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).

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

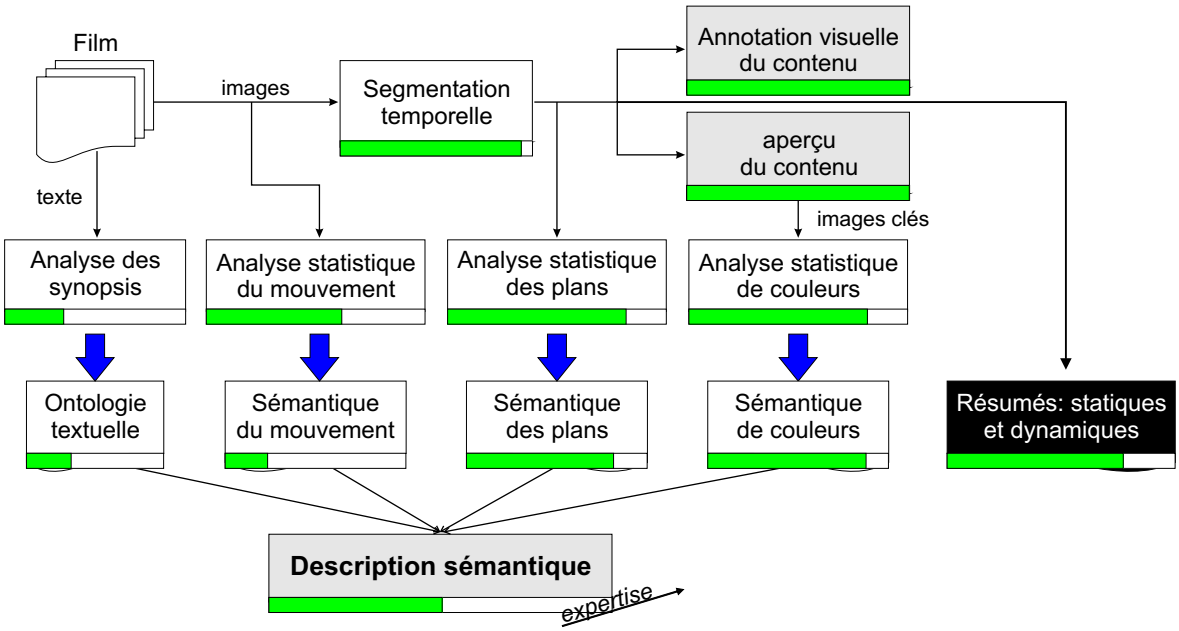

FIG. 9.1 – L'avancement des nos travaux (la ligne verte indique la quantité de travail accomplie).

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).

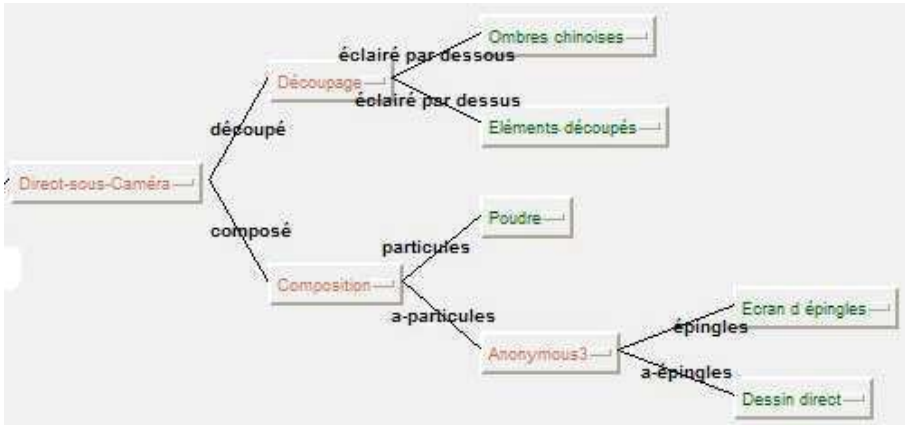

Fig. 9.2 – Extrait de l'ontologie des techniques.

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.