du déroulement de la séquence, ce qui correspond à la description du synopsis : *une armée médiévale emmenée par un chef cruel et sanguinaire fait régner la terreur*. La plupart des passages du film sont des passages *d'action élevée* (voir dans l'annotation visuelle les passages avec de très nombreuses transitions dans la Figure G.9). On note un *faible contenu mystérieux*. La présence fréquente des effets de couleurs de type SCC, liés à l'atmosphère de guerre, donne un *caractère explosif* au film (voir la Figure 7.17.a dans la Section 7.3.1).

Les couleurs élémentaires prédominantes du film sont *"Azure"*, *"Cyan"*, *"Black"* et *"Gray"*, la teinte prédominante étant le Bleu. Les couleurs prédominantes sont des *couleurs foncées* et *froides*. En ce qui concerne la saturation des couleurs on retrouve une distribution *faiblement saturée*. Les couleurs utilisées par le film sont également adjacentes et complémentaires. A partir de l'histogramme global pondéré de la séquence, nous pouvons dire que le film utilise une *variété élevée* de couleurs, puisque 187 couleurs différentes se retrouve dans le film sur un total de 216 couleurs que comporte la palette "Webmaster". D'autre part la diversité couleur est *faible* car le film n'utilise que quelques couleurs élémentaires (voir les histogrammes couleurs dans la Figure G.9).

Annexe H

Comparatif des films d'animation en utilisant les gamuts sémantiques

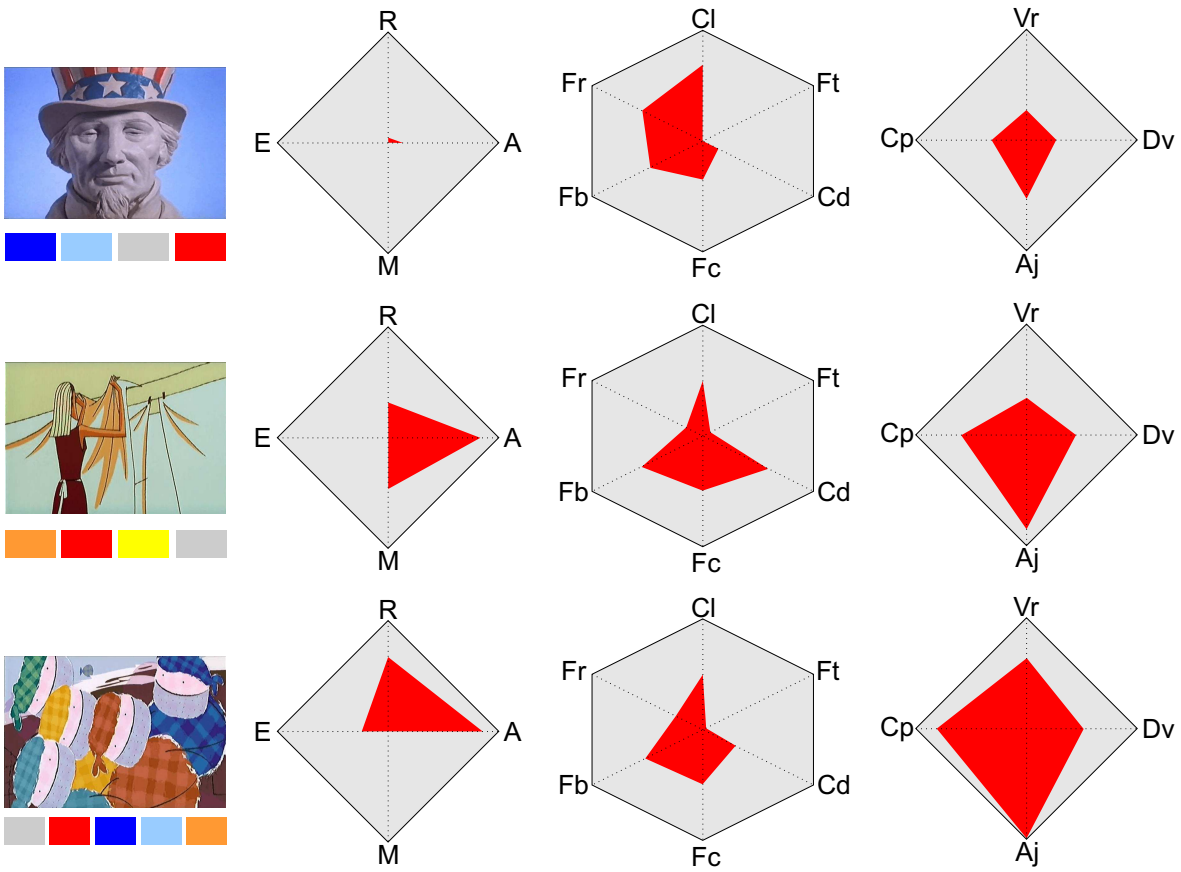


Fig. H.1 – Les gamuts sémantiques G^p, G^c et G^{rl} (voir la Section 7.6) obtenus pour les films suivants (ordre de haut en bas) : "Amerlock", "Casa" et "Circuit Marine" (les couleurs élémentaires prédominantes de chaque film sont illustrées en dessous de l'image).

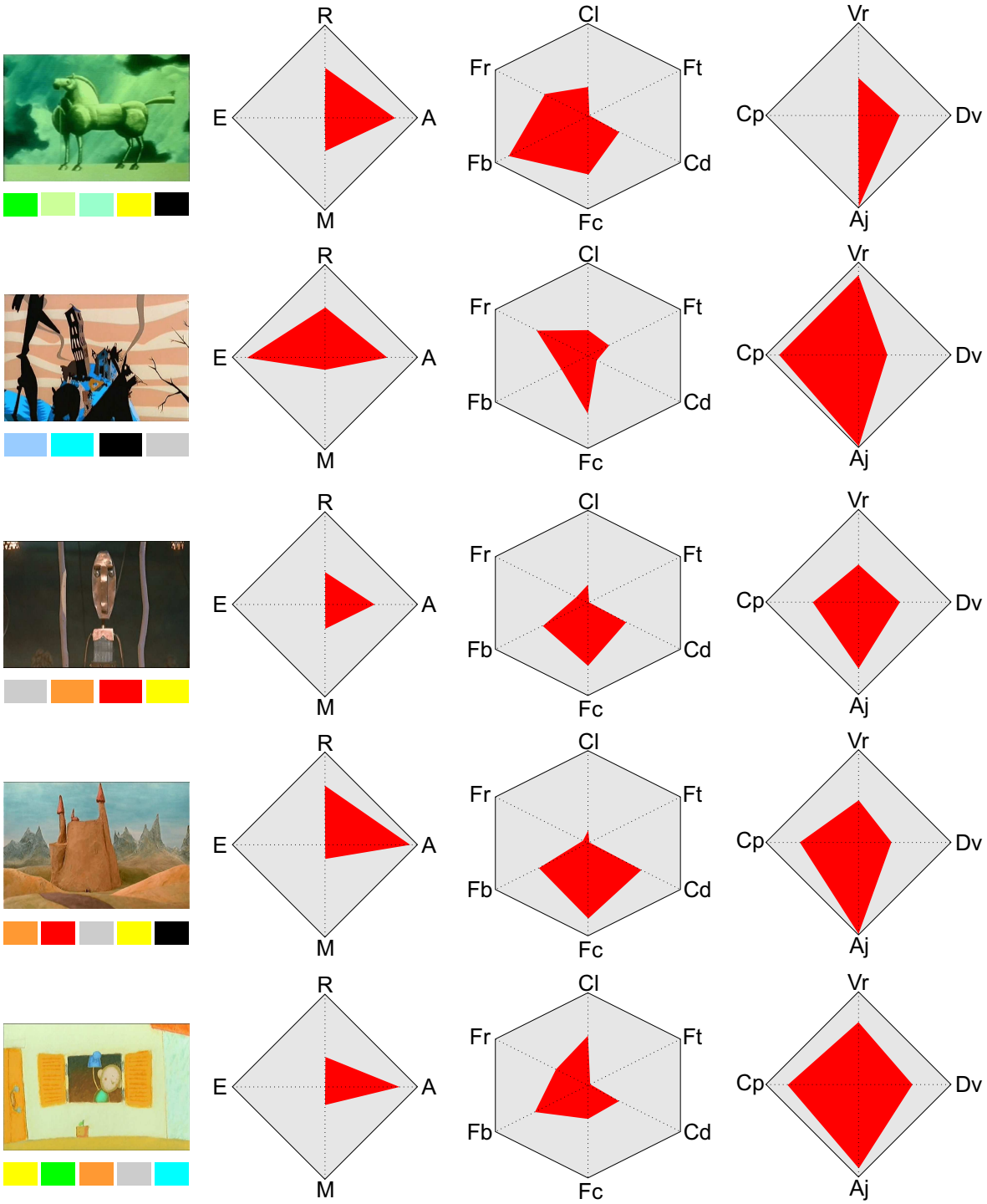


FIG. H.2 – Les gamuts sémantiques G^p, G^c et G^{rl} (voir la Section 7.6) obtenus pour les films suivants (ordre de haut en bas) : "Tamer of Wild Horses", "François le Vaillant", "La Cancion du Microsillon", "Le Château des Autres" et "Le Trop Petit Prince" (les couleurs élémentaires prédominantes de chaque film sont illustrées en dessous de l'image).

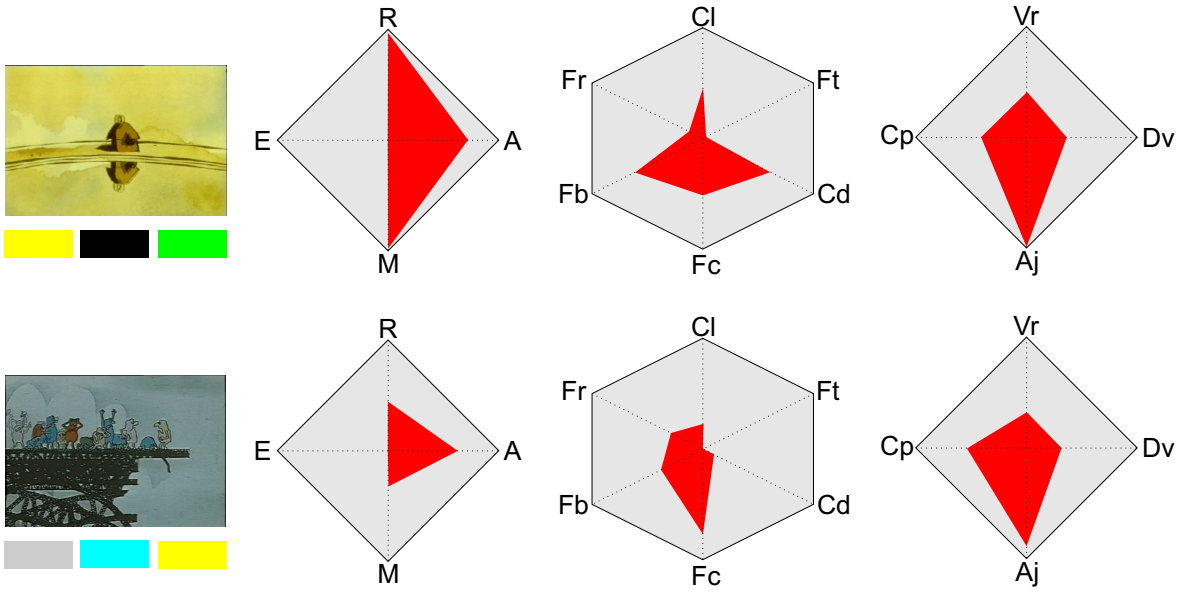


FIG. H.3 – Les gamuts sémantiques G^p, G^c et G^{rl} (voir la Section 7.6) obtenus pour les films suivants (ordre de haut en bas) : "Le Moine et le Poisson" et "Och, och" (les couleurs élémentaires prédominantes de chaque film sont illustrées en dessous de l'image).

Annexe I

Le logiciel : "Animation Movie Analysis Tool"

Les travaux effectués pendant cette thèse nous ont amenés à développer un certain nombre d'outils logiciels dédiés au traitement des séquences d'images. Le projet a eu comme point de départ la plateforme du logiciel "open source" VirtualDub 1.4.12 de Avery Lee, développé sous Microsoft Visual C++. VirtualDub est un logiciel qui permet d'effectuer les traitements de base sur des vidéos au format MPEG1 (navigation, montage, amélioration, etc.). Suite aux limitations constatées sur cette plateforme de développement (les plus importantes étant la difficulté de programmation, le manque d'une interface facilement accessible, le manque de compatibilité avec les standards MPEG2 ou MPEG4, des fonctions proposées ne répondant pas aux besoins d'un système d'indexation) nous avons transféré le projet sous Borland C++ Builder.

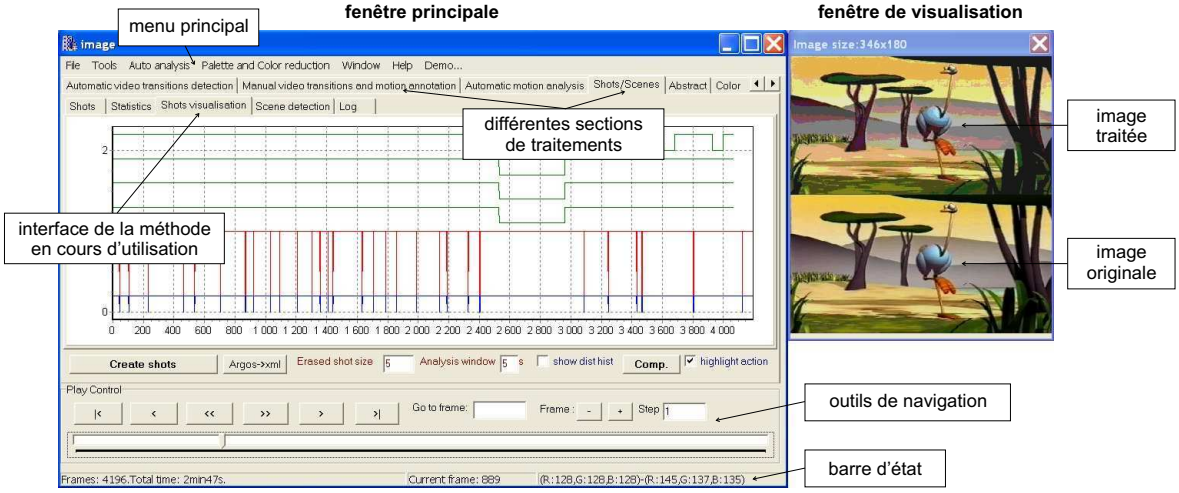


Fig. I.1 – L'écran principal du logiciel "Animation Movie Analysis Tool".

Nous avons repris depuis le départ la construction d'une interface visuelle de traitement vidéo. Nous avons ensuite implanter dans cet environnement toutes les méthodes et algorithmes que nous avons proposés pendant cette thèse. Le résultat est le logiciel : "*Animation Movie Analysis Tool*" qui est d'abord un *outil permettant les différentes tâches spécifiques* au domaine de l'indexation vidéo, mais construit sous la forme d'une bibliothèque de fonctions ce qui en fait *une plateforme de développement et d'analyse*. Cet outil propose à l'utilisateur un

ensemble vaste et évolutif de fonctions de traitements spécifiques aux séquences d'images (des fonctions génériques de manipulation et de traitement vidéo jusqu'aux méthodes spécifiques au traitement de films d'animation afin de tester nos approches).

Dans la suite nous allons illustrer ses principales fonctionnalités. La fenêtre principale du logiciel est présentée par la Figure I.1. Le logiciel comporte plusieurs *sections de traitement* :

Segmentation temporelle

Cette section permet le découpage en plans de la séquence par la détection des transitions vidéo et des effets des couleurs. Elle propose les outils suivants (les méthodes ont été présentées dans le Chapitre 2, voir deux exemples dans la Figure I.2 et I.3) :

- la détection des "cuts" avec les méthodes "4histogrammes" (parcours séquentiel et adaptatif) et "2derivées",
- la détection des "fades" ("fade in" et "fade out"),

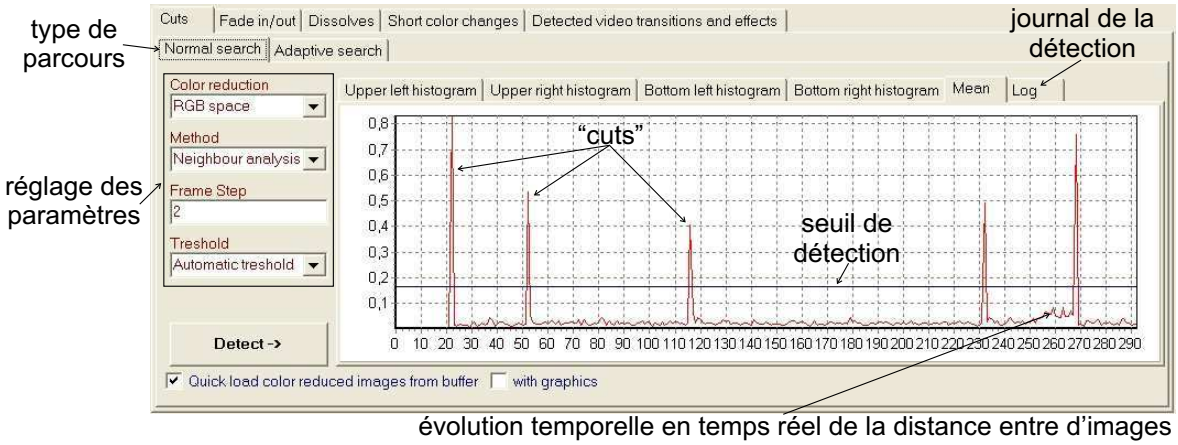


FIG. I.2 – Écran de détection de "cuts".

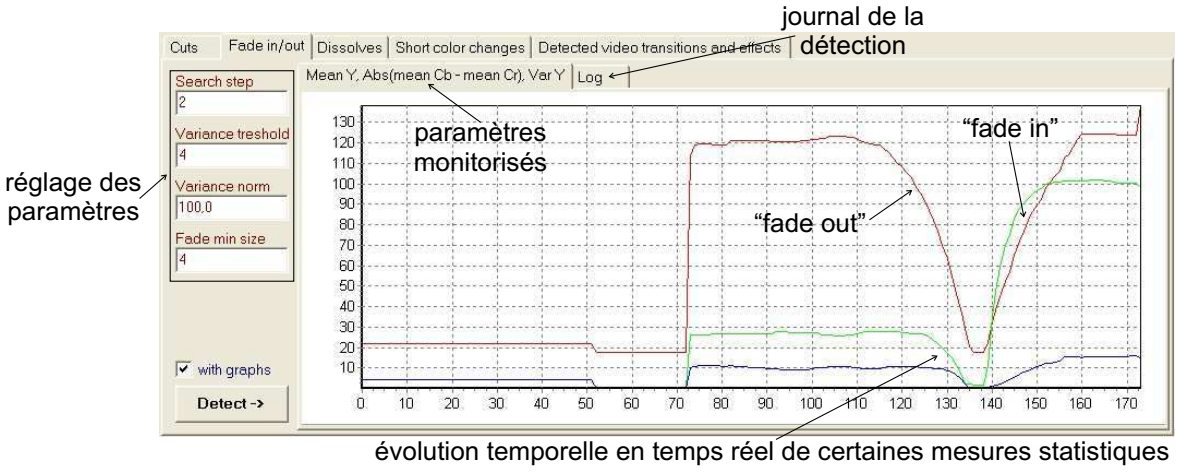


FIG. I.3 – Écran de détection de "fades".

- la détection des "dissolves",

– la détection de changements bref de couleurs ou SCC,
– un outil permettant la centralisation de toutes les transitions vidéo de la séquence.

L'annotation manuelle de la séquence

Cette section permet la construction d'une vérité terrain effectuée par annotation manuelle de la séquence (annotation des transitions et du mouvement). Elle propose de plus le calcul automatique des taux de détection (rappel, précision, erreur de détection, etc.), la représentation graphique des résultats et un ensemble d'outils permettant l'évaluation manuelle des résultats (voir la Figure I.4).

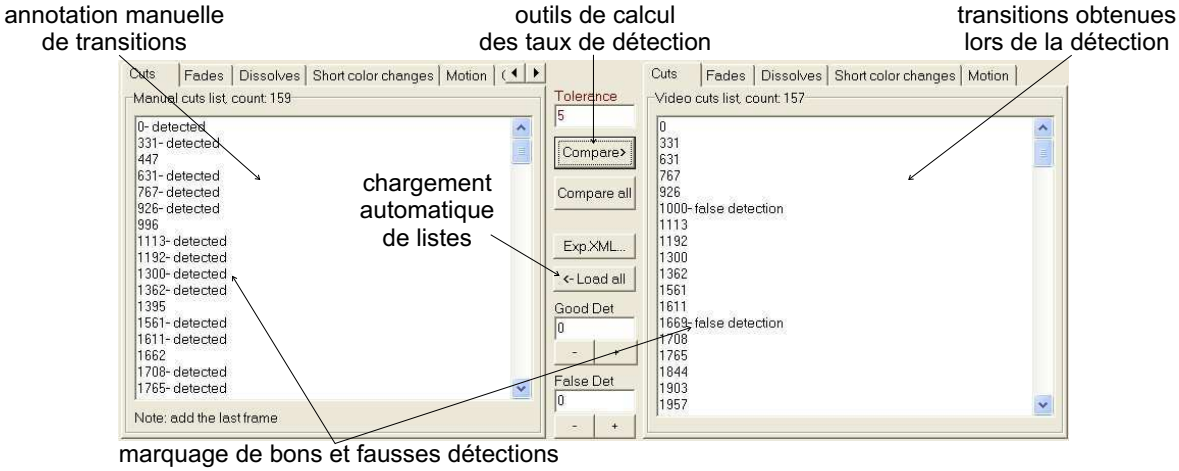


FIG. I.4 – Outil d'annotation manuelle.

L'analyse du mouvement

Cette section permet l'analyse et la caractérisation du mouvement global de la séquence (méthodes présentées dans le Chapitre 3). Les fonctionnalités proposées sont les suivantes (voir la Figure I.5) :

– estimation du mouvement basée sur l'analyse de blocs de pixels (recherche complète, recherche logarithmique et la recherche du standard H.263+),

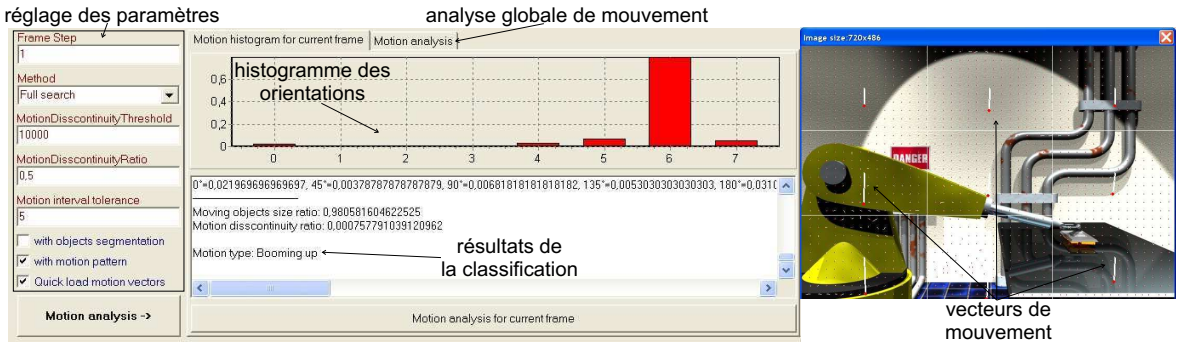


FIG. I.5 – Écran de l'analyse du mouvement.

– la classification du mouvement de la caméra pour l'image courante,
– la détection de segments de mouvement pour la séquence entière,
– la détection de "cuts" par la méthode "mdiscont".

L'annotation visuelle du contenu

Cette section permet d'analyser et de caractériser la structure temporelle de la séquence. Les méthodes proposées ont été présentées dans le Chapitre 2 (segmentation), Chapitre 5 (détection de scènes), la Section 7.3.1 (caractérisation sémantique) et la Section 6.2.1 (analyse de l'activité intra plan). Les fonctionnalités proposées sont les suivantes (voir la Figure I.6) :

– la construction des plans par agrégation des transitions vidéo détectées,
– l'analyse de l'activité intra plan par le calcul de l'histogramme des distances cumulées,
– la détection de passages d'action par l'analyse du rythme de la séquence,
– la caractérisation statistique et sémantique du contenu en terme d'action
– l'annotation visuelle des transitions,
– la détection de scènes.

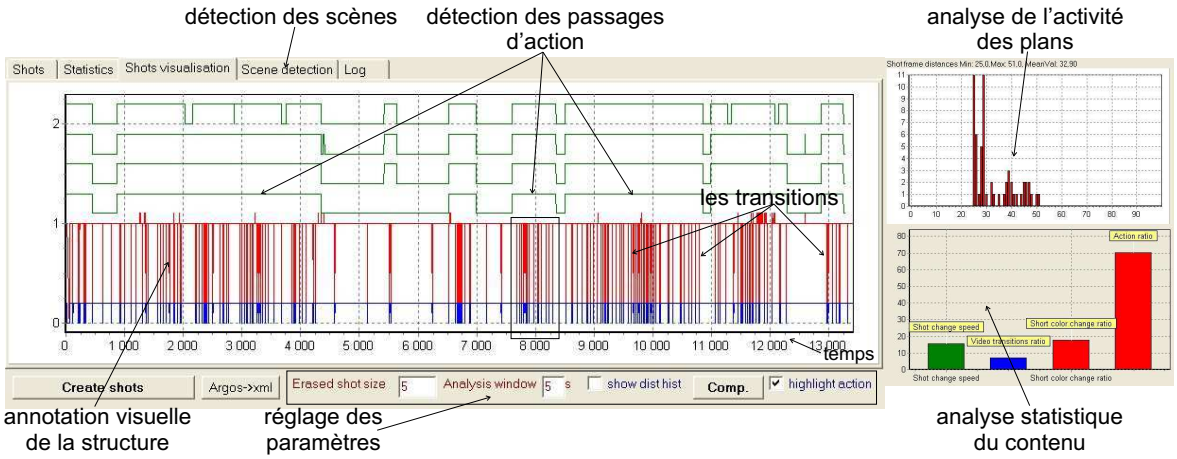


FIG. I.6 – Écran de l'annotation visuelle du contenu.

La construction de résumés

Cette section permet la construction de différents types de résumé de la séquence (les méthodes ont été présentées dans le Chapitre 6). Les fonctionnalités proposées sont les suivantes (voir la Figure I.7) :

– la construction de résumés statiques en images (une image par plan, résumé adaptatif par l'approche développée dans [Ott 05]),
– la construction de résumés dynamiques en mouvement (une sous séquence par plan, "bande-annonce"),
– la construction d'un résumé compact adaptatif en quelques images,
– l'analyse des résumés ainsi obtenus.

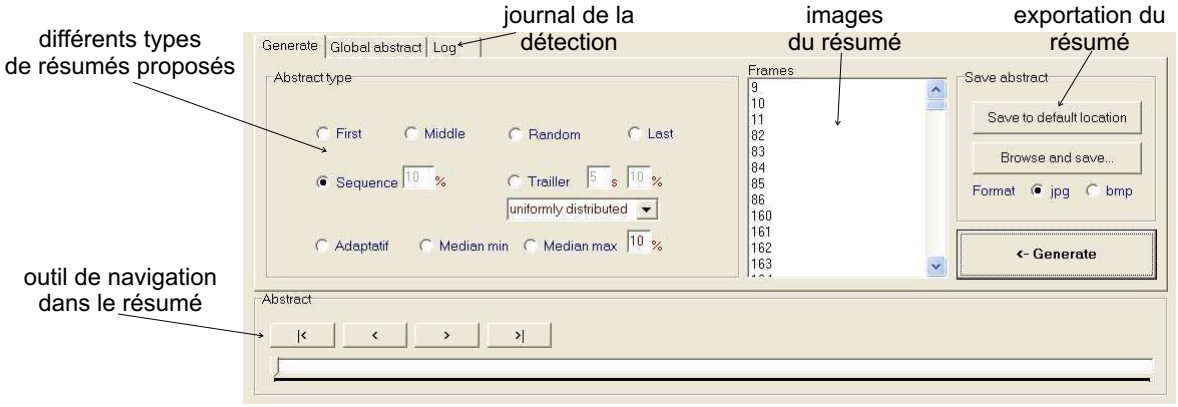


FIG. I.7 – L'écran de l'outil de construction de résumés.

L'analyse de couleurs

Cette section permet d'analyser la distribution des couleurs de la séquence (les méthodes utilisées ont été présentées dans le Chapitre 4 et Section 7.2). Les fonctionnalités proposées sont les suivantes (voir la Figure I.8) :

- le calcul de l'histogramme global pondéré pour l'analyse de la distribution globale des couleurs,
- le calcul de l'histogramme des couleurs élémentaires,
- la caractérisation statistique et sémantique de la distribution des couleurs.

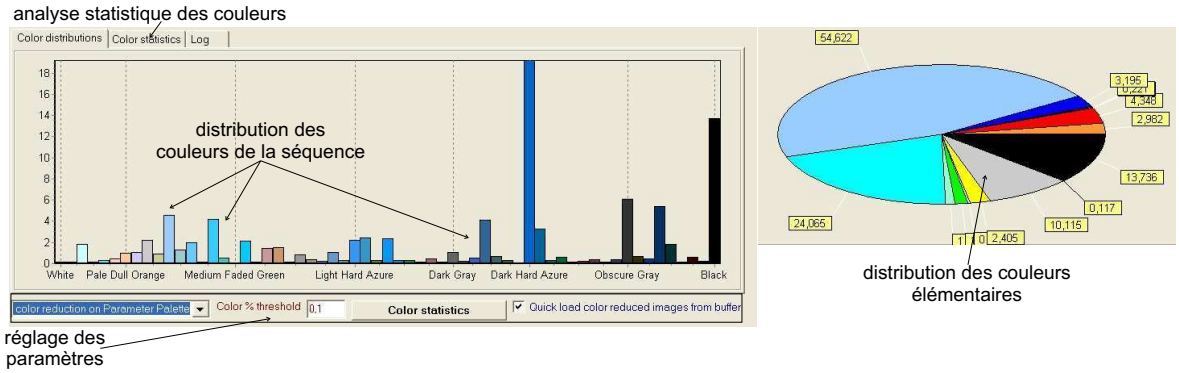


FIG. I.8 – Écran de l'outil d'analyse de la distribution des couleurs.

Autres outils

Nous avons également proposé un ensemble d'outils et de fonctions de traitement pour aider à l'analyse (voir quelques exemples dans la Figure I.10) :

- un outil de traitement automatique à partir de scripts. A l'aide d'un langage dédié il permet d'exécuter et de sauvegarder les résultats pour une liste de tâches de traitement,
- un outil de définition et d'analyse des palettes de couleurs utilisées dans la réduction des couleurs,
- un outil de visualisation de la palette des couleurs utilisées par l'image en cours,

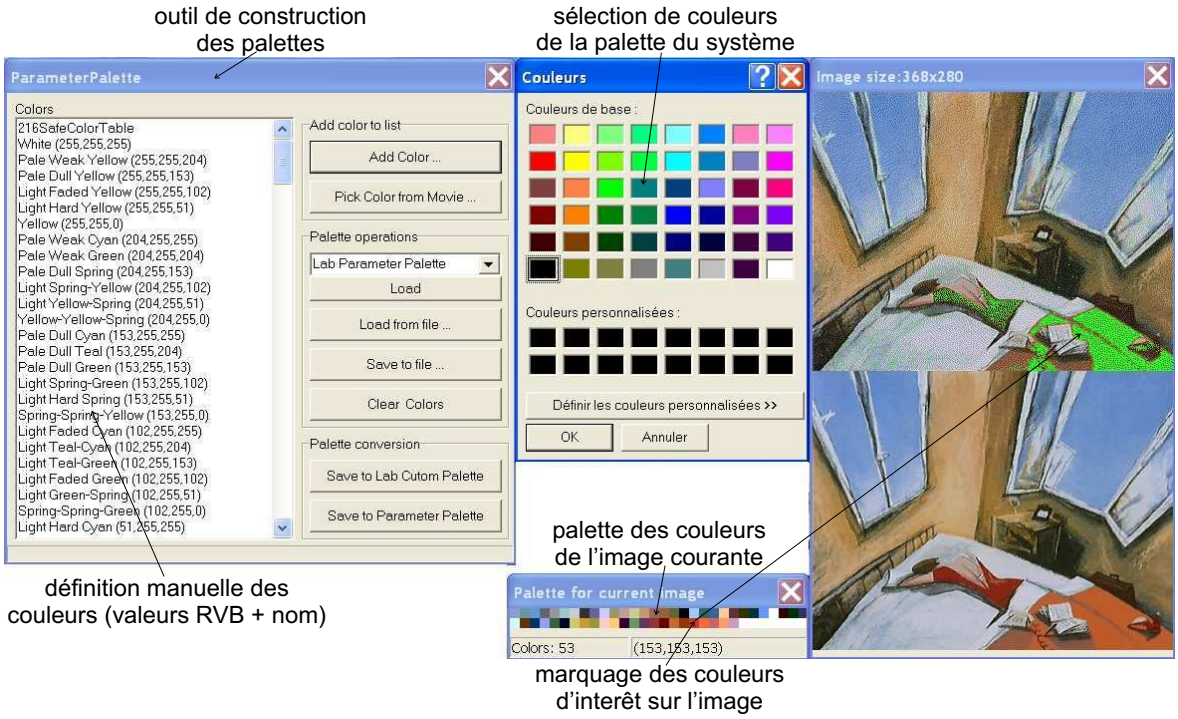


FIG. I.9 – Outil de construction et d'analyse des palettes.

- différents algorithmes de réduction des couleurs (voir Section 2.4.2),
- un filtrage médian en niveaux de gris et couleurs,
- une détection de contours,
- la conversion dans différents espace de couleurs : XYZ, YCbCr, TSL, TSV, Lab et la séparation des différentes composantes de l'espace,
- un outil de démonstration facilement accessible pour l'utilisateur (les paramètres sont réglés automatiquement).

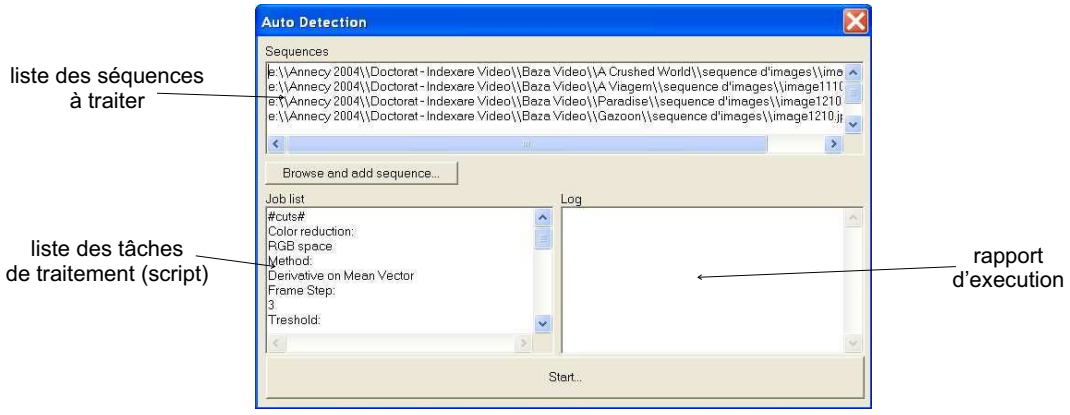


FIG. I.10 – Outil de traitement automatique à partir de scripts.