HABILITATION A DIRIGER DES RECHERCHES

Spécialité : Sciences et Technologies de l'Information

Contribution à la comparaison d'images pour l'évaluation des traitements, la reconnaissance de formes et l'indexation des séquences d'images

Didier COQUIN

Soutenue à Annecy-le-Vieux, le 6 décembre 2007, devant :

Jenny Benois-Pineau, Professeur des Universités, LABRI – Université de Bordeaux 1 (rapporteur)

Philippe Bolon, Professeur des Universités, LISTIC - Université de Savoie (examinateur)

Jean-Michel Jolion, Professeur des Universités, LIRIS – INSA de Lyon (examinateur)

Annick Montanvert, Professeur des Universités, GIPSA-Lab - UPMF (rapporteur)

Sylvie Philipp-Foliguet, Professeur des Universités, ETIS – ENSEA Cergy-Pontoise (rapporteur)

Thierry Pun, Professeur des Universités, Computer Vision and Multimedia Laboratory – Université de Genève (examinateur)

Polytech'Savoie
Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance
Université de Savoie

LISTIC Polytech'Savoie
laboratoire d'informatique systèmes, traitement de l'information et de la connaissance

UNIVERSITE CHAMBERY ANNECY DE SAVOIE

Remerciements

Je tiens tout d'abord à remercier Jean-Michel Jolion, Professeur à l'INSA de Lyon, délégué général du PRES (Pôle de Recherche et d'Enseignement Supérieur) à l'Université de Lyon, de m'avoir fait l'honneur de présider le jury et d'avoir mis en valeur mes contributions.

Que madame Jenny Benois-Pineau, Professeur à l'Université de Bordeaux I, ainsi que madame Sylvie Philipp-Foliguet, Professeur à l'Ecole Nationale Supérieure de l'Electronique et de ses Applications de Cergy Pontoise, trouvent ici le témoignage de ma gratitude pour avoir accepté d'examiner ce travail et d'en être rapporteurs. Leurs remarques, conseils et échanges sur la comparaison d'images et l'indexation des séquences d'images ont été pour moi d'une grande utilité.

Je veux également remercier Annick Montanvert, Professeur à l'Université Pierre-Mendès-France à Grenoble, d'avoir accepté d'être rapporteur de ce travail. Son analyse, ses remarques constructives et pertinentes ont été très enrichissantes et ont permis l'amélioration de ce mémoire.

Je remercie également Thierry Pun, Directeur du laboratoire Computer Vision and Multimedia, Professeur à l'Université de Genève, d'avoir accepté d'examiner ce travail. Ses remarques sont certainement la source de travaux futurs.

Je souhaiterais enfin adresser mes remerciements à Philippe Bolon, directeur du LISTIC, Professeur à l'Université de Savoie. Par ses compétences mais aussi la confiance qu'il m'a accordée, il m'a donné l'envie d'apprendre et de progresser. Durant ces années de collaboration, il a su se montrer disponible.

Merci à tous ceux qui ont contribué d'une manière ou d'une autre, à créer au sein de l'équipe *Traitement de l'Information*, un cadre de travail propice à l'étude et à la réalisation de ce travail. Je remercie particulièrement, Sylvie Galichet, Patrick Lambert, Lionel Valet, Gilles Mauris, Eric Benoit, Emmanuel Trouvé et Reda Boukezzoula pour leurs conseils avisés.

Avec une certaine émotion, j'exprime mes plus profonds remerciements à mes parents qui ont su me donner toutes les chances pour réussir. Qu'ils trouvent dans la réalisation de ce travail l'aboutissement de leurs efforts ainsi que l'expression de ma reconnaissance. A ma mère qui nous a quittés hélas trop vite ...

Merci également à notre fille Lou, qui a su se "passer" de son papa durant les soirées, week-end et une grande partie des vacances d'été.

Si je ne devais remercier qu'une personne ce serait sans hésitation Sophie. Je ne la remercierai jamais assez pour son indéfectible soutien. Je lui dédie ce travail et je me réjouis chaque jour qu'elle ait accepté de partager ma vie.

Table des matières

Première partie

Document de présentation

Chapitre 1

Document de Présentation

1.1 Curriculum Vitae

État Civil

Nom : COQUIN	Laboratoire LISTIC
Prénom : Didier	Tel : 04 50 09 65 46
Né : le 22 Novembre 1961 à Rennes (35)	`didier.coquin@univ-savoie.fr`
Situation personnelle : Marié, 1 enfant	`http ://www.listic.univ-savoie.fr`

Formation et Diplômes

09/87 - 09/91	**Thèse de doctorat en Traitement du Signal et Télécommunications :** "Segmentation et analyse d'images pour la classification automatique : application au Zooplancton". LASTI, ENSSAT, Lannion, Université de Rennes I. Mention très honorable. Directeur de thèse Michel Corazza et co-directeur Kacem Chehdi
09/86 - 07/87	**DEA de Traitement du Signal et Télécommunications**, Université de Rennes I, Mention Bien. Stage de recherche effectué à l'ENSTBr sous la responsabilité de C. Roux. Titre : "Extraction automatique des contours cellulaires de l'endothélium cornéen".
09/85 - 07/86	**Maîtrise EEA**, Université de Rennes I, Mention AB.
09/84 - 07/85	**Licence EEA**, Université de Rennes I.
09/81 - 07/84	**Deug A : Mathématiques et Physique**, Université de Rennes I.

Emplois occupés

09/87 - 09/88	Assistant Associé à l'ENSSAT Lannion, Université de Rennes I.
10/88 - 09/90	ALER à l'ENSSAT Lannion, Université de Rennes I.
10/90 - 08/92	ATER à l'ENSSAT Lannion, Université de Rennes I.
Depuis oct. 92	Maître de conférences à l'IUT d'Annecy : $61^{\text{ème}}$ section. de 1992 à 1998 : au sein du département GEii depuis sept. 1998 : au sein du département R&T (Réseaux et Télécommunications).

Recherche au Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance (LISTIC), Polytech'Savoie site d'Annecy, Université de Savoie.

Principales responsabilités

Recherche

Membre du comité de programme du congrès annuel SPIE International Symposium on Optomechatronic Technologie (Computer Vision Systems) depuis 2005 et co-chair de ce congrès.

Membre du comité de pilotage de l'action 3 : Systèmes Complexes pour l'Analyse et le Traitement d'Images (SCATI) avec Régis Clouard et Rémy Mullot, du thème B : (Image et Vision) du GDR ISIS, depuis Juillet 2007.

Enseignement

Responsable du module Algorithmique et Programmation C en 1ère année au département GEii de 1995 à 1998.

Responsable des modules de Mathématiques en 2ème année (Statistiques et Probabilités, Analyse-Vectorielle et Série) au Département GEii de 1996 à 2001.

Responsable des Cours-TD-TP de Télécommunications de première année de 1998 à 2005.

Responsable des modules de Télécommunications : T1 (Fondamentaux de télécommunications et transmissions : 60h) et T3 (Téléphonie : 30h) de 1ère année en R&T depuis septembre 2005.

Responsable du module de Maths (Statistiques et Probabilités : 30h) de 2ème année en R&T depuis septembre 2004.

Responsabilités administratives et collectives

depuis 1995 : Membre élu de la Commission de Spécialistes 61ème, puis de la commission de spécialistes 61ème/63ème sections puis de la 27ème/61ème sections de l'Université de Savoie.

de 1998 à 2005 : Membre nommé à la Commission de Spécialistes 61ème section de l'INPG.

de 2002-2005 : Membre élu au Conseil d'Administration de l'IUT d'Annecy.

de 1996 à 2000 : Adjoint du directeur du DEA d'Automatique Industrielle, associé à l'école doctorale des Sciences pour l'Ingénieur EEA, en collaboration avec l'INSA de Lyon, l'Université Claude Bernard Lyon I, et l'Ecole Centrale de Lyon.

depuis 1992 : Responsable des admissions (recrutement à l'IUT dans les départements GEii et R&T).

depuis 2003 : Coordinateur national du Groupe Télécom, pour l'ensemble des 28 départements d'IUT en Réseaux et Télécoms de France et Métropole.

Membre du comité de programme du 1er Workshop pédagogique Réseaux&Télécoms, 12-16 Novembre 2007, Saint Pierre, Ile de la Réunion, (http ://workshop.iut-rt.net/)

1.2 Résumé des activités de Recherche

1.2.1 Organisation générale

Mon activité de recherche fait apparaître 2 périodes principales :

– **Sept. 1986 - Sept. 1992** : cette période couvre celle de mon DEA [RD-1] (stage effectué à ENST de Bretagne au sein de l'équipe de Traitement des images sous la

responsabilité de C. Roux et G. Cazugel), de ma thèse [RD-2] et de ma deuxième année d'ATER. Mon activité, effectuée au LASTI (ENSSAT Lannion, Université de Rennes I), concernait l'analyse des images non texturées et les méthodes de classification. Ces travaux ont été réalisés en collaboration avec IFREMER et ont donné lieu à 3 publications internationales [CI-1], [CI-2], [CI-3] et 1 publication nationale [CN-1].

– **Depuis Sept. 1992** : A mon arrivée à l'Université de Savoie (UdS), je me suis intégré au sein de l'équipe "Traitement d'Images" du LAMII. Puis, en Janvier 2003, le LISTIC est né de la fusion du LAMII et du LLP. Je travaille désormais au sein de l'équipe "Traitement de l'Information" du LISTIC qui étudie les aspects de **fusion** pour la maîtrise de l'information. Durant ces 15 années (1992-2007), j'ai encadré 8 stagiaires DEA-Master et co-encadré deux Thèses et demie. Mes travaux de recherche s'articulent autour du thème général qu'est **la comparaison des images** pour l'analyse et l'évaluation des méthodes de traitement, pour la reconnaissance des formes et enfin, pour l'indexation de séquences vidéos.

1.2.2 Présentation des principaux axes

Contexte et objectif

Début des années 1990, les développements en analyse d'images sont essentiellement effectués sur des images en niveau de gris. Cette situation est en grande partie due à la prépondérance des caméras monochromes (les caméras couleur tri-CCD ont un coût très élevé) et à la puissance limitée des moyens de calculs. Les images couleurs, et plus généralement les images multi-composantes (un pixel est caractérisé par un vecteur d'attributs) sont encore peu exploitées. Durant ces années de recherche, mes travaux ont suivi l'évolution de la technologie en matière de caméra et également en terme de puissance de calcul des ordinateurs. J'ai d'abord travaillé sur des images en niveau de gris puis, sur des images de profondeur-réflectance (images multi-composantes) puis, sur des images couleurs et enfin, sur des séquences d'images couleurs.

Un problème d'intérêt théorique et pratique en traitement d'images est la comparaison de deux images de même nature. Comparer des images cela signifie établir des ressemblances ou des différences entre les images. Mais selon le type d'image et selon l'utilisation que l'on souhaite en faire, les techniques de comparaison sont bien différentes. Parmi les outils permettant d'étudier les formes ou les structures, je me suis intéressé aux techniques issues de la géométrie discrète. Ce choix se justifie par la volonté d'adapter l'outil de mesure à la nature discrète des images. J'ai développé une mesure de dissimilarité qui prend en compte les déplacements spatiaux et les déplacements radiométriques. Cette mesure permet d'analyser la performance d'un traitement de manière locale ou globale et de s'adapter à la forme du pixel ou du voxel.

J'ai choisi de développer mes activités de recherche autour du thème de **la comparaison d'images** en vue d'évaluer les traitements ou bien de reconnaître des formes et à un plus haut niveau, d'indexer des séquences d'images.

Contributions et principaux résultats obtenus

J'ai choisi de privilégier plus particulièrement quatre axes de recherche :

Premier axe : Mesure de dissimilarité reposant sur des opérateurs locaux

de distances : Lorsque les conditions d'acquisition de l'image sont bonnes, ou lorsque l'on a pu utiliser un bon opérateur de prétraitement, il est possible de faire, par exemple, une segmentation de l'image. L'analyse de cette image nécessite une caractérisation des entités ainsi mises en évidence. La difficulté de cette analyse provient de la nature discrète de l'image numérique qui entraîne une déformation des structures observées par rapport à leur apparence réelle dans l'espace continu. Une autre source de difficulté étant la forme du pixel qui, à l'époque, était rectangulaire et non carrée lors d'acquisition d'images par des systèmes de vision industrielle standards. Les outils d'analyse devaient donc tenir compte de cette source d'anisotropie supplémentaire. Nos efforts ont alors porté sur la mise au point d'opérateurs de distance et sur l'étude quantitative des effets de discrétisation de structures continues.

Les opérateurs locaux de distances constituent un moyen rapide, peu consommateur de mémoire, pour le calcul approché des distances euclidiennes entre objets observés au moyen d'une image numérique. Le principe consiste à approximer le calcul de la distance euclidienne entre deux pixels en considérant que le chemin minimal est formé de déplacements élémentaires auxquels sont affectés des coefficients de pondération. Le calcul de distance nécessite alors que des opérations locales d'addition dans un voisinage. Il est donc possible d'obtenir simplement et rapidement des cartes de distances. En reprenant la démarche proposée par Borgefors [Borgefors 86] qui consiste à considérer des déplacements continus sur une trajectoire de référence, mais en l'adaptant au cas de trajectoires de références circulaires, il est possible d'optimiser le choix des coefficients au sens de la minimisation de l'écart maximal entre distance euclidienne et distance calculée par l'opérateur local.

Dans le cas de maillages rectangulaires, nous avons déterminé les expressions analytiques des coefficients optimaux en fonction de la largeur relative du pixel et ce, pour des opérateurs 3x3, 5x5, et 7x7 [CN-2], [R-1]. Pour un maillage carré, on retrouve les valeurs proposées par Verwer [Verwer 91]. En utilisant des opérateurs en nombre entier, on accélère les traitements et on limite la quantité de données à stocker. Il faut pour cela choisir un facteur d'échelle pour lequel la perte de précision reste faible. L'intérêt de cette approche est de pouvoir effectuer directement une **analyse de forme** (par étude du squelette ou de l'axe médian par exemple) sans devoir procéder à un rééchantillonnage de l'image.

Nous avons étendu ce mode de calcul en 3D en considérant des maillages parallélépipèdiques. Les coefficients optimaux d'un opérateur 3x3x3 ont été déterminés et l'erreur par rapport à la distance euclidienne analysée [CI-4]. L'optimisation de l'opérateur local de distance réalisée en 2D puis en 3D nous a amenés à généraliser notre approche à des opérateurs locaux de distance cubiques et non cubiques en maillage parallélépipédique. En adoptant la même démarche, nous avons déterminé les expressions analytiques des coefficients optimaux en fonction des paramètres du voxel, pour les opérateurs cubiques NxN et NxNxN et non cubiques MxN puis MxNxV [CI-5]. Les performances de ces opérateurs sont analysées et comparées. L'intérêt de travailler avec des opérateurs non cubiques réside dans le gain de temps pour des performances similaires.

Ces travaux ont abouti à la soutenance de la thèse de Yousra Chehadeh en septembre 1997 (co-encadrement à 50% avec Ph. Bolon) [TH-1] qui a fait l'objet de 3 communications internationales et 2 communications nationales.

L'avantage de notre approche est l'extension possible à des espaces de dimension plus élevée. Dans [DEA-3] nous avons donné les expressions analytiques des coefficients d'un masque cubique 3x3x3x3x3. Son exploitation demandait trop de temps et les données étaient volumineuses à sauvegarder. Pour l'appliquer aux images couleurs, nous sommes passés du

domaine 5D (x, y, R, V, B) trop "volumineux", au domaine 3D (x, y, C) en utilisant un réseau de neurones de Kohonen unidimentionnel circulaire - la variable C représentant une pseudo-couleur. Ce réseau possède la particularité de conserver les propriétés de voisinage en passant d'un espace à l'autre. La transformation fournit une image dans laquelle les pseudo-couleurs ne sont pas équidistantes. Nous avons donc développé un opérateur local de distance non-stationnaire qui possède la particularité de s'adapter à la hauteur entre ces pseudo-couleurs [CI-8] et [CI-9].

Ces différents opérateurs ont été utilisés pour la **comparaison d'images** et l'évaluation de traitements. Ils donnent des résultats encourageants.

Deuxième axe : Evaluation des performances des méthodes de traitements d'images : L'élaboration de traitements évolués passe par la mise au point d'outils de **comparaison** et d'**évaluation** de performances aux niveaux local et global. A partir de l'expérience acquise depuis plusieurs années sur les espaces discrets, nous avons proposé une méthode qui permet de comparer quantitativement des images en niveau de gris, en calculant un critère de dissimilarité [R-2]. Un tel critère permet alors de mesurer l'écart entre le résultat d'un traitement et une référence, ou de comparer entre eux les résultats de plusieurs traitements. Ce critère repose sur le calcul de la distance de Baddeley étendue aux cas des images en niveau de gris. La mesure de dissimilarité, établie à partir de nos opérateurs de distances, nous a permis d'analyser les performances de différents traitements (filtrage et segmentation). Cela nous a permis de mettre en évidence l'influence du filtrage sur les contours de l'objet, là où la distorsion est plus importante. Par rapport aux mesures classiques telles que le PSNR, la mesure introduite ici prend mieux en compte les informations de transitions entre régions. Nous avons alors étudié les propriétés de la mesure de dissimilarité pour l'évaluation de différents types de filtrage. Cette méthode a été comparée à 5 autres mesures objectives de dissimilarité. Nous avons montré que notre opérateur de dissimilarité était stable vis-à-vis du bruit, permettait de discriminer les différents types de filtrage et était tolérant à de petites variations de formes consécutives au filtrage [R-2], [CI-7], [CN-4], [CN-5].

Afin de **comparer** les images couleurs entre elles, en terme de variation de l'intensité lumineuse dans un espace à deux dimensions (le support XxY), nous avons adapté la mesure de dissimilarité aux images couleurs. Nous avons utilisé pour cela un opérateur local de distance non-stationnaire, possédant la particularité de s'adapter à la hauteur entre les pseudo-couleurs. Nous avons étudié les propriétés de cette nouvelle dissimilarité dans le cas d'un déplacement linéaire d'objet dans l'image, mais aussi dans le cas d'une variation radiométrique linéaire de l'une des trois composantes (luminance, intensité et saturation) de l'image couleur. Notre dissimilarité réagit pratiquement linéairement contrairement au critère RMS [CI-8] et [CI-9].

Nous avons également travaillé sur l'**évaluation** des méthodes de segmentations en utilisant les informations données par des indicateurs de performances et en utilisant une méthode d'évaluation basée sur un système de neurones flous. Nous avons constaté que le comportement des méthodes de segmentation est très différent (d'une méthode à l'autre) car il tient compte de la complexité des images utilisées (nombre d'objets dans l'image, distribution des couleurs, ...). En conséquence, les indicateurs de performances donnent des valeurs très variables, ce qui rend l'interprétation encore plus délicate. Ayant pris en considération ces aspects, nous avons réalisé une méthode d'évaluation qui permet de résoudre la complexité de ce problème et d'interpréter les informations fournies par les indicateurs de performances sur la base de la puissance de calcul et de la capacité d'apprentissage des réseaux de neurones.

En effet, en imposant *a priori*, pour chaque image segmentée, une évaluation basée sur la perception humaine, nous pouvons exploiter la puissance du système d'évaluation neuronal et sa capacité d'expertise par rapport à la décision initiale. Des tests ont également montré qu'un réseau de neurones flous permet d'améliorer significativement la vitesse de convergence par rapport à un réseau de neurones classique [DEA-4].

Enfin, nous avons travaillé sur la segmentation asservie d'images couleurs, à l'aide d'indicateurs de performance flous. Les travaux de thèse de Nadia Bouloudani portaient sur l'évaluation symbolique des méthodes de traitements d'images. La description symbolique permet de donner une évaluation simplifiée de la qualité d'un traitement [CN-6].

Troisième axe : Reconnaissance de formes : Dans tout processus de reconnaissance de formes, une étape importante et incontournable est l'étape de décision ou d'identification. Elle consiste à affecter l'objet qui se présente à l'une des classes d'apprentissage. Là encore, l'étape de décision se fera le plus souvent par **comparaison** à un alphabet.

Le squelette et l'axe médian sont deux outils performants pour l'analyse de formes, puisqu'ils permettent de décrire les propriétés générales des objets et de réduire l'image initiale à une représentation plus compacte. Nous avons alors tout naturellement proposé une méthode d'extraction du squelette en utilisant des opérateurs de distance adaptés au maillage rectangulaire [CI-6] et [0-1]. Le squelette est extrait de la carte de distance par suivi des lignes de crêtes, ou par extraction de l'axe médian et connexion de ce dernier. Le squelette résultant est un ensemble de points pondérés. Le poids de chaque point représente sa distance au fond de l'image et cette information de distance est très utile pour la reconstruction.

Nous avons également travaillé sur la reconnaissance de gestes dynamiques de la main. Ce travail se situe dans le contexte du contrôle d'un poste de travail par reconnaissance des gestes de la main d'un opérateur. Des mesures primaires des angles des articulations des doigts de la main sont effectuées à l'aide d'un gant numérique. Au même niveau, un second module de traitement basé sur un système de vision vient enrichir la prise de décision, permettant en particulier, la prise en compte du mouvement global de la main. La caméra fait l'acquisition du geste mais ne peut, dans certaines situations, faire la distinction entre des positions différentes des doigts de la main. De même, le gant ne peut pas, par exemple, faire la distinction entre le pouce mis en haut ou en bas. La collaboration et la fusion des informations issues de ces deux capteurs (gant et caméra) permet de lever ces ambiguïtés.

La méthode que nous avons développée, dans la partie vision, repose sur la combinaison de deux approches : une première basée sur l'histogramme des orientations du gradient et une seconde utilisant la superposition des squelettes de la main et de l'avant-bras, calculée sur l'ensemble de la séquence. Dans notre application, un geste dynamique est caractérisé par une séquence comportant de 30 à 50 images. L'histogramme des orientations du gradient est utilisé comme une signature statique calculée sur la première et la dernière image de la séquence. Ces signatures statiques permettent de délimiter le geste dynamique. La superposition des squelettes est utilisée comme une signature dynamique calculée sur la séquence. Elle permet de résumer le geste en une seule image. En revanche, elle ne permet pas de traduire l'information sur la chronologie du geste.

Le principe de reconnaissance du geste repose sur la **comparaison** entre les signatures du geste courant et les signatures de séquences d'apprentissage correspondant à un alphabet de gestes connus. Le geste reconnu est associé à la distance la plus petite. Pour la signature statique nous avons utilisé la distance euclidienne entre les histogrammes des orientations du gradient. Pour la signature dynamique, comme nous devions comparer deux images binaires représentant la superposition des squelettes de la séquence courante et des séquences

d'apprentissage, nous avons utilisé la distance de Baddeley, précédemment développée.

Nous avons développé un système qui permet de fusionner les mesures issues des deux sources (gant numérique et système de vision). Le gant donne des informations sur la position des doigts et de la main alors que système de vision donne des informations sur le mouvement général du bras. Ces mesures peuvent être redondantes ou complémentaires, nous avons alors défini une stratégie pour fusionner ces résultats et prendre une décision finale. Ce système a été implémenté pour commander un mini-robo sur un alphabet de 10 gestes. Il fonctionne en temps réel. Ces travaux ont fait l'objet de plusieurs publications [R-4], [R-3], [CI-15] et [CN-7].

Quatrième axe : Caractérisation et Indexation de séquences vidéo : Cette dernière thématique, récemment abordée, est actuellement une préoccupation importante de la communauté Image. La problématique liée aux images statiques s'est ainsi transposée dans des termes assez semblables aux séquences vidéo. Cependant, la difficulté est accrue car il faut prendre en compte l'aspect temporel et faire face à des tailles de données encore plus importantes. Les récents travaux sur les séquences vidéo se sont essentiellement concentrés sur le découpage en plans, le résumé automatique, l'analyse du mouvement et le suivi d'objets caractéristiques, la recherche par l'exemple et la navigation. Les travaux développés s'inscrivent dans cette perspective, la spécialité tenant à la recherche de caractéristiques de nature sémantique, domaine délicat et encore peu abordé dans la littérature. Les recherches engagées portent sur l'extraction d'attributs (analyse bas-niveau) et la représentation symbolique de ceux-ci (analyse haut-niveau). Là encore, la **comparaison** entre images, entre caractéristiques issues des images, des plans et des scènes, joue un rôle prépondérant pour l'indexation de séquences vidéo.

- 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 d'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 comporte de deux étapes : une première étape dont le degré de granularité est l'image, 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 de 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, avec deux objectifs principaux : la robustesse et le caractère automatique, pour réduire le plus possible l'intervention humaine dans le système. Ce niveau a permis de construire des caractéristiques liées au rythme et à 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, ceci permettant de détecter les transitions. Pour l'analyse des couleurs, nous avons proposé une signature globale de la séquence prenant en compte son aspect temporel. Cette approche est basée sur l'utilisation d'une palette particulière de couleurs associée à un dictionnaire des noms de couleurs. Elle prépare ainsi l'analyse sémantique [CI-17], [CI-14], [CN-9], [CN-8].

- Analyse de haut-niveau : La détection de scènes et la construction de résumés constituent une sorte d'étape intermédiaire entre la caractérisation bas niveau et la descrip-

tion sémantique. Le développement de certaines mesures de similarité entre le contenu des plans, pour la **comparaison** de ceux-ci, nous 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 à la tâche de navigation. A ce niveau, notre apport consiste en la proposition d'un résumé intelligent, similaire à la "bande-annonce" d'un film, et en le développement d'une méthodologie permettant la construction de résumés compacts constitués seulement de quelques images représentatives de la séquence [CI-19].

- 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 traduction d'une séquence en 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 sur 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. Notre contribution principale réside 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 impliquant le jugement humain quant à leur pertinence. 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. Ces travaux ont abouti aux publications suivantes [CI-14],[CI-16],[CI-20],[CI-21] et [O-3]. Notons que ces travaux sont développés en coopération avec la CITIA (Cité de l'image en mouvement) qui nous a fourni une base de données de séquences vidéo issues du Festival Internationale du Film d'Animation. Il se déroule tous les ans à Annecy.

Dans le cadre de l'indexation vidéo, nous avons également mis en place une stratégie reposant sur des techniques de mesures de distances entre images par **comparaison** des histogrammes couleurs des images extraites des plans. A partir d'une distance entre histogrammes couleur des images, nous calculons, pour chaque image du plan, la distance cumulée aux autres images. L'analyse des histogrammes, de ces distances cumulées moyennes, permet une sélection des images appropriées au contenu du plan. Cette technique fournit un résumé compact permettant à l'utilisateur d'avoir une impression globale du film sans avoir à parcourir trop d'images [CI-21].

Enfin, nous avons proposé 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. Cette tâche est nécessaire au moteur de recherche d'une base de données vidéo. Les caractérisations acquises de chaque film sont illustrées par une représentation graphique inspirée de la construction des gamuts de couleurs d'un dispositif de restitution d'images couleurs (écran ou imprimante). Ainsi, pour chaque film, nous avons associé 3 gamuts sémantiques : un gamut des plans, un gamut des propriétés couleurs,

et un gamut de la richesse couleur et des relations entre couleurs. Ce type de représentation visuelle compacte permet de se faire une idée globale de 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 des films analysés. Les films ayant des caractéristiques sémantiques différentes auront des formes de gamuts sémantiques différentes et inversement. Une mesure de similitude peut être définie pour la comparaison de ces gamuts sémantiques.

L'ensemble de ces travaux a donné lieu à la publication de 5 articles dans des revues, 3 articles dans des ouvrages, 23 communications dans des congrès internationaux avec actes et comité de lecture, 8 communications dans des congrès nationaux avec actes et comité de lecture, et 3 articles dans la revue de l'Université Polytehnica de Bucarest.

Deux thèses ont été soutenues, celles de Yousra Chehadeh en 1997 [TH-1] et de Bogdan Ionescu en 2007 [TH-2] et 8 stages de DEA-Master ont été effectués.

1.2.3 Encadrements

Les travaux présentés sont la plupart du temps liés à l'encadrement de stages de DEA-Master et de Thèses

Thèses

[TH-1] Y. CHEHADEH : "Opérateurs locaux de distance en maillages rectangulaire et parallélépipédique : application à l'analyse d'images", Thèse de doctorat, Université de Savoie, **1er Octobre 1997**. Jury : D. Barba (président), A. Montanvert, F. Préteux (Rapporteurs), T. Redarce, Ph. Bolon, D. Coquin (co-encadrement à 50% avec Ph. Bolon).

[TH-2] B. IONESCU : "Caractérisation symbolique de séquences d'images : Application aux films d'animation", Thèse de doctorat, Université de Savoie, soutenance le **7 mai 2007**. Jury : T. Petrescu (président) M. Rombaut , C. Gordan (rapporteurs), C. Fermandez-Maloigne, V. Buzuloiu, C. Vertan, P. Lambert, D. Coquin. (co-encadrement à 50% avec P. Lambert)

[TH-3] N. BOULOUDANI : "Evaluation symbolique des méthodes de traitements d'images". Septembre 2001 à Juillet 2003 (démission pour raisons médicales). (co-encadrement à 50% avec P. Lambert)

La thèse de B. Ionescu a été réalisée en cotutelle avec l'Université POLITEHNICA de Bucarest. On peut noter que B. Ionescu est venu faire un stage de 5ème année en 2002, durant lequel il a travaillé sur la reconnaissance de gestes de la main. Ces thèses ont donné lieu à des publications (voir paragraphe 1.4), le doctorant apparaît en 1er auteur, si la publication est directement tirée de son travail.

Après la thèse, Y. Chehadeh a obtenu un emploi d'Ingénieur en informatique en région Parisienne, et B. Ionescu occupera un poste de maître assistant à l'Université POLITEHNICA de Bucarest à la prochaine rentrée universitaire (septembre 2007).

1.2.4 Stages de DEA et Master

[DEA-1] M. CIUC : "Méthodes d'extraction du squelette d'un objet en maillage rectangulaire ", DEA Automatique Industrielle, Université de Savoie, Juillet 1996. (encadrement

à 100%). M. CIUC a fait sa thèse au laboratoire en cotutelle avec l'Université POLITEHNICA de Bucarest, sur le traitement d'images multicomposantes : Application à l'imagerie couleur et radar. Il est depuis octobre 2002 maître assistant à l'Université POLITEHNICA de Bucarest.

[DEA-2] F. FILLION-ROBIN : "Traitement d'images de profondeur-réflectance", DEA Automatique Industrielle, Université de Savoie, Juillet 1998. (encadrement à 100%).

[DEA-3] A. ONEA : "Utilisation des opérateurs de distances pour la comparaison des images couleur", DEA Automatique Industrielle, Université de Savoie, Juillet 1999. (encadrement à 100%). A. ONEA a poursuivi en thèse à l'Université de Poitiers.

[DEA-4] Benone IONESCU : " Evaluation de segmentations avec indicateurs de performances en utilisant un système de classification neuronal", DEA Automatique Industrielle, Université de Savoie, Septembre 2001. (encadrement à 100%). Benone IONESCU travaille en tant qu'Ingénieur informaticien en France.

[DEA-5] M. BIARDEAU : " Poste de travail interactif basé sur le geste", DEA Automatique Industrielle, Université de Savoie, Juillet 2003. (encadrement à 100%). Après avoir travaillé, M. BIARDEAU est actuellement en thèse CIFRE avec la société SERT (fusion multi-capteur pour la coulée continue).

[DEA-6] O. DJAMIAI : "Segmentation d'images asservie à l'aide d'indicateurs de performance flous", co-encadrement (50%) avec P. Lambert, DEA Automatique Industrielle, Université de Savoie, Juillet 2003. O. DJAMIAI est retourné travailler dans son pays.

[DEA-7] S. PHIN : "Fusion d'informations issues d'un gant numérique et d'une caméra pour la reconnaissance de gestes", co-encadrement (50%) avec E. Benoit, Master ITI, Université de Savoie, Juillet 2004. S. PHIN est retourné travailler dans son pays.

[DEA-8] L. OTT : "Mesures de distance entre images, plans et scènes dans les films d'animation", co-encadrement (50%) avec P. Lambert, Master ITI, Université de Savoie, Juillet 2005. L. OTT est actuellement en thèse à l'Université de Strasbourg.

1.2.5 Participation à un Jury de Thèse

En tant qu'examinateur, j'ai été membre du jury de la thèse de : C. Gobinet : " Application de techniques de séparation de sources à la spectroscopie Raman et à la spectroscopie de fluorescence ", thèse de doctorat de l'Université de Reims, Champagne Ardenne. 27 mars 2006.

1.2.6 Animation et rayonnement scientifique

Animation

Nos travaux sur l'évaluation des performances en traitement et analyse d'images ont été présentés au niveau national à plusieurs reprises dans le cadre du GdR ISIS, GT3 " systèmes de segmentation " et " évaluation en traitement d'images ". Plus récemment, j'ai présenté nos travaux sur la reconnaissance de gestes lors des journées du GdR ISIS & AS 70 du département STIC du CNRS sur la "Perception, Modélisation et Interprétation du Geste Humain", les 27 et 28 Mars 2003, [J-5]. Nous participons également au GT 3.5 " Indexation et recherche d'informations Multimédia" dans le cadre de notre activité concernant l'analyse de séquences d'images. Nos travaux ont été présentés lors de la journée du 24 septembre 2004 [J-6], et au cours de l'Ecole d'Hiver sur l'Image Numérique Couleur [J-8]. A partir de Juillet

2007, je remplace P. Lambert au comité de pilotage de l'***action 3*** (**SCATI**) du thème B du GDR ISIS.

Sur le plan régional, de 1998 à 2000, puis de 2000 à 2003, j'ai participé, au niveau du LAMII, aux projets ACTIV 1 et 2 (Archivage Couleur Traitement d'Images et Vision).

J'ai participé au projet BQR (2002-2004) relatif à l'étude de postes de travail interactifs basés sur des mécanismes de reconnaissance de gestes et de postures. Une caméra fait l'acquisition du geste, la séquence est traitée, puis un module de reconnaissance interprète le geste permettant de guider un mini-robot mobile [CN-7], [CI-13], [CI-15], [R-3], [R-4].

J'ai également participé durant 2 ans (2002 à 2004), aux 2 projets BQR (1 an chacun) sur l'indexation ontologique des documents audio-visuels, pour la mise en place de méthodes permettant la caractérisation et l'analyse des films d'animation. L'objectif à plus long terme, que nous poursuivons, est de fournir des outils logiciels de recherche, de navigation ou d'exploitation dans une base de films numérisés (utilisation des péri-textes, en se basant sur des ontologies, et utilisation des images). Les données sont fournies par le CICA (Cité de l'image en mouvement) qui organise tous les ans à Annecy, le Festival International du Film d'Animation.

J'ai participé en 2006, au projet BQR sur un système coopératif de fusion d'informations pour l'interprétation d'images 3D, dans lequel j'interviens au niveau de l'évaluation de performances (évaluation de la qualité des résultats obtenus, et évaluation de la coopération entre l'homme et le système). Ces travaux ont permis de modéliser à l'aide de la méréotopologie, la coopération Homme/Système de façon à structurer et organiser l'ensemble des actions et interactions qui interviennent dans une coopération, travaux qui ont fait l'objet d'une publication au congrès international Information Fusion 2007 [CI-21].

Depuis début 2007, je participe au projet LIMA (Loisir et IMAge), du cluster ISLE (Informatique Signal Logiciel Embarqué, de la région Rhône Alpes. Ce projet s'attaque à deux problématiques complémentaires. La première concerne la caractérisation des séquences vidéo ou des films dans un but d'indexation servant à la recherche, la navigation ou l'exploitation de bases de données. La seconde est liée à la création de contenus graphiques 3D ainsi qu'à leur visualisation. Ces deux problématiques, souvent regardées comme distinctes, tendent à se rapprocher et l'un des objectifs du projet LIMA est de favoriser ce rapprochement.

Relations Internationales

Depuis plusieurs années, le laboratoire a engagé des actions de coopération de type bilatéral au niveau européen. Les opérations les plus actives concernent les échanges avec la Roumanie (Prof Buzuloiu) et avec l'Allemagne (Prof Bohner). Elles se sont traduites par des échanges bidirectionnels de chercheurs, doctorants et stagiaires, ainsi que par la participation croisée à des enseignements de type école doctorale ou séminaire. C'est dans ce cadre que :

- J'ai participé, en tant qu'invité, à 4 écoles de printemps ETASM, organisées par le professeur Buzuloiu, durant lesquelles j'ai assuré 12H de cours sur l'approche géométrique en analyse d'images : application à la comparaison d'images couleur, à l'Université POLITEHNICA de Bucarest.
- Des travaux, engagés depuis plusieurs années avec M. Bohner (Université de Sciences appliquées de Kaiserslautern en Allemagne), ont fait l'objet d'une publication acceptée pour le congrès International on Information Fusion'03 [CI-12]. Ces travaux portaient sur l'analyse des images de profondeur-réflectance. Nous avons développé un système expert pour fusionner les résultats des différentes segmentations issues des images de

réflectance et des images de profondeur.

Participation à des comités de lecture et d'organisation

J'exerce la fonction de relecteur pour les revues nationales ou internationales :
- Traitement du Signal
- IEE Proceedings on Vision Image and Signal Processing
- Pattern Recognition Letters
- IEEE Transactions on Image Processing
- IEEE Transactions on Instrumentation & Measurement

J'ai également été membre des comités de lecture pour les colloques RFIA'00, ISSPA'03, ACIVS'04, SPIE-JOM'05, AEI-IAE'06.

Je suis également membre depuis 3 ans du comité de programmes de la conférence SPIE ISOT 2007 (International Symposium on Optomechatronic Technologies), qui a eu lieu en décembre 2005 à Sapporo au Japon, conférence durant laquelle j'ai animé la session "***Face and Gesture***". C'est à Boston (USA) que cette conférence s'est déroulée en octobre 2006. Elle s'est tenu du 8 au 10 Octobre 2007 à Lausanne, avec ma participation en qualité de co-chair d'une session de ***Computer Vision Systems II*** (http ://www.isot07.org/index.php) et je suis également éditeur associé à ces proceedings [Eds-1].

Activités en liaison avec le secteur industriel

Depuis 2004, je suis responsable des transferts de technologie et du support scientifique de nos méthodes de fusion d'informations, auprès de la société de vidéo surveillance **EBOO Scanner**. Le travail porte sur la fusion des informations issues de différentes caméras en niveaux de gris, couleur, et infra-rouge. Nous travaillons actuellement sur un système de pilotage automatique de caméra de type PTZ, pour la reconnaissance et le suivi de voitures à l'entrée d'un parking. Ce système permet l'ouverture automatique des portes, après reconnaissance de la plaque d'immatriculation. Ce travail fait l'objet d'une communication au congrès ISOT 2007 [CI-23].

J'ai également participé en 2005, à un projet "confidentiel" sur l'analyse d'image avec la société TEFAL [Rp-1]. Il s'est poursuivi en 2006 par l'encadrement d'un stagiaire de fin d'études d'école d'ingénieurs qui a réalisé un prototype. Ce dernier est actuellement testé par le service marketing de TEFAL auprès d'une cible de consommateurs.

1.3 Responsabilités administrative et collective

Les responsabilités présentées correspondent à celles que j'ai exercées ou que j'exerce encore.

depuis 1995 : Membre élu à la Commission de Spécialistes $61^{\text{ème}}$, puis de la commission de spécialistes $61^{\text{ème}}/63^{\text{ème}}$ sections puis de la $27^{\text{ème}}/61^{\text{ème}}$ sections de l'Université de Savoie.

de 1998 à 2005 : Membre nommé à la Commission de Spécialistes $61^{\text{ème}}$ section de l'INPG.

de 2002-2005 : Membre élu au Conseil d'Administration de l'IUT d'Annecy.

de 1996 à 2000, adjoint du responsable du DEA Automatique Industrielle. Le LAMII a été fortement impliqué dans le DEA AI, cohabilité entre l'INSA Lyon, l'Université de Lyon 1,

l'Ecole Centrale de Lyon et l'Université de Savoie. Durant 4 années, l'Université de Savoie en a été l'établissement principal, et Ph. Bolon en a porté la responsabilité. Ensemble, nous nous sommes occupés chaque année, du recrutement, de l'organisation, du planning des cours, et du jury associé à ce DEA.

de 2002 à 2005, membre élu au Conseil d'Administration de l'IUT d'Annecy.

En 1999, j'ai réalisé une expertise à l'EST (école supérieure de technologie) de Meknès (Maroc) concernant l'ouverture d'une option Télécommunications dans le département Génie Electrique.

depuis 1992, Responsable des admissions (recrutement à l'IUT).

depuis 2003, je suis coordinateur national du **Groupe Télécom**, pour l'ensemble des représentants des 28 départements d'IUT en Réseaux et Télécoms de France et Métropole. Cette fonction a consisté entre 2004 et 2005, à rédiger l'ensemble du nouveau programme pédagogique national (PPN) pour les modules liès aux télécommunications. Ce programme a été approuvé par la CPN (commission pédagogique nationale) pour être publié au Journal Officiel du 13 Août 2005. Maintenant, je poursuis l'organisation de réunions du groupe Télécom destinés à parfaire notre enseignement grâce à un échange d'expériences. C'est l'occasion annuelle de faire un bilan de chacun des modules, de discuter les difficultés rencontrées, de la nécessité et du choix de nouveaux équipements. C'est une volonté commune et dynamique de transmission de compétences qui se traduit également par la présentation de nouvelles technologies par les participants, enseignants et professionnels.

1.4 Liste des travaux et publications

1.4.1 Revues d'audience internationale avec comité de Lecture

[R-6] Ionescu B., Coquin D., Lambert P., Buzuloiu V., A Fuzzy Color-Based Approach for Understanding Animated Movie Content in the Indexing Task, Journal on Image and Video Processing, **en révision**.

[R-5] Ionescu B., Lambert P., Coquin D., Buzuloiu V., The Cut Detection Issue in the Animation Movie Domain, Journal of Multimedia, Academy Publisher, ISSN : 1796-2048, Vol. 2, Issue : 4, pp. 10-19, August 2007.

[R-4] Coquin D., Benoit E., Sawada H., Ionescu B., Gesture Recognition Based on the Fusion of Hand Positioning and Arm Gestures, Journal of Robotics and Mechatronics, Vol. 18, No. 6, 2006, pp. 751-759.

[R-3] Ionescu B., Coquin D., Lambert P., Buzuloiu V., Dynamic Hand Gesture Recognition Using the Skeleton of the Hand, EURASIP Journal on Applied Signal Processing, Vol. 2005, No. 13, 2005, pp. 2101-2109.

[R-2] Coquin D., Bolon Ph., Application of Baddeley's distance to dissimilarity measurement between gray scale images, Pattern Recognition Letters, Vol. 22, No. 14, 2001, pp. 1483-1502.

[R-1] Coquin D., Bolon Ph., Discrete distance operator on rectangular grids, Pattern Recognition Letters, Vol. 16, No. 9, 1995, pp. 911-923.

1.4.2 Edition d'ouvrages

[Eds-1] Optomechatronic Computer-Vision Systems II (Proceedings Volume), Eds. Kofman J., Lopez de Meneses Y., Kaneko S., Perez C., Coquin D., Proceedings of SPIE Volume 6718, ISBN 9780819468666, 2007, 148 pages.

1.4.3 Contributions à ouvrage

Version étendue sélectionnée après congrès

[O-3] Ionescu B., Coquin D., Lambert P., Buzuloiu V., Fuzzy Semantic Action and Color Characterization of Animation Movies in Video Indexing Task, Lecture Notes in Computer Science : Adaptive Multimedia Retrieval, Vol. 4398, Springer-Verlag, Berlin Heidelberg, 2007, pp. 119-135.

[O-1] Coquin D., Bolon Ph., Ciuc M., Quantitative assessment of two skeletonization algorithm adapted to rectangular grids, Lecture Notes in Computer Science : Image Analysis and Processing, Vol. 1310, Springer-Verlag, Berlin Heidelberg, 1997, pp. 588-595.

Chapitre d'un livre

[O-2] Coquin D., Bolon Ph., Chap. 7 : Quantitative assessment of image filtering : comparison of objective metrics, Imaging and Vision Systems : Theory, Assessment and Applications, Vol. 9, Nova Science Publishers, Inc., 2001, pp. 129-139.

1.4.4 Conférences d'audience internationale avec actes et comité de lecture

[CI-23] Coquin D., Tailland J., Cintract M., Event detection for car park entries by video surveillance, SPIE - International Symposium on Optomechatronic Technologies, Computer Vision Systems II, CD-ROM , Lausanne, Suisse, October 2007, 8 pages.

[CI-22] Ott L., Lambert P., Ionescu B., Coquin D., Animation Movie Abstraction : Key Frame Adaptive Selection Based on Color Histogram Filtering, Computational Color Imaging Workshop (CCIW'07), CD-ROM , Modena, Italy, September 2007, 6 pages.

[CI-21] Valet L., Coquin D., Jullien S., Teyssier S., A 3D image-segmented evaluation procedure in a cooperative fusion system context, 10^{th} International Conference on Information Fusion , CD-ROM , Quebec, Canada, July 2007, 8 pages.

[CI-20] Ionescu B., Lambert P., Coquin D., Buzuloiu V., Color-Based Content Retrieval of Animation Movies : A Study, IEEE International Workshop on Content-Based Multimedia Indexing, CD-ROM , Bordeaux, France, June 2007, 7 pages.

[CI-19] Ionescu B., Lambert P., Coquin D., Ott L., Buzuloiu V., Animation Movies Trailer Computation, ACM Multimedia, CD-ROM , Santa Barbara, Californie USA, October 2006, 4 pages.

[CI-18] Ionescu B., Coquin D., Lambert P., Buzuloiu V., Semantic Characterization of Animation Movies Based on Fuzzy Action and Color Information, 4th Int. Workshop on Adaptive Multimedia Retrieval, CD-ROM , Geneva, Switzerland, July 2006, 15 pages.

[CI-17] Ionescu B., Lambert P., Coquin D., Buzuloiu V., Fuzzy Color-Based Semantic Characterization of Animation Movies, Third European Conference on Color in Graphics Imaging and Vision, CD-ROM , Leeds, United Kingdom, June 2006, 5 pages.

[CI-16] Ionescu B., Lambert P., Coquin D., Buzuloiu V., Improved cut detection for the segmentation of animation movies, IEEE Int. Conf. on Acoustics Speech and Signal Processing, CD-ROM , Toulouse, France, May 2006, 4 pages.

[CI-15] Coquin D., Benoit E., Sawada H., Ionescu B., Fusion of Hand and Arm Gestures, SPIE - International Symposium on Optomechatronic Technologies, ISOT - Machine Vision, CD-ROM , Sapporo, Japon, December 2005, 11 pages.

[CI-14] Ionescu B., Lambert P., Coquin D., Darlea L., Color-Based Semantic Characterization of Cartoons, IEEE Int. Symposium on Signals, Circuits and Systems, Iasi, Roumanie, July 2005, pp. 223-226.

[CI-13] Benoit E., Coquin D., Sawada H., Distributed data fusion applied to human gesture measurement, 6th Int. Workshop on research and Education in Mechatronics (REM 2005), Annecy, France, June 2005, pp. 92-96.

[CI-12] Coquin D., Bohner M., Segmentation of Range and Reflectance Images with an Expert System, 6th Int. Conf. on Information Fusion, Cairns, Queensland, Australia, July 2003, pp. 943-950.

[CI-11] Coquin D., Bolon Ph., Ionescu B., Dissimilarity measures in color spaces, IEEE, IAPR, 16th Int. Conf. on Pattern Recognition (ICPR 2002), Quebec City, Canada, August 2002, pp. 612-615.

[CI-10] Coquin D., Bolon Ph., A new method to compute the distortion vector field from two images, IEEE, IAPR, 16th International Conference on Pattern Recognition (ICPR 2002), Quebec City, Canada, August 2002, pp. 279-282.

[CI-9] Coquin D., Bolon Ph., Onéa A., Objective metric for colour image comparison, 10th European Signal Processing Conf. (EUSIPCO'2000), Tampere, Finland, September 2000, pp. 119-122.

[CI-8] Coquin D., Bolon Ph., Onéa A., 3D Nonstationary local distance operator, IEEE, IAPR, 15th Int. Conf. on Pattern Recognition (ICPR 2000), Barcelona, Spain, September 2000, pp. 963-966.

[CI-7] Coquin D., Bolon Ph., Quantitative assessment of image filtering : comparison of objective metrics, Workshop on Advanced Concepts for Intelligent Vision Systems (ACIVS'99), Baden-Baden, Germany, August 1999, pp. 92-103.

[CI-6] Coquin D., Bolon Ph., Chehadeh Y., A skeletonization algorithm using chamfer distance transformation adapted to rectangular grids, IEEE, IAPR, 13th Int. Conf. on Pattern Recognition (ICPR 1996), Vol. 2, Vienne, Austria, August 1996, pp. 131-135.

[CI-5] Chehadeh Y., Coquin D., Bolon Ph., A generalization to cubic and non cubic local distance operators on parallelepipedic grids, Actes du 5ième colloque DGCI, Clermont-Ferrand, France, September 1995, pp. 27-36.

[CI-4] Coquin D., Chehadeh Y., Bolon Ph., 3D local distance operator on parallelepipedic grids, 4th Discrete Geometry for computer imagery, Grenoble, France, September 1994, pp. 147-156.

[CI-3] Coquin D., K. Chehdi, Pattern Recognition by Image Analysis. IEEE, IAPR, 11th International Conference on Pattern Recognition (ICPR 1992), The Hague, The Neederlands, August 1992.

[CI-2] Coquin D., K. Chehdi, Automatic identification and counting of zooplankton by image analysis, First International Conference on Electronics and Automatic Control, ICEA 92, Tizi Ouzou, Algeria, 1992.

[CI-1] Coquin D., K. Chehdi, Binarisation of various images by detecting local thresholds with validation test, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, B.C. Canada, vol. 2, 1991, pp. 611-614.

1.4.5 Conférences d'audience nationale et francophone avec actes

[CN-9] Ionescu B., Lambert P., Coquin D., Buzuloiu V., Influence de la Réduction des Couleurs sur la Détection des Changements de Plan dans les Films d'Animation, 20e colloque GRETSI, CD-ROM , Louvain-la-Neuve, Belgique,septembre 2005, 4 pages.

[CN-8] Ionescu B., Coquin D., Lambert P., Buzuloiu V., Analyse et caractérisation de séquences de films d'animation, Congrès Jeunes Chercheurs en Vision par Ordinateur (ORASIS 05), CD-ROM , Fournols, mai 2005, 10 pages.

[CN-7] Ionescu B., Coquin D., Lambert P., Reconnaissance de gestes dynamiques de la main, 19ème colloque sur le traitement du signal et des images (GRETSI'03), Vol. III, Paris, France, septembre 2003, pp. 22-25.

[CN-6] Bouloudani N., Lambert P., Coquin D., Segmentation Automatique des Images couleur à base d'Indicateurs de Performance, 8èmes journées CORESA (COmpression et REprésentation des Signaux Audiovisuels), Lyon, France, mai 2003, pp. 201-204.

[CN-5] Bolon Ph., Coquin D., Signature de la déformation entre deux images à niveaux de gris, 18ème colloque sur le traitement du signal et des images (GRETSI'01), Vol. II, Toulouse, septembre 2001, pp. 349-352.

[CN-4] Coquin D., Bolon Ph., Chehadeh Y., Evaluation quantitative d'images filtrées, 16ème Colloque sur le Traitement du Signal et des Images (GRETSI'97), Vol. 2, Grenoble, France, septembre 1997, pp. 1351-1354.

[CN-3] Coquin D., Bolon Ph., Chehadeh Y., Opérateurs de distance 3D. Application à la comparaison d'images, 15ème Colloque sur le Traitement du Signal et des Images (GRETSI'95), Vol. 2, Juan-Les-Pins, France, septembre 1995, pp. 761-764.

[CN-2] Coquin D., Bolon Ph., Comparaison d'opérateurs locaux de distance, Colloque en géométrie discrète : Fondement et Applications, Strasbourg, France, septembre 1993, pp. 182-191.

[CN-1] Coquin D., K. Chehdi, Binarisation de plusieurs classes d'images par seuillage local optimal maximisant un critère d'homogénéité, 13ème Colloque sur le Traitement du Signal et des Images (GRETSI'91), Vol. 2, Juan-Les-Pins, France, septembre 1991, pp. 1069-1072.

1.4.6 Conférences sans Acte et Journée d'études

[J-8] Lambert P., Ionescu B., Coquin D., La couleur dans les séquences d'images, Actes de l'Ecole d'Hiver sur l'Imagerie Numérique Couleur (EHINC'2007), Poitiers, Janvier 2007.

[J-7] Coquin D., Indexation et recherche Multimédia, 15 Septembre 2006 [J-6] Coquin D., P. Lambert, B. Ionescu , Indexation et recherche d'information Multimédia, présentation faite dans le cadre du GT5, 24 septembre 2004.

[J-5] Coquin D., Lambert P., Ionescu B., Reconnaissance de gestes de la main, GT5, présentation faite dans le cadre du GT5 Perception Modélisation et Interprétation du Geste Humain, 27-28 mars 2003.

[J-4] Coquin D., Numérisation, GT.5, 24-25 Janvier 2001

[J-3] Coquin D., Bolon Ph., Etude comparative de mesures de dissimilarité, présentation faite dans le cadre du GT3 (OT.3.5), 27 Janvier 2000.

[J-2] Coquin D., Bolon Ph., mesure de qualité et contrôle en segmentation : présentation faite dans le cadre du GT3 (OT.3.5), 18 Novembre 1999

[J-1] Coquin D., Deuxième journée de l'image qualité : GT8, Dijon, Juin 1999.

1.4.7 Rapports de synthèse et rapports internes

[Rp-1] Bolon Ph., Coquin D., Lambert P., Trouvé E., Vacher P., Analyse d'image, Rapport Interne LISTIC n° 05/08, Université de Savoie, 2005, 27 pages.

[Rp-2] Coquin D., Mesure de dissimilarité entre images en niveaux de gris. Rapport Interne LAMII n° 00/02, Université de Savoie, 2000, 17 pages.

[Rp-3] Coquin D., Bolon Ph., A modification of Wilson-Baddeley-Owen dissimilarity measure for gray-scale image comparison. Rapport Interne LAMII n° 00/05, Université de Savoie, 2000, 20 pages.

[Rp-4] Coquin D., Bolon Ph., A new metric for grey-scale image comparison. Rapport Interne LAMII n° 98/04, Université de Savoie, 1998, 16 pages.

[Rp-5] Coquin D., Opérateur locaux de distance en maillage rectangulaire. Rapport Interne LAMII n° 94/03, Université de Savoie, 1994, 23 pages.

[RD-1] Coquin D. Segmentation et Analyse d'Images pour la Classification automatique : application au Zooplancton. Thèse de Docteur de l'Université de Rennes I, Mention : Traitement du signal et Télécommunication. N° 685, 1991.

[RD-2] Coquin D. Extraction automatique des contours cellulaires de l'endothélium cornéen. DEA de Traitement du Signal et Télécommunications de l'Université de Rennes 1, stage de recherche effectué à l'ENSTBr, 1987.

[RD-3] Coquin D. Participation à la campagne de mesure St Privat d'Allier : Signaux induits par la foudre sur les cables de télécommunications. Rapport RP/LAA/ELR/613, CNET Lannion A, 1986.

1.5 Résumé des activités d'Enseignement

Mon activité d'enseignement présente 4 volets, correspondant à des périodes ou des établissements différents. Le premier volet correspond à l'activité d'enseignement que j'ai effectuée à l'ENSSAT, à Lannion, Université de Rennes 1, durant mes 4 années de thèse (1987-1991) et durant l'année d'ATER qui a suivi. Les deux volets suivants portent respectivement, sur mon activité d'enseignement, au sein du département GEii d'Annecy, puis au sein du département R&T (anciennement GTR). Le dernier volet porte sur des enseignements plus ponctuels directement liés à mes travaux de recherche. **volet 1 :** j'ai effectué

ma thèse sans financement du ministère. C'est avec les emplois d'Assistant Associé (192h équivalent TD durant 1 an), d'ALER (96h équivalent TD durant 2 ans) et $\frac{1}{2}$ ATER (96h équivalent TD durant 1 ans), que j'ai réussi à financer mes travaux de recherche. Lorsque je suis arrivé à l'ENSSAT, en 1987, l'école d'ingénieurs n'avait qu'une année d'ancienneté. J'ai donc participé au montage des enseignements de deuxième puis troisième année (TD et

TP en signaux et systèmes, transmissions numériques et électronique). Durant ma deuxième année d'ATER (1991-1992), ma thèse étant soutenue, j'ai effectué un plein temps d'ATER, durant lequel j'ai eu la responsabilité des deux cours suivants : Amplification et commutation (20h), et Instrumentation électronique (15h). Du fait de leur ancienneté, je ne détaillerai pas plus ces enseignements.

Dans les tableaux suivants, les enseignements dont j'ai eu la responsabilité sont indiqués en **gras**.

volet 2 : en tant que Maître de Conférences au département GEii de l'IUT d'Annecy. Cette activité, correspondant à mon service statutaire, débuta en octobre 1992 avec comme mission de participer aux enseignements du département et prendre en charge un groupe de projet de deuxième année. Les principaux éléments de cette activité sont résumés dans le tableau ci-dessous.

Période	**Public**	**Matière**	**Contenu**	**Nbre d'heures/an**
1995-2001	Dépt GEii 2ème année	**Maths CM/TD**	Base de probabilités, statistiques descriptives, estimation et tests de loi, fiabilité	25h /an
1996-1998	Dépt GEii 1ère année	**Informatique CM/TD/TP**	Base de l'algorithmie et de l'informatique industrielle (68HC11, Langage C)	70h /an
1995-1998	Dépt GEii 3ème année Sport-Etudes	**Maths CM/TD**	Analyse vectorielle, algèbre matricielle, série numérique	50h /an
1992-1997	Dépt GEii 2ème année	**Projet**	Responsable d'un groupe TP en projet	30h/an
1992-1998	Dépt GEii 2ème année	Electronique TD/TP	Modulations AM et FM, circuits bouclés, analyse spectrale	50h/an

volet 3 : J'ai rejoint, en septembre 1998, le département Réseaux et Télécoms qui souffrait alors d'un fort déficit en enseignants (ouverture d'un 3ème groupe en 1ère année) ceci, en plein accord avec le département GEii. J'ai apporté mes compétences en Télécommunications, en prenant la responsabilité des Cours-TD-TP de 1ère année, et en participant activement aux Télécommunications en 2ème année et à l'Informatique de 1ère année.

Que ce soit au sein du département GEii ou au sein du département R&T, la responsabilité d'un module recouvre la conception du cours et des TD/TP ainsi que l'animation de l'équipe pédagogique, chaque groupe de TD ou de TP ayant un enseignant différent. Cette manière de fonctionner assure à la fois continuité (un seul responsable pour les trois formes d'enseignement) et diversité (plusieurs intervenants).

Je privilégie, dans le cadre de mes travaux pratiques, une approche concrète des compétences à acquérir. C'est par conséquent sur du matériel dédié et non par seule simulation, qu'ils apprennent à utiliser correctement les appareils de mesures, à analyser les résultats, à comprendre le fonctionnement des systèmes, ...

Période	**Public**	**Matière**	**Contenu**	**Nbre d'heures/an**
1998-2004	Dépt R&T 1ère année	**Télécom CM/TD/TP**	Télécommunication analogique et numérique	120h /an
depuis 2005	Dépt R&T 1ère année	**Télécom CM/TD/TP**	Module T1 : Fondamentaux des télécoms et transmission	60h /an
depuis 2005	Dépt R&T 1ère année	**Télécom CM/TD/TP**	Module T3 : Téléphonie, PABX, RNIS	30h /an
1998-2007	Dépt R&T 2ème année	Télécom	TD/TP Traitement du signal, Transmissions numériques, Haut débit SDH, GSM, ADSL	60h /an
depuis 1998	Dépt R&T 2ème année	**Maths CM/TD**	Base de probabilités, statistiques descriptives, estimation et tests de loi	30h /an
depuis 1998	Dépt R&T 1ère année	Informatique TD/TP	Bases de l'algorithmie et de l'informatique (langage C, Java), programmation Web : HTML	60h/an

volet 4 : en tant que vacataire en Maîtrise EEA puis en filière Ingénieurs et autres interventions ponctuelles Cette activité débuta deux ans après mon arrivée à l'IUT d'Annecy. Ces vacations ont d'abord été assurées en Maîtrise EEA puis à l'Ecole Supérieure d'Ingénieurs d'Annecy (ESIA, maintenant Polytech'Savoie). Les principaux éléments de cette activité sont résumés dans le tableau ci-dessous.

Période	**Public**	**Matière**	**Nbre d'heures/an**
1994 - 1997	Maîtrise EEA	TD de Traitement du signal	20h/an
1996 - 1998	DEA AI	Cours de Traitement d'images Vision industrielle 2D et 3D	15h /an
depuis 2000	ESIA 2ème année PAI	TD de Traitement du signal	12h/an
2002 - 2004	5ème année de la faculté de Bucarest	Cours en Traitement d'images : Géométrie discrète (Ecole de printemps organisée par l'Université Politehnica de Bucarest)	10h /an
2003 - 2005	INSA Lyon 5ème année	Cours en Traitement d'images Vision industrielle	8h /an
2004	Ecole doctorale SPI	Cours de Reconnaissance des formes	4h/an
2006	Ecole doctorale SPI	Cours d'initiation au Traitement et analyse des images	4h/an

C'est à travers ces dernières responsabilités d'enseignement en traitement des images, à Bucarest, à L'INSA de Lyon, en DEA ou à l'école doctorale que j'ai réellement pu établir un lien avec mes activités de recherche. Ces interventions ont nécessité la rédaction de supports pédagogiques. Il faut également noter que le cours de DEA a donné lieu à une expérience pédagogique assez originale à l'époque, de visio-conférence entre Annecy et Lyon. L'objectif de mon enseignement est de préparer efficacement nos étudiants à leurs activités professionnelles.

1.6 Projet en cours et Perspectives

Les travaux développés durant ces dernières années et au travers des différents DEA-Master et de la première thèse encadrée ont essentiellement porté sur l'analyse des images, plus particulièrement sur l'évaluation des traitements et la comparaison des méthodes par analyse des résultats.

Suite à ces travaux, plusieurs pistes mériteraient d'être approfondies. Tout d'abord, une mesure de dissimilarité qui combinerait des informations quantitatives et qualitatives, en associant l'avis de l'utilisateur, permettrait d'apporter plus de subjectivité. Pour répondre à cette approche, un axe de recherche intéressant serait l'étude de descripteurs qui permettent de caractériser l'impression qui se dégage d'une image, ce que les japonais appellent le "*Kansei*". Ce type d'étude, peu encore développé en France, connaît un fort intérêt au Japon depuis plusieurs années. Il s'agit de donner la description d'une image en des termes proches de la perception humaine : *cette image dégage une impression de bonheur*. Cette approche avait été initiée lors de la thèse de Nadia Bouloudani. Les applications potentielles de ces travaux sont multiples. On peut citer en particulier la recherche dans une base d'images. La plupart des travaux existants s'appuient sur des mesures de similitude entre des attributs de forme, de texture ou de couleur. Il serait extrêmement riche, pour répondre de manière plus pertinente aux requêtes, d'ajouter des descripteurs de ce type. Plus récemment avec la deuxième thèse soutenue, nous nous sommes penchés sur la caractérisation symbolique de séquences d'images et avons initié des travaux dans ce sens. Là aussi, des descripteurs issus du "*Kansei*" apporteraient une richesse supplémentaire.

Une autre piste de recherche qui me semble intéressante est le **rebouclage** dans les systèmes de traitement d'images. En effet, toute méthode de traitement d'images, ou plus généralement, tout système nécessite des paramètres. Or, il est bien souvent difficile d'ajuster ces paramètres et ceux-ci sont, la plupart du temps, liés à l'application. En utilisant une mesure de dissimilarité, et le jugement d'un utilisateur, nous pourrions comparer les résultats successifs, guider l'utilisateur et ajuster les paramètres pour tendre vers un "*bon*" résultat. Ma participation plus active à l'action 3 (SCATI) du thème B du GDR ISIS et les résultats encourageants que nous avons obtenus dans le projet BQR portant sur "*un système coopératif de fusion d'informations pour l'interprétation d'images 3D*", vont dans ce sens.

Une dernière piste qui me tient à cœur est l'évaluation de performance des systèmes de fusion d'informations, plus particulièrement, l'évaluation de ceux associés aux traitements des images. Il serait très intéressant de définir une méthodologie générique qui permette d'évaluer la performance des systèmes de fusion d'informations. Notamment, cette évaluation devra permettre de comparer des différentes méthodes de fusion, de mieux comprendre l'apport de chaque paramètre de la méthode et leur interaction sur la sortie du système et de mieux comprendre l'apport de chaque système de fusion, dans des systèmes coopératifs.

Les méthodes subjectives ont démontré leurs intérêts mais nécessitent des conditions d'expérimentation parfois lourdes (tests sur une base de données importante, bonnes conditions de visualisation, durée importante pour effectuer l'évaluation, choix du nombre d'utilisateurs expérimentés ou non, ...). Les méthodes objectives nécessitent quant à elles une vérité terrain, pour qu'elles soient pertinentes. Mais cette vérité terrain est parfois difficile à obtenir.

Nous proposerons une mesure objective qui évaluera la quantité relative d'information qui est transférée de l'image d'entrée vers l'image de sortie du système de fusion. Nous pensons, par exemple, à une mesure de dissimilarité entre images, ou une mesure sur la préservation des contours ou des régions, ou une mesure basée sur des statistiques locales calculées sur une portion de l'image. Afin de valider de manière significative la méthode de fusion, un mécanisme de tests appropriés de comparaison subjective-objective sera défini, par exemple, à partir d'un classement subjectif, d'un vote, etc. Nous appliquerons la méthode développée, par exemple, à l'analyse d'images tomographiques ou d'images multi-spectrales, permettant ainsi l'amélioration du résultat de la segmentation issue d'un système coopératif de fusion d'informations.

C'est vers ces trois dernières pistes que je souhaite m'engager dans mes futurs travaux de recherche.

Deuxième partie

Description des travaux de recherche

Chapitre 2

Introduction

2.1 Contexte des travaux

Les travaux présentés dans ce mémoire ont été effectués entre 1992 et 2007 à l'université de Savoie, au LAMII (Laboratoire d'Automatique et de MicroInformatique Industrielle) devenu le LISTIC (Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance) après un regroupement de laboratoires. L'équipe dans laquelle j'ai travaillé, actuellement dénommée "Traitement de l'Information", a pour objectif à moyen terme, de fournir des éléments permettant de réaliser l'adaptation ou la reconfiguration en temps réel des **Méthodes de Fusion**, sur la base de critères de performance ou de comparaison. Le traitement des images est un des champs d'investigation privilégié.

Avant d'entrer dans le sujet proprement dit, il est important de comprendre la nature des objets que nous allons manipuler. Intéressons-nous à la **notion d'image**. Qu'est-ce qu'une image ? Une image est une représentation imprimée d'un sujet quelconque. Cela signifie qu'une image nécessite un support sur lequel elle sera visible. Ainsi une photographie papier, une peinture sont des exemples d'images au même titre qu'une image numérique affichée sur un écran d'ordinateur. Informatiquement, une image est une représentation numérique en mémoire d'un sujet imprimé sur une rétine artificielle (matricielle, comme le capteur d'un appareil photographique numérique ou la scène virtuelle d'une image de synthèse, comme le capteur optique du télécopieur, du photocopieur ou du scanner). Nous allons donc travailler sur des ensembles de nombres numériques codés sur un ordinateur.

A partir d'une image, nous pouvons extraire deux types d'information. Le premier type d'information est appelé **niveau syntaxique** (ou graphique). Il nous donne des renseignements sur la scène que représente l'image. Le second se nomme **niveau sémantique**. C'est la phase d'interprétation de l'image qui varie d'une personne à une autre en fonction des connaissances de chacun et du contexte d'observation. Il faut prendre en compte dans la sémantique, le contexte social et sociétal de l'observateur de l'image. Une image n'a pas la même signification selon la société de l'observateur. Les différentes cultures et les modes de vie peuvent influencer l'interprétation d'une scène. L'aspect syntaxique et l'aspect sémantique sont importants pour la recherche d'images et nous verrons plus tard qu'il est très difficile de décoder le niveau syntaxique pour arriver au niveau sémantique (c'est pourtant le plus important problème à résoudre). Les images existent depuis la nuit des temps et leur nombre croît exponentiellement. Mais sans la vision, sans ce formidable outil optique qu'est l'œil, il n'y aurait pas d'images. La vision est le plus important de nos cinq sens. C'est celui qui nous permet de percevoir notre environnement et d'interagir avec lui. La vision coordonne notre attention, nos mouvements, nos réactions, elle oriente nos décisions. Elle permet de différencier les couleurs, les formes, les textures, les visages, les objets, les scènes. Elle nous donne beaucoup plus d'informations que n'importe quel autre moyen de description. Si plusieurs personnes observent une image, tous perçoivent la même image mais chacun l'interprète à sa façon, selon ses connaissances, son passé, son contexte social et sociétal et sa propre vision du monde.

Il existe un autre problème : celui de la non adéquation entre la représentation informatique et la signification dans le monde réel. En effet, quand nous observons par exemple une image, nous

"voyons" des objets et une scène mais pas une série de pixels. Nous appliquons effectivement une grille sémantique sur l'image. Mais au niveau du codage informatique nous n'avons qu'une série de vecteurs qui décrivent l'image en traduisant simplement les couleurs point par point. Ce décalage est connu aujourd'hui, dans le milieu de la recherche informatique, sous l'appellation de "**fossé sémantique**".

En rejoignant en 1992 l'axe *Vision Industrielle* mis en place par Philippe Bolon, j'ai choisi d'orienter mon activité de recherche autour de la **comparaison des images**. Pour réduire ce fossé sémantique, nous avons travaillé directement sur l'image - images fixes et séquences d'images - afin d'en extraire une information nécessaire à sa caractérisation. Nous avons également travaillé sur la nature discrète des images dans le but de développer des mesures de similarité dans les espaces discrets. L'objectif de ces travaux est d'évaluer la performance au niveau local et global de traitements évolués.

Lorsque les conditions d'acquisition d'images sont bonnes, ou lorsque l'on a pu utiliser un bon opérateur de prétraitement, il est possible de faire, par exemple, une segmentation de l'image, de façon relativement simple. L'analyse de cette image nécessite une caractérisation des entités ainsi mises en évidence. La difficulté de cette analyse provient de la nature discrète de l'image numérique qui entraîne une déformation des structures observées par rapport à leur apparence réelle dans l'espace continu. De plus, lorsque les images sont acquises avec des systèmes de vision industrielle standards de l'époque (1992), la forme du pixel est rectangulaire et non carrée. Les outils d'analyse doivent tenir compte de cette source d'anisotropie supplémentaire. Nos efforts ont porté, au début, sur la mise au point d'opérateurs de distance et sur l'étude quantitative des effets de discrétisation de structures continues.

Mes premiers travaux ont donc consisté à définir des opérateurs de distance discrète en maillage rectangulaire. En parallèle j'ai implémenté une méthode de segmentation d'images (*méthode de Nakagawa : méthode de seuillage dynamique intégrant des informations de localisation*) en vue de la comparaison de méthodes de segmentation que l'on retrouve dans le chapitre X du livre "**Analyse d'images : Filtrage et Segmentation**" ; ouvrage collectif coordonné par J.P. Cocquerez et S. Philipp-Foliguet [Bolon 95]. Tout naturellement, nous avons voulu utiliser ces opérateurs de distance pour **comparer** les résultats de méthodes de traitement d'images et pour **évaluer leurs performances**. Puis, nous nous sommes tournés vers des applications plus complexes comme, par exemple, l'extraction de caractéristiques issues de la fusion de traitements d'images de réflectance et de profondeur ou encore, la reconnaissance de gestes dynamiques de la main à partir de la fusion des informations issues d'une caméra et d'un gant numérique. Dernièrement, nous avons travaillé sur l'indexation de séquences d'images. Toutes ces applications nécessitent à un moment ou à un autre **la comparaison des images**.

2.2 Travaux développés

Pour améliorer un traitement complexe ou une méthode de fusion, nous avons besoin de **comparer** le résultat par rapport à une attente. Dans le cadre applicatif du traitement des images, nous nous sommes fixés les objectifs suivants :

- définir une mesure de dissimilarité qui permette de comparer les images deux à deux (de façon globale puis locale) ;
- tenir compte du caractère discret des images numériques, et par conséquent, développer des opérateurs de distance en maillage rectangulaire pour le 2D et parallélépipédique pour le 3D pour s'adapter aux capteurs de vision utilisés ;
- fournir des informations sur la qualité du résultat ;
- permettre à l'utilisateur de comprendre les résultats qui lui sont proposés afin d'agir sur les paramètres de la méthode et ainsi d'affiner le résultat ;
- comparer les séquences d'images et pour ce faire, extraire des informations pertinentes qui permettent une indexation de ces séquences d'images par une description sémantique,
- proposer une mesure de distance entre les plans, les scènes d'une séquence d'images.

C'est au travers des deux thèses que j'ai co-encadrées, ainsi qu'au travers des 8 stages de DEA-Master et de 2 stages d'étudiants venus dans le cadre de collaborations avec l'étranger, que nous avons tenté de répondre à ces objectifs. Ma participation à de nombreux projets, m'a également permis d'appliquer et de valider les méthodes que nous avons développées.

Tout d'abord, nous avons travaillé sur la nature discrète des images par le biais de la thèse de Yousra CHEHADEH portant sur **les opérateurs locaux de distance en maillages rectangulaire et parallélépipédique avec comme application, l'analyse des images** [Chehadeh 97]. Nous nous sommes intéressés aux distances discrètes, en particulier, aux distances de chanfrein définies par des opérateurs locaux de distance. La distance entre deux pixels (voxels) peut être calculée par propagation de l'information locale donnée par l'opérateur. Après avoir comparé les méthodes d'optimisation les unes par rapport aux autres, nous avons proposé une méthode d'optimisation adaptée au maillage rectangulaire en 2D et nous l'avons étendue en 3D au cas des maillages parallélépipédiques. Cette méthode s'inspire des méthodes existantes et essaie d'en tirer les avantages. Ces opérateurs servent à calculer l'image de distance, ou la carte de distance, d'une image binaire. Les applications de ces opérateurs et des images de distance sont nombreuses aussi bien en 2D qu'en 3D.

Nous avons proposé une application des opérateurs locaux de distances 3D à la comparaison d'images 2D en vue de l'évaluation des algorithmes et des opérateurs de traitement d'images. Deux cas ont été considérés : le filtrage et la compression des images. Pour cela, nous avons proposé une **mesure de dissimilarité** entre deux images à niveaux de gris et étendu ce principe aux images couleurs. Ce critère global est fondé sur la distance moyenne de Baddeley, qui calcule la distance moyenne entre deux ensembles binaires.

Nous avons également proposé une méthode d'extraction du squelette d'une forme à partir des images de distance. Nous avons appliqué cette méthode dans un système de reconnaissance des gestes dynamiques de la main à partir de la fusion des informations issues d'un gant numérique et d'une caméra. Pour analyser un geste de la main, nous avons défini la signature dynamique comme étant la superposition des squelettes extraits de la séquence. La mesure de dissimilarité, développée précédemment, a été utilisée pour comparer les signatures dynamiques entre elles.

La thèse de Bogdan IONESCU quant à elle, porte sur **la caractérisation symbolique de séquences d'images avec comme application, les films d'animation** [Ionescu 07]. Sa spécificité tient à la recherche de caractéristiques de nature sémantique, domaine délicat et encore peu abordé dans la littérature. L'objectif est de fournir, à terme, des outils logiciels efficaces permettant la recherche et la navigation dans des bases de séquences d'images. Les films d'animation, dans le contexte du "Festival International du Film d'Animation d'Annecy", servent de support applicatif aux méthodes mises en œuvre.

Le système d'analyse sémantique envisagé dans cette thèse comporte deux axes fondamentaux d'analyse : une première analyse de bas niveau alimentant une seconde analyse de plus haut niveau, de nature symbolique.

Dans l'analyse de bas niveau, trois types d'informations sont recherchés. D'abord, le film est découpé en ses unités de base (les plans vidéo) obtenues par détection des transitions vidéo. Les particularités des films d'animation (grande variabilité, absence de règles de montage, effets couleurs, etc.) ont nécessité le développement d'approches spécifiques pour la détection des "cuts" (transitions brutales), des "fades" (transitions graduelles), et des SCC (ou "short-color-change", effet couleur spécifique). Ensuite, une estimation du déplacement par une analyse par blocs est effectuée pour caractériser le mouvement global dans le film. Enfin, l'analyse de la distribution des couleurs permet de caractériser chaque séquence par une signature couleur. Avec comme point de départ ces trois types d'informations, le contenu du film est caractérisé par un certain nombre de mesures statistiques concernant : la structure des plans (fréquence des changements, présence d'effets particuliers, etc.), le mouvement de la caméra (mouvement de translation, zoom, mouvement d'objets, etc.), la distribution des couleurs (couleurs prédominantes, couleurs complémentaires, etc.). En complément, une annotation visuelle est mise à la disposition des experts pour illustrer d'un seul coup d'œil la structure du film. A partir de ces informations de bas niveau, un certain nombre de résumés "intelligents" sont proposés comme, par exemple, la bande-annonce du film. Ces résumés sont construits en utilisant une sélection d'images clés obtenues après **comparaison** et analyse des histogrammes couleurs.

Dans l'analyse sémantique, les informations de bas niveau acquises dans la première étape sont transformées en utilisant des représentations symboliques et, selon les cas, des combinaisons de ces informations symboliques à l'aide de règles floues. Ces représentations et ces règles sont construites en intégrant une connaissance fournie par les experts du domaine. C'est cette connaissance experte qui permet de hausser le niveau sémantique des descripteurs. Les films d'animation sont caractérisés du point de vue de la structure (rythme, action, mystère, explosivité) et de la distribution des couleurs (importance des couleurs prédominantes, diversité/variété, contrastes, etc.). Pour montrer la puissance discriminatoire des descriptions sémantiques proposées, nous les avons utilisées pour classifier une partie d'une base de films d'animation (52 films). Ces travaux sont les premiers que nous avons entrepris dans le domaine, et sont donc un premier pas vers la caractérisation sémantique des films d'animation.

2.3 Organisation du mémoire

Le mémoire est divisé en quatre parties principales.

La première partie (chapitre 3) cherche à dégager les caractéristiques générales de la comparaison des images. Le domaine est vaste et nous avons choisi d'organiser cette présentation en deux temps :

- les domaines liés à la comparaison d'images
- les méthodes de comparaison d'images

Ce chapitre met en évidence ***deux pistes*** pour comparer les images. ***Une première*** porte sur la **comparaison directe**. Elle nécessite une mesure de dissimilarité. Celle-ci sera basée sur l'utilisation d'opérateurs locaux de distance exposés dans la seconde partie (chapitre 4) de ce mémoire. Nous présenterons notre réflexion autour des opérateurs locaux de distance et exposerons notre démarche en vue de l'optimisation de ceux-ci.

La troisième partie (chapitre 5) décrit les mesures de dissimilarité que nous avons définies à partir de nos opérateurs locaux de distances, pour la comparaison d'images binaires, à niveau de gris, puis couleur.

La seconde piste développée pour comparer les images repose sur la **comparaison indirecte** à partir d'attributs extraits des images. Nous avons utilisé cette voie pour comparer les séquences d'images par l'intermédiaire de résumés ou par une représentation visuelle des attributs grâce à la définition de "gamuts sémantiques". Nous apporterons un début de réflexion sur le sujet et présenterons des méthodes de construction automatique de résumés de séquences d'images adaptées au cinéma d'animation ainsi qu'une méthode de comparaison de "gamuts sémantiques".

Enfin, nous conclurons ce mémoire par nos perspectives (Chapitre 7) vis-à-vis de nos préoccupations actuelles.

Chapitre 3

La comparaison d'images

Résumé : *Ce chapitre expose la problématique de la comparaison d'images pour l'évaluation de certains traitements, pour la reconnaissance des formes et pour l'indexation de séquences d'images. Nous positionnons nos travaux par rapport à ce vaste domaine.*

3.1 Introduction

L'essor du traitement de l'image et son utilisation dans de nombreux secteurs d'activité aussi différents que la biométrie, la médecine ou la vidéo surveillance ont donné naissance à de nombreuses bases de données d'images fixes ou de séquences d'images. Le support informatique des images permet notamment de les comparer. Cela peut être utile à plusieurs titres :

- pour **visualiser les différences ou les similitudes** entre les images, et pour les quantifier automatiquement, tant spatialement que radiométriquement. Ainsi, certains traitements pourront être comparés, ce qui permettra d'analyser leurs performances, de régler les paramètres de l'algorithme, d'ajuster les règles utilisées pour la fusion des informations, de régler les seuils de la partition floue, etc. Nous pensons par exemple à l'analyse de certaines méthodes de filtrage, de certaines méthodes de compression ou encore de certaines méthodes de segmentations pour lesquelles, il est difficile de juger de la performance par simple visualisation du résultat, ...
- pour **visualiser l'évolution dans le temps d'un phénomène** en comparant des images prises à des instants différents (images multi-dates par exemple) et on étudie l'évolution des structures d'une date à l'autre (télé-détection : suivi de l'évolution de la déforestation dans certaines régions d'Afrique ou d'Amazonie, suivi du trait de côte, suivi de l'évolution des glaciers, aide au diagnostic dans le domaine médical, ...). En comparant des images dans une séquence d'images, cela permet, par exemple, d'accéder aux informations de mouvement des objets se déplaçant dans une scène ou d'extraire une information pertinente de la séquence. Là encore, on mettra en œuvre des méthodes de recherche d'objets par comparaison soit en recherchant la forme dans sa globalité, soit en recherchant certaines caractéristiques de l'objet.
- pour **exploiter la redondance des informations recherchées** pour permettre un traitement plus robuste (image acquise dans plusieurs bandes spectrales, utilisation de plusieurs sources d'acquisition de la même scène (images multi-spectrales, images couleur, images de profondeur-réflectance, ...).
- pour **rechercher des images** dans une base de données. On pense alors aux deux techniques sous-jacentes : la recherche d'images à partir du texte associé à l'image (première étape d'indexation) et la recherche d'images par leur contenu (c'est-à-dire à partir des données extraites de l'image elle-même). En effet, ces bases de données sont souvent trop volumineuses pour être annotées de manière manuelle. Elles nécessitent alors le développement de méthodes d'exploitation automatiques comme la recherche d'images à partir d'une requête. Dans ce cas particulier de la recherche d'images, les méta-données (texte descriptif, ...) sont parfois absentes ou inadaptées. Il est alors nécessaire de se baser sur le contenu des images pour effectuer la recherche. Plusieurs méthodes existent, les unes basées sur la comparaison de descripteurs des images (une

certaine forme de signature de l'image), et les autres à partir de la comparaison de points ou zones d'intérêt issus de l'image.
- pour **rechercher un certain genre de films dans une base de données**. Là encore, il est, à l'heure actuelle, très difficile de retrouver des films selon certains critères comme, par exemple, un film contenant une scène de poursuite. De nombreux travaux sont réalisés dans ce domaine, le but étant l'indexation automatique des films par leur contenu.

Notons toutefois que cette liste n'est pas exhaustive car elle n'aborde que certains domaines de traitement d'images sur lesquels nous avons travaillés, et que je vais vous présenter dans ce mémoire.

Selon le type d'image, les **techniques de comparaison** sont bien différentes. Les images binaires qui peuvent provenir de capteurs binaires, d'une extraction de contours, d'une binarisation, etc. , semblent les plus faciles à comparer puisque le nombre de niveaux n'est que de deux. Elles présentent néanmoins des difficultés spécifiques. La pauvreté de leurs attributs, et parfois la complexité des images font que si certains outils ont été développés, par exemple dans le cas de formes simples, leur description reste difficile et leur comparaison aussi. Les images binaires sont présentes dans de nombreux processus et leur comparaison s'avère donc utile [Baudrier 05]. Les images en niveaux de gris (256) sont plus riches parce qu'elles contiennent plus d'informations que les images binaires, ce qui demandera une caractérisation et un traitement différents. La plupart du temps, une segmentation de ces dernières rend la comparaison plus aisée, dans un contexte déterminé. Mais d'une manière générale, comparer deux segmentations reste une tâche encore délicate car les objectifs de la segmentation diffèrent d'une application à l'autre. La difficulté de la comparaison augmente avec le nombre de niveaux c'est pourquoi, dans le cas des images "couleurs", la comparaison est d'autant plus dure (*une image couleur comportant en effet plus de* 16 *millions de couleurs*). La difficulté de la comparaison augmente également avec le nombre d'images. En effet, les techniques de recherche d'une image dans une base de données seront différentes de celles utilisées pour la recherche d'un film dans une base de données, ou encore d'un événement particulier dans une séquence d'images (vidéo-surveillance).

Le terme "***comparer***" signifie littéralement "*examiner, établir les ressemblances ou les différences qui existent entre des personnes ou des objets*". Comparer des images signifie donc établir des ressemblances ou des différences entre les images. Cela soulève plusieurs problèmes :
- Le premier est le **choix de l'espace** dans lequel est effectuée la comparaison. Devons-nous comparer l'ensemble de l'image, pixel à pixel, ou devons-nous extraire des caractéristiques de l'image, de la séquence d'images puis, comparer ces caractéristiques entre elles ? Ces caractéristiques devront-elles être numériques ou symboliques ? En effet, pour des applications qui peuvent être coûteuses en temps de calcul, il est intéressant de ne comparer que ce qui est nécessaire si l'on recherche, par exemple, la rapidité. Ce problème donne des contraintes particulières qui demandent une étude approfondie.
- Le deuxième problème est **la manière de comparer** les images, par rapport à quelle référence (s'il y en a une), par rapport à quels critères, par rapport à quelle mesure, qui réalise cette comparaison, la machine, l'homme ? Cette comparaison donne-t-elle un résultat quantitatif ou qualitatif ?
- Enfin le troisième porte sur **l'exploitation de cette comparaison**. A quoi sert le résultat de cette comparaison ? Qui en a l'usage ? Le résultat de la comparaison sert-il à ajuster les paramètres de la méthode ? La méthode fournit-elle des résultats exploitables par la machine, exploitables par un expert ?

Ces trois problèmes soulevés par la **comparaison** d'images ou de séquences d'images vont être abordés dans ce mémoire.

Notons toutefois que le terme "***comparer***" est différent du terme "***évaluer***" qui signifie "*apprécier la valeur, le prix, l'importance d'une chose*". Dans le contexte du traitement des images, le terme "***évaluer***" signifie qualifier un traitement, un algorithme, une méthode, savoir si tel algorithme est plus performant que tel autre. L'évaluation des algorithmes de traitements d'images est un problème qui a également toute son importance de nos jours étant donné l'explosion des méthodes et des techniques dont nous disposons sur le sujet. Nous pouvons trouver une étude approfondie sur cette question dans [Zouagui 04], [Philipp-Foliguet 05a], et [Rosenberger 06]. La comparaison se situe en amont de l'évaluation car pour évaluer, il faut avant tout comparer. Ces travaux sont donc connexes et nous verrons que dans nos travaux, nous avons évalué certains traitements par une méthode de comparaison.

3.2 Les domaines liés à la comparaison d'images

La comparaison d'images intervient dans de multiples domaines mais dans ce mémoire, nous nous sommes intéressés principalement aux trois domaines suivants : **l'évaluation d'algorithmes de traitements d'images**, **la reconnaissance de formes** et **l'indexation de séquences d'images**. Nous positionnerons nos travaux par rapport à ces trois domaines.

3.2.1 L'évaluation d'algorithmes de traitement d'images

Étant donné que comparer revient à établir des ressemblances ou des différences, la comparaison sert donc à l'évaluation. Devant le foisonnement de méthodes développées depuis plusieurs décennies dans les domaines aussi variés que la segmentation, la compression, le tatouage, l'interprétation, la détection d'objets, la reconstruction 3D, l'indexation, ... , il est apparu urgent de définir des stratégies et des méthodes d'évaluation afin que, pour une application donnée, nous puissions utiliser le meilleur algorithme qui soit, ajuster de la meilleure façon les paramètres de la méthode et enfin extraire des informations qui permettent de guider l'utilisateur sur le résultat obtenu.

Différents programmes de recherche ou conférences adoptent cette méthodologie pour identifier l'algorithme le plus efficace de l'état de l'art pour une tâche donnée. Nous citons quelques exemples liés aux traitements des images :

- TREC Video Retrieval Evaluation[1]. (TRECVID) : la série de conférences de TREC est commanditée par le National Institute of Standards and Technology (NIST) avec l'aide et l'appui d'autres organismes gouvernementaux des États-Unis. Le but de la série de conférences est d'encourager la recherche documentaire en fournissant une grande base de données, des procédures de tests et un forum pour les organismes qui souhaitent comparer leurs résultats. En 2001 et 2002, la série de conférences TREC a commandité un axe consacré à la recherche dans la segmentation automatique, l'indexation et la récupération du contenu de vidéo numérique. Depuis 2003, cette voie est devenue une évaluation indépendante (TRECVID) avec un atelier de deux jours ayant lieu juste avant la conférence. Elle met à la disposition des participants des bases de données vidéo, des métriques d'évaluation pour différentes tâches (détection d'un individu dans la vidéo, segmentation en plans...) et des vérités terrain générées par des experts ;
- IEEE International Workshop on Performance Evaluation of Tracking and Surveillance[2] (PETS) : conférence internationale créée en 2000 sur le thème spécifique de la vidéosurveillance et de son évaluation ;
- PASCAL : est un réseau d'excellence européen qui organise notamment une compétition appelée VOC[3]. (de l'anglais "Visual Object Classes") sur la détection et la reconnaissance d'objets en mettant à disposition des participants des vérités-terrain, une base d'images et des métriques ;
- TECHNOVISION[4] : le Ministère de la Recherche et le Ministère de la Défense ont lancé un

[1] http ://www.-nlpir.nist.gov/projects/trecvid
[2] http ://www.cvg.rdg.ac.uk/VS/
[3] http ://www.pascal-network.org/challenges/VOC
[4] http ://www.recherche.gouv.fr/technologie/infotel/technovision.htm

programme (2005-2007) appelé "Technologies de la Vision" visant la création d'une dynamique de l'évaluation de technologies de vision par ordinateur. Différents projets ont été retenus et procèderont, en adoptant l'approche d'évaluation par diagnostic, à la comparaison d'algorithmes de la littérature dans différents domaines (ARGOS[5] : émissions TV et vidéosurveillance, IMAGEVAL[6] : indexation d'images, ROBIN[7] : détection et reconnaissance d'objets, ...).

Comme nous avons pu le voir au travers de ces différents programme d'évaluation, il faut une base de données annotées, une vérité terrain, une métrique et un protocole d'évaluation bien défini. Afin d'avoir le jugement le plus pertinent possible, ***la base de tests*** doit être très importante. ***La référence ou vérité terrain*** associée est par conséquent assez fastidieuse à obtenir. La précision de la référence est aussi primordiale car elle conditionne grandement la qualité de l'évaluation d'un algorithme. Enfin, ***les métriques utilisées*** pour comparer les résultats attendus et ceux obtenus par un algorithme ont une influence non négligeable sur le résultat de l'évaluation. Même si cette approche est sans doute la plus complète, elle nécessite beaucoup de rigueur dans la définition de ces 3 éléments, ce qui en fait une approche lourde à mettre en place.

Les travaux relatés dans [Rosenberger 06] permettent de définir trois axes d'investigation sur l'évaluation des algorithmes de traitements d'images, à savoir :

- **l'évaluation de performances** qui consiste à émettre un jugement qualitatif ou quantitatif sur un algorithme de traitement d'images par la pertinence d'un critère d'évaluation ;
- **l'évaluation par adéquation** qui est très utilisée en ingénierie de conception de système et qui consiste à voir si l'algorithme remplit bien le cahier des charges, d'un point de vue fonctionnel ;
- **l'évaluation par diagnostic** qui consiste à apprécier le comportement d'un algorithme de traitement d'images sur une série de tests dont on connaît la vérité terrain. La métrique utilisée pour faire la comparaison conditionne fortement le résultat de l'évaluation.

Rappelons brièvement les différents critères d'évaluation qui existent :

- **des critères d'évaluation supervisée** qui exploitent une information supplémentaire (par exemple celle d'un expert) pour réaliser un jugement, ce qui accroît la pertinence de l'évaluation au prix d'un effort supplémentaire pour la concevoir.
- **des critères d'évaluation non supervisés** qui sont complètement automatisables, ce qui est un avantage indéniable, mais qui donnent néanmoins un jugement moins fiable. Le principe des critères non supervisés est de quantifier la qualité du résultat d'un traitement d'images à partir d'un calcul statistique.
- **des critères hybrides** qui, de par leurs conceptions, exploitent les avantages des deux axes précédents. L'objectif des critères hybrides est de reproduire une évaluation supervisée par un critère non supervisé.

Un autre axe d'investigation est de **modéliser un traitement** en le découpant en blocs fonctionnels. Dans [Zouagui 04] il est proposé un modèle fonctionnel de segmentation d'images et plusieurs propositions d'évaluation de la segmentation d'images. Le modèle proposé est composé de cinq blocs fonctionnels élémentaires (Mesures, Critère, Evolution et Modification) enchaînés au cours d'un processus itératif. Ce modèle a été appliqué à deux méthodes de segmentation d'images, l'une basée sur un contour actif de type "bulle discrète" et l'autre, sur une approche Markovienne. Un tel modèle permet, selon les auteurs, de rendre plus lisibles les méthodes de segmentation et offre des perspectives intéressantes pour faciliter le choix, le développement et l'implantation du processus de segmentation pour des applications de vision. L'objectif étant d'arriver à définir un modèle générique pour la segmentation d'images qui, idéalement, devrait permettre de représenter n'importe quelle méthode existante et donc faciliter la **comparaison structurelle des méthodes**, l'évaluation de l'originalité d'une méthode, la proposition de nouvelles méthodes, l'implantation logicielle des nombreuses

[5] http ://www.irit.fr/argos
[6] http ://www.imageval.org
[7] http ://robin.inrialpes.fr

méthodes existantes en factorisant les sous-structures communes à plusieurs méthodes, la mise en place, l'évaluation et donc l'optimisation d'une méthode destinée à une application particulière.

Quelles que soient les méthodologies utilisées, il est très important de définir des **critères d'évaluation** et des **mesures de comparaison**, à la base de tout jugement quantitatif ou qualitatif d'une méthode ou d'un traitement. Nous avons proposé dans [Coquin 01a] une mesure de dissimilarité entre images à niveaux de gris et l'avons appliquée comme mesure globale permettant de quantifier, tant spatialement que radiométriquement, les déformations apportées par différents traitements d'images comme le filtrage et la compression d'images [Coquin 01b]. Nous avons étendu cette mesure de dissimilarité au cas des images couleurs [Coquin 00b]. Cette partie sera développée dans le chapitre 5. Récemment, nous avons proposé une mesure de dissimilarité entre plans, scènes, pour construire automatiquement des résumés à partir d'une séquence d'images [Ott 07]. Nous avons également participé partiellement à la campagne d'évaluation ARGOS pour tester nos méthodes de détection de "cut" [ARGOS 06], et de découpage en plans des séquences de films d'animation. Nous présenterons les différents résultats que nous avons obtenus dans le chapitre 5. Nos travaux sur la dissimilarité entre images à niveaux de gris ou couleurs ne sont pas suffisamment rapides pour servir de métrique lors de campagnes d'évaluation de méthodes comme, par exemple, dans TECHNOVISION, TRECVID .... En effet, durant ces campagnes d'évaluation, les métriques utilisées sont, le plus souvent, basées sur la précision (mesure de la quantité de fausse détection) et le rappel (mesure de la quantité de bonne détection), et non sur une mesure pixel à pixel.

3.2.2 La reconnaissance de formes

Comme nous l'avons évoqué au chapitre 2, il existe un ***fossé sémantique*** entre l'homme et l'ordinateur. En effet, l'homme est capable de **reconnaître** dans l'image de la figure 3.1 une voiture et pour les habitués une "Jaguar Type E". Par contre, l'ordinateur aura beaucoup de difficulté à reconnaître qu'il s'agit d'une voiture. Pour cela, il faudra doter l'ordinateur de facultés d'analyse et de jugement. Il faudra lui donner des outils lui permettant d'extraire la voiture de l'ensemble du paysage (constituant le fond) et des paramètres de formes nécessaires à la reconnaissance d'une voiture.

FIG. 3.1: Reconnaissance d'objets.

Si le rôle de l'intelligence artificielle est d'opérer dans les domaines de la connaissance et de la compréhension, celui de la reconnaissance de formes est de simuler la perception. Son but est de doter l'ordinateur d'"organes" de sens, pour capter l'information extérieure sous des formes variées, la simplifier et la catégoriser.

L'objet de la reconnaissance des formes est donc l'identification automatique d'une forme à partir d'un vecteur de paramètres qui la représente. Derrière cette définition simpliste se cachent les difficultés suivantes : comment saisir l'information, comment la représenter convenablement, comment décider dans cet espace de représentation à quelle classe elle appartient ; ces classes devront avoir été obtenues par apprentissage préalable dans cet espace. C'est l'ensemble de ces problèmes que les techniques de reconnaissance de formes tentent de résoudre [Duda 01] [Suykens 03].

L'approche couramment suivie pour construire un module de reconnaissance des formes s'articule autour des trois phases suivantes :

- une phase de **localisation** de l'objet, dont le but est d'isoler l'objet à analyser de son environnement,
- une phase de **représentation** qui consiste à extraire un ensemble de traits caractéristiques. Le but de cette phase est de représenter l'objet par un même ensemble mathématique, tel qu'un point dans l'espace $\mathbb{R}^p$ (p étant le nombre de mesures servant à caractériser l'objet), un mot d'une grammaire, ...
- une phase de **décision** dans laquelle on se référera à une partition de l'ensemble mathématique pour décider si l'élément représentant l'objet appartient à telle ou telle classe.

La localisation de la forme à reconnaître nécessite l'acquisition d'informations extérieures à l'aide de capteurs les plus précis possible. Cela permet de préserver les subtilités du phénomène étudié. La difficulté qui surgit d'emblée est que la représentation directement issue des instruments de mesure est extraordinairement volumineuse. La redondance des données, définie *a posteriori*, subsiste une fois la reconnaissance effectuée mais nous ne savons pas *a priori* situer la redondance [Hyvärinen 01]. Il n'en reste pas moins qu'il faille transformer les données issues des capteurs pour se placer dans un espace de représentation permettant les calculs. En pratique cela signifie deux choses :

- la représentation doit être économique (en terme d'espace mémoire),
- l'espace doit posséder des propriétés mathématiques suffisantes pour que l'on puisse y effectuer des opérations, faire de l'apprentissage (si nécessaire) et prendre des décisions.

Il faut donc trouver un espace de représentation dans lequel la complexité des calculs à effectuer soit raisonnable [Hagedoorn 99]. Le choix de cet espace est un compromis entre, d'une part, les possibilités mathématiques et algorithmiques qu'il offre et, d'autre part, la fidélité de la description des formes originales qu'il assure. Une définition de ces propriétés souhaitables est donnée dans [Devijver 82]. L'étude systématique des espaces de représentation en fonction de leurs propriétés mathématiques que l'on cite ci-dessous est développée dans [Simon 84].

- deux observations successives de la même forme doivent produire le même point dans l'espace de représentation,
- deux formes quelconques différentes doivent être représentées par deux points différents,
- une légère distorsion sur une forme doit conduire à un léger déplacement de sa représentation, par rapport à une métrique donnée, pour plus de robustesse.

Il existe deux grandes familles d'espaces de représentation, qui induisent deux types de méthodologies en reconnaissance de formes :

- **les méthodes statistiques** : on effectue p mesures sur une forme qui est représentée par un vecteur de l'espace $\mathbb{R}^p$. La décision d'appartenance à une classe se fait en utilisant les propriétés des distributions des distances dans cet espace (modélisation d'une classe par un nuage de points de type gaussien par exemple), ou partitionnement *via* des réseaux de neurones, etc.
- **les méthodes structurelles** : la reconnaissance se fait, non par projection dans un espace vectoriel, mais en cherchant à traduire la structure dans un espace de représentation adapté. Par structure, on entend ici des lois d'assemblages des éléments de base pour former un ensemble construit. Les outils dont on dispose ne viennent plus des statistiques mais de domaines comme la théorie des langages, la théorie des graphes, etc. Ce qui est déterminant pour les méthodes structurelles, c'est la façon dont les primitives se composent pour structurer la forme totale.

Parmi toutes les techniques de classification, on distingue :

- La **classification supervisée** dans laquelle un expert a fourni le modèle exact des classes à obtenir,
- La **classification non supervisée** dans laquelle le nombre de classes, inconnu *a priori*, est déduit directement des données.

On distingue ensuite :
- La **classification avec apprentissage** dans laquelle on entraîne le classifieur à l'aide d'un ensemble de données connues *a priori*. Cet entraînement a pour but d'adapter les sorties du classifieur en fonction des entrées qu'on lui soumet.
- **La classification sans apprentissage** où le classifieur travaille directement sur les données sans aucune connaissance préalable.

On trouve enfin des méthodes de classification :
- **paramétriques** dans lesquelles on fournit un certain nombre de paramètres au classifieur qui vont influencer le résultat de la classification.
- **non-paramétriques** dans lesquelles le classifieur doit se débrouiller seul pour classer les données sans aide extérieure.

Dans le cas de la classification supervisée, un expert identifie les classes de données parmi lesquelles on classe ensuite les données existantes. Dans le cas non supervisé, les classes sont construites en fonction des données, selon l'algorithme de classification employé.

Dans le contexte de la reconnaissance de formes, différents axes ont été développés : une méthode générale consiste à extraire des descripteurs d'images ou de formes qui soient invariants à différentes transformations et à **comparer** ces descripteurs entre eux par une mesure de similarité.

Notre contribution : nous avons été amenés à travailler dans le domaine de la reconnaissance de formes pour les deux applications suivantes consistant :
- à analyser des **gestes dynamiques** de la main,
- à détecter et à suivre une voiture à l'entrée d'un parking pour commander l'ouverture de la porte d'entrée après l'identification (collaboration avec la société de vidéo-surveillance EBOO-SOLUTION[8]).

L'ensemble de ces deux applications ont en commun l'analyse ***de séquences d'images***. Le volume de données est donc très important. Pour ces deux applications, nous avons développé la même stratégie à savoir : **isoler** la forme recherchée, trouver des **descripteurs** permettant de bien la caractériser, définir une **mesure de dissimilarité** qui permettent de prendre la décision finale, lors de l'étape de classification. Ces deux applications seront développées dans le chapitre 5.

3.2.3 L'indexation d'images et de séquences d'images

Les premiers systèmes de recherche d'images utilisaient des **mots-clés** associés aux images pour les caractériser. Les méthodes basées sur la recherche textuelle permettaient de retrouver les images contenant ces mots-clés. Plusieurs moteurs de recherche comme par exemple, Google[9] et Lycos[10], proposent ces recherches d'images basées sur le texte. Ils s'appuient sur le principe simple que dans une page web, il y a une forte corrélation entre le texte et les images présentes. Le principal problème de ces recherches par mots-clés est le risque d'un résultat complètement hors sujet.

Une solution consiste à ne pas utiliser les mots-clés et donc à considérer l'image et uniquement l'image pour effectuer les recherches. Cette méthode s'appelle la recherche d'images par le **contenu** ou CBIR (Content-Based Image Retrieval). En règle générale, les systèmes de recherche d'images par le contenu fonctionnent par **comparaison** d'un vecteur descripteur d'une image requête avec les vecteurs descripteurs des images de la base selon une métrique donnée.

Google avait inauguré, il y a quelque temps, un concept à la fois étonnant et ludique pour tenter d'améliorer ses performances dans ses résultats de recherche d'images. Pour rappel, le principe

[8] http ://www.eboo-solutions.com/index.htm
[9] http ://image.google.fr/
[10] http ://www.lycos.fr

consistait à faire participer deux internautes dans le cadre d'un jeu afin d'optimiser l'indexation des images disponibles sur le web. Un grand nombre d'images sont ainsi validées par une intervention humaine qui, jusqu'à l'heure actuelle, reste la meilleure option pour indexer des images. Toutefois, le travail est considérable et une masse importante de contenus graphiques sur Internet restera toujours à indexer.

Plusieurs chercheurs de l'université américaine de Pennsylvanie ont alors développé un système de reconnaissance et de description automatisée d'images. L'ALIPR (Automatic Linguistic Indexing of Pictures Real-Time) est un système qui décrit les images à l'aide d'un vocabulaire de plus de 300 mots, pour l'instant en langue anglaise. Ces mots-clés permettent de référencer les images avec des tags spécifiques sans intervention humaine.

Le processus d'indexation de l'image se déroule en **comparant** l'image à plusieurs dizaines de milliers d'autres contenues dans une base de données. L'analyse suggère alors 15 mots-clés correspondant au mieux à l'image. James Wang, professeur à l'université d'Etat de Pennsylvanie, déclare "avoir entraîné le système de reconnaissance à reconnaître des concepts et des objets afin d'indexer automatiquement des images rencontrées pour la première fois". Il ajoute que "la grande majorité des analyses donne 15 tags décrivant correctement l'image", ce qui permet une grande précision dans la recherche [Li 06].

Un des problèmes centraux de l'indexation d'images par le contenu est la difficulté que pose le choix d'une représentation pertinente des images pour la création de primitives visuelles significatives et fiables capables de traduire le contenu sémantique de la base. Le but est de **mesurer la ressemblance** avec les descripteurs correspondant à la requête. En général, ces primitives sont regroupées en trois classes : les descripteurs liés à la **couleur** (l'histogramme couleur), les descripteurs de **textures** (matrice de coocurrence, indices de direction principale et de rugosité, filtre de Gabor, ondelettes) et les descripteurs de **formes** (descripteurs de Fourier, les moments, points caractéristiques). L'extraction de primitives pertinentes est un problème qui ne connaît pas de solution dans le cas général car il semble qu'il n'existe pas d'attributs qui puissent modéliser une base selon tous les points de vues (pour des raisons multiples qui tiennent à la subjectivité de la requête de l'utilisateur).

Décrivons brièvement les différents systèmes de recherche d'images CBIR ("Content-Based Image Retrieval") et les caractéristiques de chacun d'eux. Historiquement, le premier système de recherche est **QBIC** [Flickner 95] d'IBM dans lequel la recherche est basée sur l'indexation des textures des régions d'images dans l'espace couleur de Munsell amélioré. **Photobook** [Piccard 96], du MIT Media Lab (Massachussets Institute of Technology), propose une recherche possible sur trois critères différents : l'apparence, le contour et la texture. Les deux premiers critères utilisent une décomposition de Karhunen-Loève des régions des images de la base. A partir de cette transformée, les vecteurs propres des images sont utilisés pour la recherche et l'indexation. En ce qui concerne la recherche de texture, elle est basée sur une localisation des pics de fréquence de la transformée de Fourier des images. Ce système est très efficace pour les bases d'images spécialisées. Dans le système **Virage** [Bach 96], la localisation spatiale des couleurs associée à la détection de textures des régions peut être pondérée pour affiner les résultats d'une recherche. Le système **BlobWorld** [Carson 99] de l'université de Berkeley en Californie travaille sur des régions homogènes issues de l'image en procédant à une recherche à partir d'une région d'apprentissage. **Cortina**[11] [Quack 04] utilise des descripteurs issus de la norme MPEG-7 et des mots issus du texte autour des images dans les pages web, pour construire son index d'images. Le regroupement d'images est réalisé avec l'algorithme des k-plus proches voisins. La requête est soit une requête par mot-clé, soit une requête par image exemple. **Imedia**[12] est un système de recherche développé à l'INRIA. Il est basé sur une recherche à partir d'une image requête utilisant la couleur, la texture et la forme. Un système de bouclage de pertinence adapte la recherche suivant les modifications apportées aux résultats par l'utilisateur au cours de la recherche. **Kiwi**[13] est un système développé à l'INSA de Lyon. Il est basé sur une analyse des images et l'extraction de points d'intérêt multirésolution des images. Le système **Retin** a été développé à l'ENSEA de Cergy-Pontoise. Il utilise des attributs de couleur dans l'espace Lab et des attributs de

[11] http ://vision.ece.ucsb.edu
[12] http ://www-rocq.inria.fr/imedia//cbir-demo.html
[13] http ://telesun.insa-lyon.fr/kiwi

texture (filtres de Gabor) [Fournier 01] [Fournier 02]. Sa principale force est son système très efficace de bouclage de pertinence. **Windsurf** [Ardizzoni 99] est basé sur la décomposition en ondelettes des images, suivie d'une segmentation des régions à l'aide des nuées dynamiques, et d'une extraction d'attributs de couleur et de texture. Les régions de l'image requête sont ensuite comparées selon la distance de Mahalanobis pour donner les images les plus proches de la requête. Tous ces systèmes sont malheureusement difficilement comparables car ils travaillent tous avec une base d'images différente. Ils offrent tous une interface conviviale à l'utilisateur qui doit néanmoins entrer quelques paramètres avant de lancer la recherche.

La difficulté dans les systèmes de recherche d'images par le contenu est d'associer une valeur sémantique à une image. À partir des pixels qui représentent une information bas-niveau, il est très difficile d'arriver à l'interprétation haut-niveau de l'image. La figure 3.2 montre à quel point ce pas est difficile à franchir puisqu'à l'heure actuelle, reconnaître la présence ou l'absence d'un certain animal dans une image est un problème encore difficile à résoudre. Dans l'étape de segmentation, les pixels sont associés pour former des régions de différentes textures. Ces textures définissent les objets qui composent la scène conduisant à l'interprétation de l'image. Ainsi, donner un sens à une image signifie qu'à partir d'une suite de pixels, on va être capable de définir les objets présents dans la scène. Or il n'existe pas de technique de reconnaissance capable de recréer ce processus d'analyse qu'un enfant de trois ans arrive à faire au premier coup d'œil.

FIG. 3.2: Illustration du fossé sémantique.

Les attributs vont représenter l'image et dépendent du contexte. Il existe donc deux familles d'attributs, les attributs globaux qui sont calculés à partir de l'image entière, et les attributs locaux qui sont calculés sur une région de l'image considérée. Leur choix est déterminant pour la suite de la méthode. Si les attributs sont mal choisis, la méthode de classification donnera de mauvais résultats. Comment choisir de bons attributs ? Il n'y a pas de réponse générale à cette question car le choix des attributs va dépendre de ce que l'on souhaite classer. Les attributs sont en général choisis par un expert du domaine des images de la base. L'expert justifie le choix des attributs par son expérience pour les caractéristiques qui lui semblent importantes et pour le champ applicatif de la méthode de recherche. Suivant l'application développée, l'expert pourra choisir différents attributs. Le choix des attributs est fortement dépendant des images de la base. Ainsi, les attributs qui donnent d'excellents résultats sur une base d'images peuvent donner des résultats médiocres sur une autre base. Il n'y a pas d'attributs universels donnant de bons résultats sur n'importe quelle base d'images.

Les systèmes d'indexation de séquences d'images ou CBISR ("Content-Based Image Sequence Retrieval") correspondent à l'extension temporelle des systèmes CBIR. Le premier problème qui se pose est la taille des données. En effet, un film (ou une séquences d'images) à lui seul est équivalent à une base contenant plusieurs dizaines de milliers d'images. Si dans un système CBIR deux images contenant 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 considère 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. Mais en plus, ces images ne peuvent pas être

considérées indépendamment les unes des autres, car elles sont liées temporellement. 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.

Dans la thèse de **Bogdan Ionescu**, nous avons travaillé sur des indicateurs sémantiques/symboliques pour l'indexation de films d'animation du CICA. Ces travaux présentent une première réflexion sur le sujet et seront détaillés dans le chapitre 6.

Avant tout, dressons un rapide panorama des méthodes de comparaison des images. L'objectif de cette partie étant de présenter les principales méthodes de comparaison d'images et, étant donné, la diversité et le nombre important de travaux disponibles dans le domaine, nous n'avons pas jugé utile de réaliser une bibliographie exhaustive. Nous pensons que cette partie doit plutôt rassembler les références pertinentes permettant d'appuyer les propos de ce mémoire.

3.3 Les méthodes de comparaison d'images

Les méthodes de comparaison d'images reposent le plus souvent sur la sélection de descripteurs discriminants dans le cadre d'une base d'images donnée. Les descripteurs donnent en général de l'information sur la couleur, la texture, et les formes extraites de l'image [Smeulders 00b]. Leur choix, qui conditionne l'efficacité de la méthode, constitue une étape délicate de l'indexation [Antani 02b]. Ces descripteurs sont aussi utilisés pour les images en niveaux de gris à l'exception, bien sûr, de celui de la couleur. Une fois les descripteurs sélectionnés, ils sont agencés pour chaque image afin de former une signature. Ces signatures permettent d'indexer la base et sont donc utilisées pour la recherche d'images. Les signatures des images de la base sont alors comparées à celle de la requête à l'aide d'une **mesure de dissimilarité**. Cette dernière est construite en fonction des descripteurs choisis qui peuvent être traités séparément ou agrégés sous forme d'un vecteur. Détaillons maintenant les descripteurs, les signatures et les mesures de dissimilarité les plus connus.

3.3.1 Les descripteurs des images

La couleur est très souvent le premier descripteur employé pour la comparaison d'images. Plusieurs travaux ont déjà prouvé qu'il s'agissait d'un descripteur efficace [Smeulders 00b]. Une technique très utilisée pour la couleur est l'intersection d'histogrammes [Swain 91]. Les histogrammes sont faciles et rapides à calculer, robustes à la rotation et à la translation. Cependant, l'utilisation d'histogrammes pour l'indexation et la recherche d'images pose cinq problèmes [Gong 98] :

- ils sont de grande taille par conséquent, il est difficile de créer une indexation rapide et efficace en les utilisant tels quels,
- ils ne donnent pas d'informations sur la position des couleurs dans l'image,
- ils sont sensibles à de petits changements de luminosité, ce qui est problématique pour comparer des images similaires acquises dans des conditions différentes,
- ils sont inutilisables pour la comparaison partielle des images (objet particulier dans une image), puisque calculés globalement sur toute l'image.
- ils ne contiennent pas l'information temporelle utile pour traiter les séquences d'images.

Plusieurs approches ont été engagées pour améliorer ces problèmes. La première approche consiste à ajouter des informations spatiales aux histogrammes. Dans [Stricker 96], les auteurs ont divisé une image en cinq blocs se superposant, fixés à l'avance, et ils ont extrait les trois premiers moments

d'inertie de chaque bloc pour créer un vecteur de descripteurs. Pass et Zabih [Pass 96] ont ajouté de la cohérence spatiale dans les histogrammes. Un pixel est cohérent s'il appartient à une région validée par la segmentation et est incohérent sinon (c'est le cas les pixels situés hors des régions segmentées). Les valeurs de l'histogramme sont alors divisées, en fonction des pixels, en deux classes : classe cohérente et classe incohérente. La comparaison entre deux histogrammes revient à faire la comparaison entre les valeurs des histogrammes dans les classes correspondantes. Huang et al. [Huang 97] ont proposé le corrélogramme et l'auto-corrélogramme. [Chen 99] utilise des histogrammes "augmentés" qui sont calculés en ajoutant des informations statistiques, comme la moyenne, l'entropie, la variance, etc. , sur les distances entre les pixels. 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.

La deuxième approche consiste à rechercher d'autres espaces de couleur comme, par exemple, l'espace CIELab, qui soient proches de la perception humaine, l'espace RVB n'étant pas l'espace le mieux adapté. 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 de 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 qu'habituellement, ces méthodes cherchent un compromis entre la qualité de la préservation des couleurs et le temps de calcul, compromis qui dépend du type d'application.

Parmi les autres approches d'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 distribution des rapports entre les couleurs ou "color ratio models" (ces rapports sont calculés sur les contours de l'image).

D'autres méthodes proposent de caractériser la séquence à travers des vecteurs de caractéristiques locales des couleurs de l'image et en étudiant leur évolution temporelle. 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.

Notre contribution porte sur la définition d'une mesure de dissimilarité entre images couleur et sur la définition de critères sémantiques des couleurs en vue de l'indexation de films d'animation.

Les caractéristiques de formes

Les descripteurs de formes sont complémentaires de la description de la couleur. Des études ont été faites pour les rendre robustes aux transformations géométriques comme la translation, la rotation et le changement d'échelle. Nous distinguons deux catégories de descripteurs de formes : les descripteurs basés sur les régions et ceux basés sur les frontières. Les premiers font classiquement référence aux moments invariants [Hu 62] [Derrode 99] et sont utilisés pour caractériser l'intégralité de la forme d'une région. Cependant, il est difficile de relier les valeurs des moments de grand ordre aux caractéristiques locales des formes. La seconde approche porte sur une caractérisation des contours de la forme (coefficients de Fourier, excentricité, nombre d'Euler, ...) [Hagedoorn 99]. Ces méthodes sont tributaires d'une détection de contours (sélection des zones de fortes variations) et pour des images peu contrastées ou bruitées, elles ne sont pas appropriées. Pour les images binaires, les variations de niveaux de gris sont toujours d'amplitude égale à 0 ou à 1. L'information concernant les contours est donc pauvre. La détection peut se faire grâce à la topologie discrète, avec l'utilisation des 4-voisinages ou 8-voisinages [Leonard 91]. Il est possible de distinguer de cette manière l'intérieur et la frontière des objets et de les étiqueter. Cependant, la détection d'objet dans les images binaires n'est pas fiable lorsque les images ne sont pas composées de formes simples par exemple, lorsqu'elles comportent beaucoup de traits. A partir d'une image binaire, il est facile d'extraire le squelette d'une forme. Le squelette est un descripteur très utilisé car c'est une représentation compacte de la forme.

Dans le cas des systèmes d'indexation d'images, les paramètres caractérisant les formes sont analysés dans le domaine spatial de l'image. Dans le cas des séquences d'images, le déplacement des objets dans la scène est une caractéristique importante. Celui-ci 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, etc.) est comparée à l'efficacité des descripteurs basés sur les régions de pixels (moments invariants, moments de Zernike). Pour améliorer l'invariance, une collaboration entre différents types de descripteurs est proposée et testée : moments invariants et descripteurs de Fourier, ou moments invariants, 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.

Notre contribution porte sur la définition de la signature d'un geste dynamique de la main utilisée pour la reconnaissance de gestes dans une séquence d'images.

Les caractéristiques de texture

La texture a plusieurs définitions. La notion de texture est utilisée pour traduire un aspect homogène de la surface des niveaux de gris d'un objet sur une image. Dans [Unser 95], l'auteur présente la texture comme une structure disposant de certaines propriétés spatiales homogènes et invariantes par translation. Dans [Gagalowicz 83], l'auteur considère la texture comme "*une structure spatiale constituée de l'organisation de primitives ayant chacune un aspect aléatoire, donc une structure hiérarchique à deux niveaux*". La prise en compte de ce type de caractéristiques pour représenter globalement ou partiellement une image est courante et discriminante dans de nombreux cas. A titre d'exemple, ce type d'information est utilisé pour l'indexation d'images satellitaires [Li 97]. Les descripteurs de texture en indexation apparaissent dans le système de recherche d'images par le contenu d'IBM QBIC [Niblack 93]. Les caractéristiques en question sont le grain, le contraste et l'orientation. Les travaux d'Aksoy et Haralick [Aksoy 98] sur la texture et en particulier, les matrices de co-occurrences et les différents indices qui peuvent en découler ont également servi de base pour l'indexation. De nombreuses méthodes sont référencées dans la littérature pour la décomposition de l'image et le calcul de caractéristiques dites de texture. Parmi les plus connues, on trouve la décomposition paramétrique Wold 2D d'abord utilisée par Liu et Picard [Liu 96] puis reprise notamment dans le système Photobook [Pentland 96] ou encore par Stoica et al. [Stoica 98]. Afin d'estimer la similarité entre des matrices de co-occurrences, quatre caractéristiques extraites de ces matrices sont largement utilisées : l'énergie, l'entropie, le contraste et le moment inverse de différence. Il existe aussi d'autres méthodes pour analyser les textures dont celles basées sur les filtres de Gabor et celles basées sur les décompositions en ondelettes. Après avoir appliqué la transformation de Gabor sur une image, une région de texture est caractérisée par la moyenne et la variance des coefficients de transformation. Un vecteur de caractéristiques est construit en utilisant ces caractéristiques comme composants.

Notre contribution : dans le projet BQR[14], nous avons analysé la structure des fibres dans des images tomographiques 3D (une radiologie d'une pièce, en quelque sorte). Pour ce faire, nous avons utilisé les matrices de co-occurrences pour localiser les zones de porosité, les zones orientées, les zones non-orientées et les zones de manque de renfort, en calculant des paramètres liés à l'homogénéité et à l'orientation des fibres. Nous avons mis en œuvre une méthode qui permet de guider l'utilisateur vers le bon choix des paramètres de réglage, permettant d'obtenir de meilleurs résultats [Valet 07].

3.3.2 Les signatures des images

Une fois les descripteurs discriminants extraits, ils sont ordonnés pour former la signature de l'image. C'est la structuration de ces descripteurs qui va former la signature d'une image. Les signatures peuvent servir à la **comparaison des images** [Philipp-Foliguet 05b], et cette comparaison doit

[14]Système coopératif de fusion d'information pour l'interprétation d'images 3D, 2006

rendre compte du degré de similarité entre les images. Deux types de structuration de l'information pour former une signature sont possibles : la première synthétise l'information de toute l'image et forme donc une **signature globale** ; la seconde tient compte des différences au sein même de l'image et rend compte de l'**information locale et partielle**.

Un système basé uniquement sur des caractéristiques globales ne peut pas donner les résultats désirés. En effet, si nous considérons une image composée de plusieurs objets ayant des caractéristiques (couleur et texture) très différentes, le vecteur de caractéristiques global extrait à partir de l'image entière perd les informations locales (les objets) et ne produit qu'une moyenne grossière du contenu de cette image. Inversement, l'analyse uniquement basée sur des caractéristiques locales risque de perdre le sens global de l'image, en submergeant celui-ci dans un flot de petits détails inutiles. Par conséquent, un compromis doit être trouvé, différent selon les applications et les requêtes individuelles, entre caractéristiques globales et caractéristiques locales.

Signature globale : L'exemple le plus simple et le plus courant de signature est l'histogramme [Brunelli 01] qui représente une approximation de la densité de probabilité de l'image dont l'intensité est vue comme une variable aléatoire. Les histogrammes sont très utilisés en recherche par le contenu, car l'histogramme d'une image est presque invariant aux rotation, translation, et changement d'échelle. Mais le principal inconvénient est la perte de l'information spatiale. Des travaux ont été entrepris pour ajouter cette information [Kharbouche 05].

Signature locale : Deux approches différentes peuvent être employées pour calculer les caractéristiques locales. La première approche consiste à diviser une image en utilisant une grille. Des caractéristiques sont alors calculées pour chaque case de cette grille. La deuxième approche consiste à segmenter l'image pour la diviser en zones locales plus proches des objets constituant l'image et ensuite calculer les caractéristiques pour chacune des régions extraites. La segmentation est une étape qu'on souhaiterait éviter en raison de tous les problèmes que pose le choix d'une "bonne" méthode de segmentation qui serait valide pour toutes les images de la base. Cependant, la division d'une image en régions ou objets d'intérêt est souvent nécessaire pour pouvoir s'attaquer ensuite à l'extraction d'informations sémantiques à partir de l'image. Plutôt que de viser une segmentation exacte (en terme d'interprétation de l'image), on préfère souvent parler de "*groupement de pixels*" [Forsyth 03] ou encore, selon les auteurs [Medioni 05] de "*segmentation faible*". Dans les deux cas, il s'agit simplement d'oublier le découpage précis en terme d'objets, pour s'intéresser au découpage de l'image en régions similaires relativement aux caractéristiques mesurées sur l'image. Cette dernière idée apparaît plus logique car l'idée même d'objet est une vue sémantique de l'image. Il est facile de conclure à la lumière des innombrables travaux en segmentation des dernières décennies que segmentation bas niveau (basée uniquement sur les caractéristiques issues de l'image) et découpage sémantique de l'image sont deux idées complètement différentes. Un objet tel que nous le percevons dans l'image n'est pas forcément homogène en termes de caractéristiques, tandis qu'une zone homogène de l'image peut très bien contenir plusieurs objets différents.

Une nouvelle représentation d'images basée sur des chaînes de symboles a été introduite dans [Simand 05]. La signature, mêlant les notions de points d'intérêt, de contraste, et d'ordre, offre une description à la fois concise et précise de l'image. La comparaison de deux images consiste en la comparaison des symboles qui composent les chaînes des signatures.

Il existe de nombreuses signatures d'image, celles-ci dépendent le plus souvent de l'application pour laquelle elles ont été construites.

Notre contribution : Dans le cadre du projet BQR[15], nous avons travaillé sur les mécanismes de **reconnaissance de gestes** dynamiques et de postures. Nous avons proposé une nouvelle méthode de reconnaissance d'un geste dynamique de la main. Un geste dynamique est caractérisé par les signatures statiques de début et fin de geste (histogramme des orientations du gradient) et par la signature dynamique (superposition des squelettes de chacune des images composant la séquence. La reconnaissance du geste se fait par une **mesure de similitude** entre ces signatures et des signatures d'un alphabet connu [Ionescu 03] et [Ionescu 05]. De plus amples détails seront donnés dans le chapitre

[15] Etude de postes de travail interactifs : 2002-2004

5.

Dans le cadre de l'**indexation de séquences d'images**, chaque séquence d'images peut être caractérisée par la distribution des couleurs prédominantes et l'agencement de celles-ci, par des paramètres liés au rythme, à l'action, par la fréquence des changements de plans, etc. En faisant la classification des séquences par rapport aux couleurs et au rythme, nous nous sommes aperçus que ces caractéristiques pouvaient servir de signature artistique d'un film. Le réalisateur d'un film d'animation emploie en effet, d'un film à l'autre, toujours la même technique de réalisation et approximativement les mêmes couleurs suivant ce qu'il cherche à exprimer. La figure 3.3 illustre le film "Le moine et le poisson". Des détails sont donnés dans la thèse de Bogdan IONESCU [Ionescu 07].

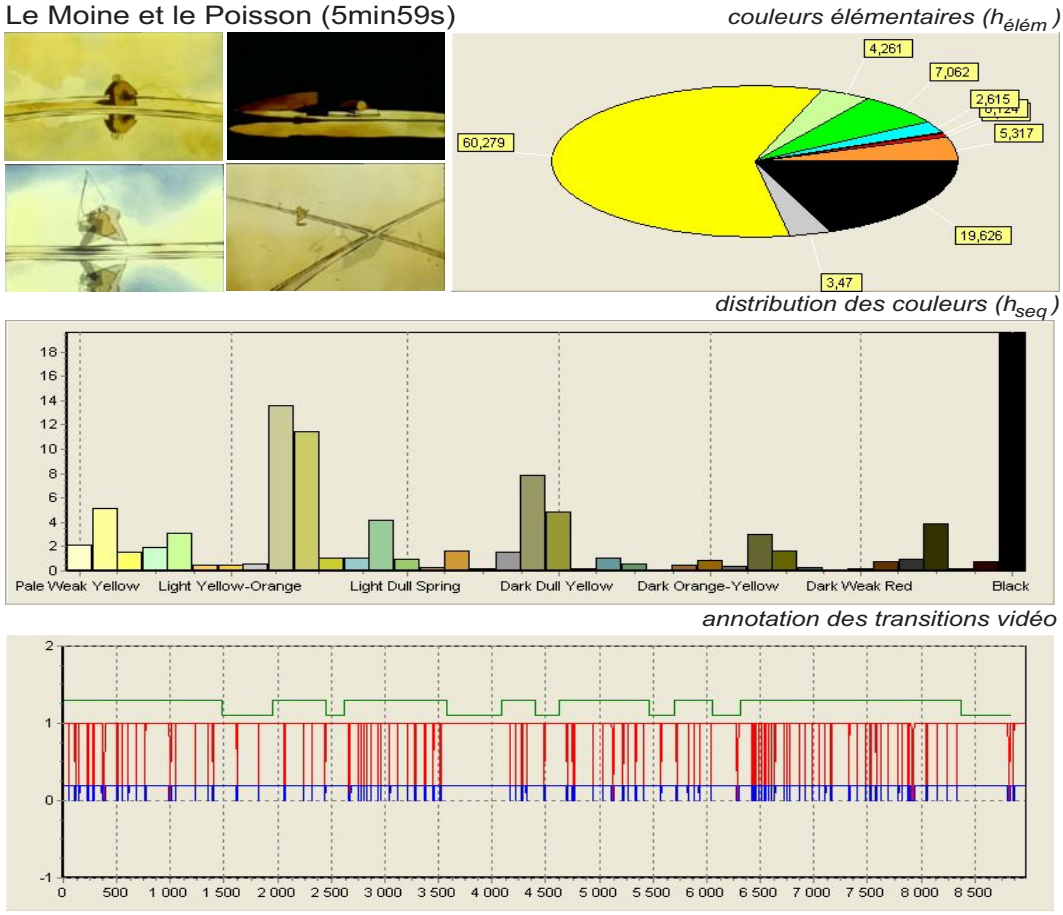


FIG. 3.3: Film "Le Moine et le Poisson" : histogramme des couleurs élémentaires, histogramme global pondéré et annotation visuelle des transitions.

Notons également qu'une tendance actuelle en recherche des images est la modélisation des images ou des objets segmentés par des graphes d'adjacence des régions qui engendre la comparaison des images par des méthodes de comparaison de graphes [Galmar 05]. Enfin pour la caractérisation des séquences d'images, l'utilisation des descripteurs normalisés $MPEG7$ est également une piste à envisager comme le montre par exemple les travaux de Kompatsiaris [Mezaris 04].

3.3.3 Les mesures de similarité

La notion de **similarité** entre deux images est une notion difficile à définir. Il faut en effet préciser sur quel(s) critère(s) cette notion de similarité se base. La figure 3.4 présente un exemple où on demande à l'utilisateur de classer les quatre images proposées en deux familles d'images.

Il y a plusieurs réponses acceptables à cette question selon le **critère de similarité** retenu. Si on décide de retenir la **couleur** comme critère, les deux familles sont (A,C) et (B,D). Si c'est la **texture** que nous retenons, on aura (A,B) et (C,D). Si c'est la **forme** que nous choisissons comme critère, les deux familles deviennent (A,D) et (B,C). Enfin, si nous choisissons la **taille**, le regroupement devient (A,C) et (B,D) comme dans le premier cas. Il est aussi possible de définir une composition de ces critères avec un résultat qui dépend de leur importance relative.

Cet exemple est volontairement caricatural mais il illustre à quel point il faut tenir compte du critère de similarité lors de la recherche. Autrement dit, il faut créer un vecteur descripteur contenant les informations selon un ou plusieurs critère(s) choisi(s) pour les besoins de l'utilisateur.

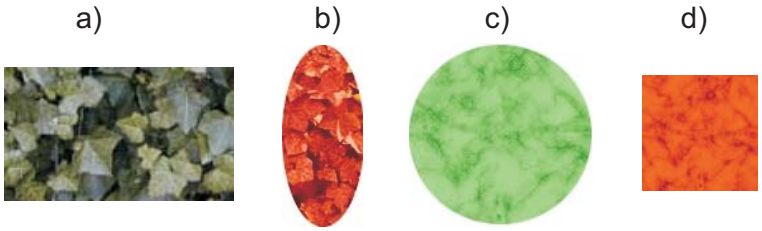


FIG. 3.4: Images à classer en deux catégories.

La **similarité** de deux images est un problème mal posé, qui nécessite pour y répondre correctement des informations supplémentaires. Il y a des choix à faire *a priori* pour résoudre ce problème. Il faut donc définir un **critère de similarité** avant de **comparer** deux images. Dans la définition de la similarité et dans l'extraction des attributs des images, l'expert du domaine joue un rôle très important. Il choisit les caractéristiques des objets et propose un modèle de comparaison des images. Il apporte son expérience du domaine comme connaissance *a priori* de la similarité entre les images.

Illustrons ces propos à la comparaison d'images couleur, et comme mesure de dissimilarité, nous avons choisi la **distance entre leurs histogrammes couleurs**. Pour une image $\mathbf{A}$ quantifiée dans un espace couleur réduit à n classes couleur $(c_1, c_2, \ldots, c_n)$, l'histogramme couleur H est un vecteur à n composantes : $(h_{c_1}, h_{c_2}, \ldots, h_{c_n})$ pour lequel h_{c_j} représente le nombre de pixels de couleur c_j dans l'image A. On a en particulier $\sum_{i=1}^{n} h_{ci} = N$, où N est le nombre de pixels de l'image.

La dissimilarité $Dissim(.)$ entre deux images $\mathbf{A^1}$ et $\mathbf{A^2}$ peut alors s'exprimer selon ces vecteurs couleur *via* toutes les distances géométriques classiques qui permettent de mesurer la différence entre les histogrammes respectifs $\mathbf{H^1}$ et $\mathbf{H^2}$, d'où $Dissim(\mathbf{A^1}, \mathbf{A^2}) = d(\mathbf{H^1}, \mathbf{H^2})$.

Le choix de la fonction de mesure $d(.)$ doit se faire de façon à respecter les propriétés classiques des distances :

- positivité : $d(\mathbf{H^1}, \mathbf{H^2}) \geq 0$,
- identité : $d(\mathbf{H^1}, \mathbf{H^1}) = 0$,
- symétrie : $d(\mathbf{H^1}, \mathbf{H^2}) = d(\mathbf{H^2}, \mathbf{H^1})$,
- inégalité triangulaire : $d(\mathbf{H^1}, \mathbf{H^3}) \leq d(\mathbf{H^1}, \mathbf{H^2}) + d(\mathbf{H^2}, \mathbf{H^3})$.

Parmi les modalités les plus connues, nous retrouvons les normes de type L_1 et L_2, qui appartiennent au cadre classique des distances de Minkowski L_p.

Norme L_1 : $d_{L_1}\left(\mathbf{H^1}, \mathbf{H^2}\right) = \sum_{i=1}^{n} \left|h_{ci}^1 - h_{ci}^2\right|$ appelée également distance de Manhattan.

Norme L_2 : $d_{L_2}\left(\mathbf{H^1}, \mathbf{H^2}\right) = \sqrt{\sum_{i=1}^{n} \left(h_{ci}^1 - h_{ci}^2\right)^2}$

Norme L_∞ : $d_{L_\infty}\left(\mathbf{H^1}, \mathbf{H^2}\right) = \max_{1 \leq i \leq n} \left|h_{ci}^1 - h_{ci}^2\right|$.

Norme L_p : $d_{Lp}\left(\mathbf{H^1}, \mathbf{H^2}\right) = \left(\sum_{i=1}^{n} \left(h_{ci}^1 - h_{ci}^2\right)^p\right)^{1/p}$

Dans la formulation de la norme L_2, nous retrouvons la classique distance euclidienne, alors que dans la norme L_∞, la distance conservée est le plus grand écart mesuré sur l'ensemble des deux vecteurs. *A priori*, ce type de métrique pourrait être adapté, mais cela serait oublier l'aspect perceptuel de telles mesures.

La figure 3.5 permet de mieux comprendre le phénomène que nous allons illustrer sur des images en niveaux de gris. Dans le premier cas (***images a et c***), nous trouvons la même image dont les conditions d'illumination ont été modifiées ; dans le second cas (***images a et b***), nous trouvons deux images différentes. Quelle que soit la métrique utilisée, le second cas est considéré comme étant toujours plus ressemblant, en ne comparant que les histogrammes.

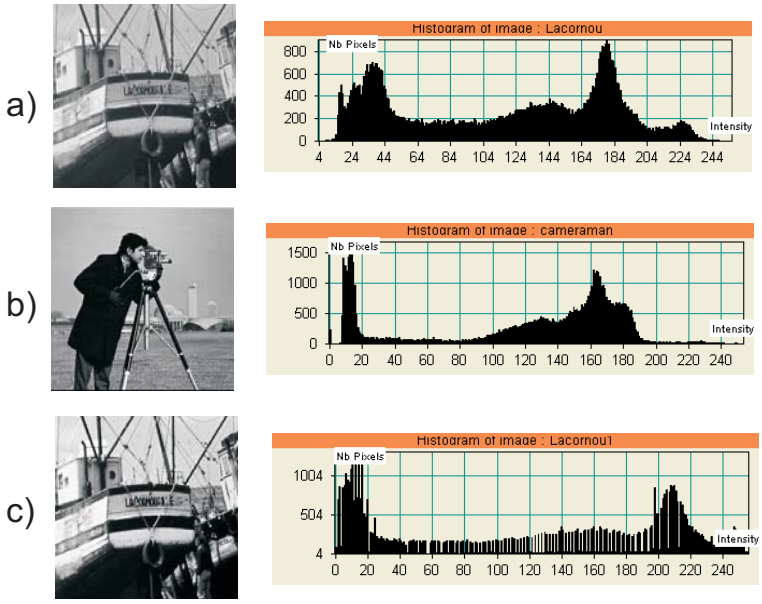


FIG. 3.5: Comparaison d'histogrammes.

Sans étendre davantage le problème de la métrologie entre attributs, ce simple exemple permet de comprendre la complexité d'associations entre un attribut et une métrique d'évaluation, cette association devant dépendre des capacités du système visuel humain pour appréhender ces variations.

Dans la littérature, il existe de nombreuses mesures de similariré que nous ne détaillerons pas ici. Nous avons simplement voulu montrer que le choix d'une mesure de similarité est quelque chose de difficile et que bien souvent, c'est l'application qui va nous guider.

Notre contribution porte sur des travaux entrepris en géométrie discrète relativement à la définition d'opérateurs locaux de distance 2D et 3D adaptés à différents maillages rectangulaires et parallélépipédiques. Ils seront utilisés lors de la définition d'une mesure de dissimilarité entre images tant en niveau de gris qu'en couleur. Ces travaux seront présentés dans le chapitre 5.

3.4 Conclusion

Que ce soit dans nos applications sur l'évaluation de méthodes de traitement d'images (filtrage, compression, ...), sur la reconnaissance des gestes de la main ou sur l'indexation de séquences d'images,

nous avons été amenés, à un instant ou à un autre, à comparer les images. Il nous semble donc nécessaire de développer une réflexion sur la **comparaison des images**. C'est une étape incontournable dans tout système automatique de traitement d'images et dans tout système pour lequel on souhaite améliorer, d'une façon ou d'une autre (par l'interaction d'un expert), le résultat.

Comme nous venons de le voir, nous pouvons **comparer** deux images de **manière directe** c'est-à-dire en définissant une mesure de dissimilarité qui permette de quantifier les différences entre ces deux images. La **comparaison indirecte** à partir d'attributs extraits de ces images est une autre manière de travailler à cet objectif.

Les questions qui restent en suspens sont les suivantes :
- comment choisir les "bons" attributs ?
- comment choisir la mesure de dissimilarité ?
- comment choisir la méthode de comparaison ?
- est-ce une comparaison de type similitude ou exactitude ?
- la comparaison doit-elle être locale ou globale ?
- est-ce une approche bas-niveau (numérique) ou haut-niveau (symbolique ou syntaxique) ?

Dans les chapitres suivants, notre réflexion portera sur la **comparaison directe des images** binaires, en niveaux de gris ou en couleur, et sur la **comparaison indirecte des images** en vue de l'analyse de séquences d'images. A terme, nous envisageons l'extension de ces pistes à l'évaluation des méthodes de fusion d'informations, thème en cours de développement au sein du laboratoire.

Chapitre 4

Les opérateurs locaux de distances

Résumé : *La comparaison directe de deux images passe par la définition d'une mesure de dissimilarité entre images. La mesure de dissimilarité sera basée sur la distance d'un point à un ensemble. Nous utiliserons des opérateurs locaux de distances pour calculer cette distance. Dans cette partie, nous allons présenter les travaux que nous avons développés sur les opérateurs locaux de distances discrètes.*

4.1 Introduction

La notion de distance est très utilisée pour décrire une forme dans une image numérique, pour comparer les formes entre elles, ou pour localiser un objet dans un volume. En analyse d'images, mesurer les distances entre objets est souvent essentiel. Dans la littérature, de nombreuses distances existent et permettent de répondre à ce problème. La distance la plus connue et la plus utilisée par les mathématiciens et les physiciens est la **distance euclidienne** car elle est adaptée au monde continu. Dans le domaine de l'analyse d'images, l'espace de travail est un espace discret. La distance euclidienne n'est donc pas forcément la mieux adaptée.

Tous les systèmes d'acquisition de données d'images fournissent des données organisées sur une grille régulière, appelées ***données discrètes***. Que ce soit pour une visualisation ou pour l'extraction d'une mesure sur ces objets discrets (paramètres de formes), les axiomes et théorèmes de la géométrie euclidienne ne sont pas directement applicables. Deux solutions s'offrent à nous : la première consiste à transposer les données discrètes dans un espace continu où ces théorèmes et ces mesures sont définis en utilisant, par exemple, le processus d'interpolation. Une seconde alternative se base sur une transposition de ces théorèmes et mesures dans l'espace discret. Ces différentes re-définitions sont le fruit de la ***Géométrie Discrète***.

Un premier intérêt à utiliser la géométrie discrète est d'éviter ainsi un changement de modèle coûteux en temps et potentiellement source d'imprécisions. D'un point de vue algorithmique, l'analyse de ces objets sur une grille régulière discrète à l'aide d'une méthodologie en nombres entiers permet, non seulement une gestion de ces incertitudes, mais aussi l'écriture d'algorithmes très efficaces, et réduit l'espace de stockage en mémoire. En effet, la manipulation de nombres entiers et les liens très forts entre les objets discrets et des théorèmes de la théorie des nombres ou l'arithmétique offrent des possibilités d'implémentation d'outils géométriques très efficaces. Derrière ces considérations d'ordre théorique se cache un intérêt pratique très important. En effet, une théorie géométrique et une algorithmique adaptées aux images nous permettent d'accélérer les processus mais surtout de proposer des outils d'analyse de formes cohérents avec la modalité des objets analysés.

Dans ce chapitre, nous rappelons ce qu'est une transformation de distance. Nous positionnerons nos travaux par rapport à ce qui a été fait et montrerons les différentes avancées dans ce domaine.

4.2 État de l'art

Si l'on devait calculer la distance de chacun des points de l'objet d'une image binaire par rapport à chacun des points du fond, on obtiendrait un coût algorithmique beaucoup trop élevé par rapport aux exigences du traitement d'images. De nombreux auteurs se sont donc penchés sur des alternatives au calcul exhaustif des cartes de distance. Pour cela, ils utilisent la cohérence spatiale d'une carte de distance qui permet la notion de propagation d'information locale. Aussi, de nombreuses transformations de distances ont été proposées dans la littérature. Une analyse bibliographique très complète est proposée dans [Cuisenaire 99a] et dans [Thiel 01]. Nous proposons ici un rapide rappel de l'état de l'art dans le domaine.

4.2.1 Transformation de distances

Le but d'une transformation de distance est de calculer la distance d'un point d'un objet à un ensemble de points de référence (par exemple le fond). La distance d'un point p de l'objet est la plus petite distance de p à n'importe quel point appartenant au fond. En d'autres termes, c'est la distance du point p au plus proche point q appartenant au fond.

Soit d une distance définie dans le plan discret $\mathbb{Z}^2$. Étant donné un pixel p et un ensemble X de pixels de $\mathbb{Z}^2$, il n'y a qu'un nombre fini de pixels x de X tels que $d(p,x)$ soit minimale. On définit ainsi la distance $d(p,X)$ de p à X comme le minimum de la distance de p aux pixels de X :

$$d(p,X) = \min\{d(p,x)/x \in X\} \tag{4.1}$$

On a :
- $d(p,X) = 0$ si et seulement si $p \in X$
- pour tout pixel $q \in X$, $d(p,X) \leq d(p,q) + d(q,X)$.

Une **transformation de distance**, notée **DT** (*pour Distance Transformation*) est l'opération qui consiste à calculer une image de distance à partir d'un ensemble de référence. Une fonction de distance est souvent définie par une formule directe sur les coordonnées des points ou par des opérations ensemblistes. Pour une DT, il peut être plus intéressant d'exprimer la distance comme le coût d'un chemin minimal sur un certain graphe pondéré. Ce graphe désigne une famille de chemins et associe un coût à chacun de ses membres. Un premier algorithme de DT est le calcul exhaustif qui, pour chaque point de l'image, teste tous les autres points pour trouver le minimum. Cette méthode est possible pour toute distance d définie par une formule directe, mais extrêmement coûteuse (en $O(L^{2n})$ pour une image de côté L en dimension n). L'idée de base pour obtenir une DT efficace est de déterminer globalement le minimum des distances par rapport à l'ensemble de référence (le fond), en propageant, à partir du contour de l'objet, des distances locales entre pixels proches. Pour faciliter la propagation locale lors de la transformation, on privilégie souvent les distances définies par chemin minimal. On peut alors déduire la valeur de distance au fond en chaque point p de l'objet, à partir des valeurs déjà connues pour chaque voisin et ceci, en prolongeant leur chemin minimal respectif jusqu'à p (algorithme de Dijkstra [Dijkstra 59]).

Il n'existe pas d'algorithme universel qui soit efficace pour toutes les familles de distances, mais des algorithmes spécifiques. On distingue les algorithmes parallèles (itératifs jusqu'à convergence) des algorithmes séquentiels (par balayages).

La complexité des DT est mesurée par le nombre de passages sur l'image et de la taille du voisinage scruté pour le calcul de chaque point. Dans une DT parallèle, l'ordre des calculs pendant une itération donnée est arbitraire (les points de l'objet sont initialisés à $+1$). On peut considérablement accélérer la convergence en choisissant judicieusement l'ordre des calculs [Montanari 68]. Une première technique est de procéder par courbes de niveaux, mais le nombre d'itérations dépend alors de l'épaisseur des objets dans l'image (complexité en $O(L^{n+1})$). Une seconde technique, encore plus rapide, est de procéder par balayages sur l'image (complexité optimale en $O(L^n)$) ; c'est le cas idéal pour un

ordinateur séquentiel, avec un nombre de balayages fixe, indépendant de l'épaisseur des objets dans l'image. Mais de tels schémas ne sont pas possibles pour toutes les distances. Nous passons en revue les DT existantes pour les distances classiques, puis nous abordons les DT de chanfrein au §4.2.3. Le but commun à toutes ces DT est d'approximer le plus rapidement possible la distance euclidienne d_E dans une image. Les distances que nous présentons ont des DT plus ou moins simples ou efficaces, approchent d_E à des degrés divers et possèdent toutes des propriétés particulières. On a donc tout intérêt à conserver cette large palette de distances et de DT pour bien adapter nos choix aux applications recherchées.

4.2.2 Transformation exacte de distance euclidienne

La distance euclidienne d_E n'est évidemment pas discrète. Si l'on veut rester dans le domaine discret, on peut considérer par exemple les fonctions d_E^2 , $round(d_E)$, où l'entier inférieur le plus proche noté $\lfloor d_E \rfloor$ et où l'entier supérieur le plus proche noté $\lceil d_E \rceil$. Si p et q sont dans $\mathbb{Z}^n$, le carré d_E^2 présente l'intérêt d'être à valeurs dans $\mathbb{Z}$. Malheureusement, d_E^2 n'est pas une distance car si on considère dans le plan les points $O(0,0)$, $A(1,0)$ et $B(2,0)$, on a $d_E^2(O,A) = d_E^2(A,B) = 1$ et $d_E^2(O,B) = 4$, mais $d_E^2(O,B) \not\leqslant d_E^2(O,A) + d_E^2(A,B)$. Pour les mêmes raisons, les fonctions $round(d_E)$ et $\lfloor d_E \rfloor$ ne sont pas des distances [Rosenfeld 68]. La cause est le non-respect de l'inégalité triangulaire, en particulier pour les petites valeurs. En revanche, il est intéressant de noter que $\lceil d_E \rceil$ est bien une distance [Rhodes 92] ; toutefois $\lceil d_E \rceil$ n'est pas une norme car $\lceil d_E \rceil((0,0),(1,1)) = \lceil \sqrt{2} \rceil = 2$ et $\lceil d_E \rceil((0,0),(2,2)) = \lceil 2\sqrt{2} \rceil = 3$. Pour mémoriser une carte de distances ou DM (*Distance Map*) calculée avec d_E, on est contraint de changer de système de représentation et cela, soit en stockant les valeurs de d_E dans une image de réels, soit en utilisant deux images d'entiers pour les coordonnées du point du fond le plus proche (en valeurs signées ou absolues), soit encore en mémorisant d dans une image d'entiers longs.

La transformation de distance euclidienne (EDT) est née dans les années 1980. Imparfaite, elle a suscité quantité de travaux pour l'améliorer ou pour élaborer d'autres stratégies. L'opération est rendue difficile par le fait que d_E n'est pas calculable localement sur une carte de distance notée DM (Distance Map). Nous trouvons une analyse bibliographique très complète des différents algorithmes successivement proposés et de leur complexité dans [Cuisenaire 99a].

La première méthode est l'algorithme SED (Sequential Euclidean Distance) de Danielsson qui opère en 4 passages séquentiels sur l'image mais produit des erreurs dans certaines configurations et n'est pas très rapide [Danielsson 80]. Une version modifiée par Ye [Ye 88] permet le calcul de l'image signée. Ragnemalm [Ragnemalm 93] propose une version de SED avec 3 passes en 2D et 4 passes en 3D. Leymarie [Leymarie 92] montre que SED devient, avec quelques changements mineurs, aussi efficace que la DT des distances de chanfrein pour un masque 3x3. Mullikin [Mullikin 92] propose un post-traitement de SED pour corriger les erreurs, de même que Cuisenaire [Cuisenaire 99b]. De son côté, Yamada [Yamada 84] a donné un algorithme parallèle qui dépend donc de l'épaisseur des objets dans l'image et est assez lent sur machine séquentielle. Le résultat est quasi-exact avec de rares erreurs selon [Forchhammer 89]. Des algorithmes parallèles utilisant des opérateurs issus de la morphologie mathématique en niveaux de gris sont proposés pour d_E [Shih 92] et d [Huang 94]. Shih et Mitchell [Shih 92] définissent une transformée de distance comme une érosion par un élément structurant dont la largeur correspond à la distance maximale dans l'image. Huang et Mitchell [Huang 94] adaptèrent cette méthode pour calculer le carré de la distance euclidienne. On trouve d'autres algorithmes parallèles dans [Eggers 97] et [Lee 97]. A titre de curiosité, Forchhammer [Forchhammer 89] présente un calcul en 2D de l'image de distance euclidienne à partir d'une image de distance de chanfrein avec une table de correspondance qui ne fonctionne que pour des petites valeurs (inférieures à $\sqrt{17}$ en utilisant la distance de chanfrein $d_{(3,4)}$ et à $\sqrt{104}$ pour une distance de chanfrein $d_{(19,27,42)}$). Une autre méthode efficace consiste à faire évoluer une chaîne du contour vers l'intérieur des objets [Vincent 91] [Ragnemalm 92] et [Eggers 98] ; le résultat est exact, mais requiert une importante structure de données.

Une nouvelle classe d'EDT utilisant le diagramme de Voronoï est proposée dans [Breu 95], [Cuisenaire 99b] et [Coeurjolly 02], mais encore une fois avec des structures de données conséquentes.

Ils considèrent un ensemble de sites R correspondant aux points discrets du complémentaire de l'objet considéré. Le diagramme de Voronoï discret de R donne, en tout point de l'objet, le point de R qui lui est le plus proche. Ainsi, pour obtenir la transformée de distance, il suffit ensuite d'étiqueter les points discrets de l'objet par leur distance au point de R donnée par le diagramme. Résoudre le problème du calcul du diagramme de Voronoï permet ainsi de résoudre le problème du calcul de la transformée en distance euclidienne. Saito et Toriwaki, proposent dans [Saito 94], un algorithme efficace qui calcule des distances exactes en considérant le carré de la distance euclidienne en toutes dimensions par l'exploitation du théorème de Pythagore. Leur DT (que nous notons PDT, pour Pythagore) est simple à implémenter car elle ne nécessite qu'un aller-retour sur chaque ligne puis sur chaque colonne. En 3D, il suffit de rajouter un aller-retour sur chaque rangée en z (et de même pour chaque dimension supérieure). Le temps de calcul varie selon l'épaisseur des objets, avec une complexité en $O(n.L^{n+1})$ pour une image de côté L en dimension n. PDT est une version spécialisée de l'algorithme multi-distance de Paglieroni dans [Paglieroni 92].

Un comparatif de ces nombreuses EDT dans [Cuisenaire 99a] permet de voir que l'algorithme PDT de Saito et Toriwaki [Saito 94] est un excellent choix si l'on veut un résultat exact. En 2D, quelques algorithmes plus récents ont une meilleure complexité théorique mais ils sont délicats à implémenter et sont pénalisés par la gestion de structures de données lourdes, ce qui les ramène expérimentalement au niveau de PDT. En 3D, PDT est imbattable pour les objets de faible épaisseur. Une solution hybride est proposée en 3D dans [Cuisenaire 99a], avec l'emploi de son algorithme PSN en 2D [Cuisenaire 99b] sur chaque coupe du volume, puis l'aller-retour de PDT sur chaque rangée en z. Cette méthode hybride est plus rentable que PDT à partir d'une épaisseur de 250 voxels.

Des travaux récents ont permis de trouver des algorithmes performants en temps d'exécution pour calculer l'EDT sur des images binaires à n-dimensions [Hirata 96] [Meijster 00] et [Maurer 03]. Notons que l'algorithme de Maurer permet de calculer des images de distance exacte, pour des voxels isotropes et anisotropes, avec un temps d'exécution qui est proportionnel au nombre de voxels de l'image. Cet algorithme nécessite de sauvegarder en mémoire la taille de la transformée de distance et un tableau dont la taille correspond à la plus grande dimension de l'image. L'algorithme de Coeurjolly [Coeurjolly 07] permet de calculer la transformée de distance euclidienne inverse, quelque soit la dimension de l'espace, en ayant un temps d'exécution optimal. C'est une optimisation de l'algorithme proposé par Saito et Toriwaki [Saito 94]. Le principe du calcul repose sur la recherche de l'enveloppe inférieure d'un ensemble de paraboles.

En conclusion, l'emploi de d_E ou de d peut se justifier en analyse d'images lorsque des propriétés d'isotropie et d'exactitude de mesure sont prépondérantes. Ceci explique le volume de travaux autour de la distance euclidienne.

4.2.3 Transformation approchée de la distance euclidienne

Dans certaines applications comme l'analyse des images médicales, images satellitaires, ou images issues de tomographie, les images comportent un volume considérable de données (plusieurs giga octets). L'exploitation de ces images a conduit certains auteurs à s'intéresser aux transformations approchées de distances pour deux raisons :

- ils ont besoin d'algorithmes extrêmement simples et rapides, consommant peu de mémoire et pouvant être facilement adaptés à un calcul en sous-images ;
- ils souhaitent obtenir une carte de distance contenant des entiers pour des problèmes de stockage et pour faciliter les calculs annexes comme le squelette, l'axe médian, ...

Les algorithmes de calcul de la distance euclidienne sont très efficaces pour des images de taille normale, mais difficilement adaptables aux cas des images volumineuses issues par exemple des milieux médical ou satellitaire. En effet, les algorithmes parallèles sont très simples et très rapides, mais multiplient au moins par deux la taille des images stockées en mémoire. On stocke le résultat de chaque passe dans une image différente de l'image de départ. Quant aux algorithmes séquentiels,

les plus rapides ne sont pas toujours facilement adaptables à une réalisation en sous-images. Les distances de chanfrein, proposées initialement par Montanari [Montanari 68], et popularisées par Borgefors [Borgefors 84], réalisent ces attentes. En effet, elles calculent une estimation entière proportionnelle à la distance euclidienne en propageant des estimations de distances locales et peuvent être très efficacement implémentées grâce à un algorithme qui effectue deux passages sur une image 2D [Rosenfeld 68]. Cette transformation calcule une valeur approchée de la distance euclidienne. Le point délicat de cette méthode est le choix judicieux du masque de chanfrein qui minimisera l'erreur entre les valeurs obtenues et la distance euclidienne.

Afin d'améliorer les **distances de chanfrein**, des efforts ont été menés dans cinq directions :

- ***diminuer la sensibilité à la rotation*** au moyen d'une meilleure approximation de la distance euclidienne. Cela a été réalisé en affectant des poids aux déplacements élémentaires de l'opérateur local de distance. Ces poids sont optimisés par rapport à un critère d'erreur, qui consiste généralement à minimiser la différence maximale entre la distance calculée et la distance euclidienne le long d'une trajectoire rectiligne [Borgefors 86] ou circulaire [Verwer 91] et [Coquin 95a].
- ***augmenter la dimension de l'espace image*** : les transformées de distance pondérées en 3D (ou distance de chanfrein) ont été introduites par Borgefors en 1984 [Borgefors 84]. Différentes approches sont possibles pour calculer les coefficients locaux, selon les modèles discrets ou continus. Les trajectoires de référence sont rectilignes ou sphériques. Les coefficients locaux d'un masque $5 \times 5 \times 5$ ont été proposés dans [Verwer 91] [Remy 00] et [Svensson 02]. Des transformations de distance dans des espaces de dimension 4, pour la squelettisation [Jonker 96] ou de dimension 5 pour la comparaison d'images couleurs [Coquin 00b], ont également été proposées. Cependant, certaines difficultés apparaissent concernant la taille mémoire utilisable pour stocker les calculs intermédiaires et mémoriser les résultats, et relativement au temps de calcul nécessaire pour obtenir le résultat.
- ***étudier des propriétés génériques*** : les distances calculées au moyen des transformées de distance doivent être des métriques, cela nécessite des conditions pour l'optimisation de l'opérateur. Ainsi, des conditions sur la géométrie des boules ont permis de traduire des contraintes exactes sur les pondérations pour qu'un masque de chanfrein induise bien une distance. La minimisation est réalisée directement dans l'espace discret par exploitation des suites de Farey [Thiel 92a] et [Thiel 92b]. De même, des conditions de semi-régularité ont été développées dans [Kiselman 96] assurant que tout chemin discret composé de certains déplacements multiples est optimal. Kiselman montre qu'une transformée de distance dans $\mathbb{Z}^n$ vérifiant les conditions d'une norme est semi-régulière. Des conditions pour obtenir une norme sont également données pour un opérateur $5 \times 5 \times 5$ [Remy 00].
- ***adapter l'opérateur local au cas des grilles non cubiques*** : les systèmes d'imagerie produisent des images qui n'ont pas nécessairement le même pas d'échantillonnage dans toutes les directions. Dans la plupart des cas, les images sont composées de voxels ayant deux côtés égaux et un troisième différent. C'est le cas par exemple en tomographie, en microscopie confocal, où le rapport entre la plus petite et la plus grande direction varie typiquement entre 1 et 10. Dans [Bolon 92], [Sintorn 01], des transformations de distance en maillage rectangulaire ont été proposées. L'étape d'optimisation des coefficients est basée la minimisation de l'erreur maximale le long d'une trajectoire de référence rectiligne. Comme la distance entre les pixels dans la direction diagonale est plus grande que la distance sur les directions verticale et horizontale, une certaine erreur relative dans la direction diagonale domine le processus d'optimisation des coefficients comparativement aux autres directions. Dans [Coquin 95a], une trajectoire de référence circulaire a été proposée et des conditions de métricité ont été données. Un autre moyen de diminuer l'erreur maximale entre la distance euclidienne et la distance locale est d'augmenter la taille du masque de l'opérateur. Une extension en 3D associée aux maillages parallélépipédiques a été développée dans [Chehadeh 95], [Sintorn 02], [Sintorn 04], puis complétée dans [Fouard 05] en construisant directement une triangulation régulière qui utilise la triangulation de Farey. Dans [Fouard 05], les coefficients entiers sont calculés en minimisant l'erreur relative entre la distance de chanfrein et la distance euclidienne, dans chaque cône régulier. Dans [Strand 05] également, nous trouvons des transformations de distances

adaptées aux grilles 3D et utilisant des voxels non-cubiques. Également, un algorithme efficace de file d'attente de pixels prioritaires pour calculer une transformée de distance dans un espace courbe, est présenté dans [Ikonen 05, Ikonen 07]. Les transformées de distance fournissent des outils pour trouver les chemins les plus courts sur des surfaces en niveaux de gris. De cette manière, une variation de surface, ou la rugosité, peut être mesurée.

– **adapter l'opérateur local aux maillages non-stationnaires** : c'est le cas des systèmes d'acquisition qui ont, par exemple, un pas d'échantillonnage non uniforme dans une direction privilégiée. Ce type d'opérateur peut être vu comme une version non-stationnaire d'un opérateur local qui s'adaptant à la grille non-cubique en 2D [Dubuisson 93] et en 3D [Coquin 00a].

Notre contribution : nous allons, dans les sections suivantes, rappeler brièvement les travaux que nous avons développés, dans la thèse de Yousra CHEHADEH soutenue en octobre 1997 [Chehadeh 97], sur les opérateurs de distances en 2D et 3D. Nous montrerons la façon d'adapter l'opérateur local au cas des grilles non-cubiques, et les extensions que nous avons faites jusqu'en 2000, avec le développement d'un opérateur local 3D non-stationnaire.

4.3 Opérateurs locaux de distances en 2D

Dans cette section, nous allons décrire brièvement le principe de calcul des opérateurs locaux de distances 2D en maillage rectangulaire. Nous reprenons la démarche adoptée par Borgefors qui consiste à minimiser l'écart maximal entre la distance euclidienne d_E et la distance de chanfrein, appelée distance locale et notée d_L [Borgefors 86].

Borgefors a proposé de minimiser l'écart maximal entre d_E et d_L le long d'une trajectoire rectiligne verticale. L'optimisation se fait en fixant un paramètre M (l'abscisse de la verticale d'équation $x = M$). Lorsque nous considérons le cas général, rien ne nous impose cette condition, car les images ne sont pas forcément de forme carrée. Nous avons décidé de réaliser cette optimisation sur une trajectoire circulaire, pour ne pas privilégier une direction particulière par rapport à une autre [Verwer 91] et [Coquin 93]. De plus, nous allons considérer une trajectoire circulaire de rayon R élevé pour assimiler le déplacement pixel à pixel à un déplacement continu et autoriser ainsi la dérivation. Cette démarche est critiquable dans la mesure où l'optimisation se fait en considérant un déplacement continu du pixel Q, alors que ce déplacement est quantifié par le pas de discrétisation de l'image. Cependant, les écarts calculés étant une fonction continue du pixel Q, l'erreur due à la discrétisation devient négligeable quand la distance entre $O(0,0)$ et $Q(x,y)$ augmente [Bolon 92].

En maillage rectangulaire, le masque des coefficients de l'opérateur local est symétrique par rapport aux diagonales. En considérant le premier quadrant (c'est-à-dire $x \geq 0$ et $y \geq 0$), chaque cône est délimité par deux déplacements élémentaires notées d_{ij}, i et j dépendant de la taille du masque. Le plus court chemin entre deux pixels situés dans ce cône sera exprimé en fonction des deux déplacements élémentaires le délimitant.

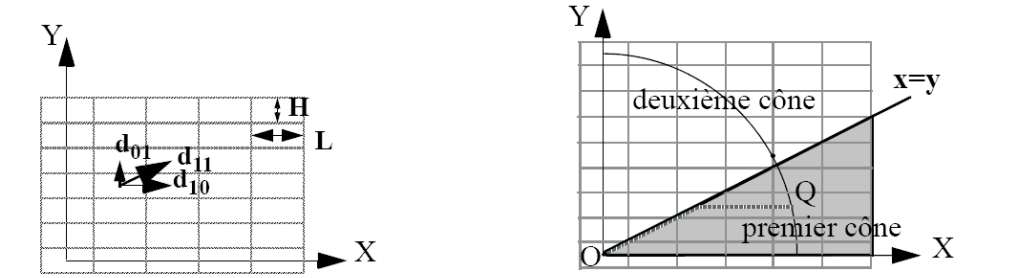
Nous détaillerons d'abord l'optimisation d'un opérateur local de distance, pour bien comprendre la démarche adoptée, puis nous proposerons une généralisation pour des opérateurs cubiques de taille quelconque ($U \times U$, avec $U = 2u + 1$, $u \geq 1$ entier) ainsi qu'une généralisation pour des opérateurs non-cubiques de taille quelconque ($U \times V$, avec $U = 2u + 1$, $V = 2v + 1$, $U \neq V$, u, v entiers).

4.3.1 Optimisation d'un opérateur local de distance

Dans cette partie nous allons détailler la démarche de l'optimisation d'un opérateur local de distance. Cette démarche s'applique à un espace de n'importe quelle dimension. Nous suivons la démarche adoptée par Borgefors [Borgefors 86] pour déterminer les coefficients réels optimaux, mais, contrairement à Borgefors qui a fait son calcul pour le pixel Q décrivant une droite verticale (*ce qui accentue l'erreur par rapport à la distance euclidienne sur la diagonale*), nous considérons, comme Verwer [Verwer 91], que le pixel Q décrit une trajectoire circulaire de rayon $R = \sqrt{(Lx)^2 + (Hy)^2}$ de manière à répartir l'erreur par rapport à la distance euclidienne sur l'ensemble des directions

possibles.

En maillage rectangulaire, l'opérateur de taille 3×3 est caractérisé par trois coefficients déterminant les déplacements élémentaires notées d_{10}, d_{11}, et d_{01}, comme le montre la figure 4.1a. Les trois déplacements élémentaires divisent le quart du cercle en deux cônes. Le premier est délimité par les déplacements élémentaires d_{10} et d_{11}. Le deuxième est celui situé entre les déplacements élémentaires d_{11}, d_{01} (figure 4.1b). Nous pouvons remarquer que les déplacements horizontaux et verticaux sont différents. D'après Montanari [Montanari 68], il existe toujours un chemin minimal formé, au plus, de deux segments de droite discrets entre les pixels $O(0,0)$ et $Q(x,y)$, quelle que soit la position de Q.



(a) Déplacements élémentaires (b) Les cônes du premier quadrant du cercle

FIG. 4.1: Opérateur local de distance de taille 3×3.

La démarche repose sur les étapes suivantes :

- **1)** Choisir la taille $U \times V$ du masque, avec $U = 2u + 1$ et $V = 2v + 1$ ce qui va permettre de définir les déplacements élémentaires d_{ij} délimitant les différents cônes (voir Fig. 4.1), avec i, j satisfaisant aux conditions suivantes :

$$Condition1 : i, j \in \{0, 1, ..., u\} \times \{0, 1, ..., v\} \qquad Condition2 : PGCD(i, j) = 1 \tag{4.2}$$

- **2)** Chaque déplacement élémentaire d_{ij} est affecté d'un poids qui doit être proportionnel au déplacement voulu.
- **3)** Exprimer dans chaque cône la distance locale $d_L(O, Q)$ entre l'origine $O(0,0)$ et un point $Q(x,y)$ décrivant une trajectoire circulaire en 2D, ou sphérique en 3D, de rayon R. Cette distance est une fonction des coordonnées du point Q et des d_{ij}.
- **4)** Calculer dans chaque cône l'erreur $E = d_L(O, Q) - d_E(O, Q)$ entre la distance locale d_L et la distance euclidienne d_E.
- **5)** Le but est de calculer les valeurs optimales des coefficients d_{ij} qui minimisent l'erreur maximale produite par E. L'erreur est extrêmale aux bornes des cônes (on note E_1, E_2, ..., E_n, ... ces erreurs) ou lorsque la dérivée première de E par rapport à x ou y s'annule (on note E_{max} cette erreur). On calcule l'expression de ces erreurs pour chaque cône.
- **6)** On cherche les solutions de l'équation $E_{max} = -E_1 = -E_2 = ... = -E_n$. C'est une manière de minimiser cette erreur maximale [Rosenfeld 66] et [Borgefors 86].
- **7)** La résolution des équations précédentes permet d'obtenir l'expression des coefficients d_{ij}.
- **8)** Enfin, on calcule la valeur de l'erreur maximale, notée E_{max}.
- **9)** Le masque est composé des déplacements élémentaires d_{ij} et est utilisé dans le calcul de la transformation de distances.

Pour plus d'information, se rapporter à l'Annexe A, qui détaille l'ensemble de ces calculs pour un opérateur local de distance de taille 3×3, et pour un opérateur de taille 5×5 se reporter à [Coquin 95a].

La figure 4.2 présente une image de distance au fond, calculée sur une image binaire (les points du fond étant étiquetés à 0) en utilisant un opérateur 5×5, par un algorithme séquentiel en deux passages [Rosenfeld 66] et [Borgefors 84].

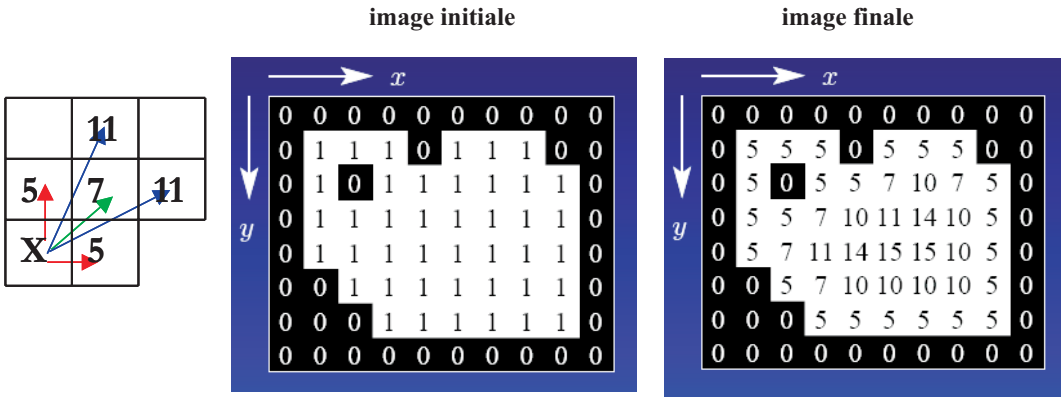


FIG. 4.2: Image de distance au fond par un opérateur de taille 5×5.

4.3.2 Généralisation : Opérateur cubique de type $U \times U$

En utilisant la démarche précédente, il est possible de trouver l'expression des coefficients pour un masque de taille quelconque. En maillage rectangulaire, un opérateur de taille $U \times U$ ($U = 2u+1$, $u \geq 1$) est défini par les coefficients d_{ij}. Les indices i et j sont les coordonnées des pixels voisins sur les axes OX et OY respectivement. Ils vérifient les deux conditions suivantes :

$$Condition1 : i, j \in \{0, 1, ..., u\} \qquad Condition2 : PGCD(i, j) = 1 \tag{4.3}$$

Les coefficients d_{ij} définissent les directions principales correspondant aux déplacements élémentaires (*dans les directions (i,j)*), sachant qu'une direction ne doit coïncider avec aucune autre. On note également que les indices i et j vérifient la condition nécessaire et suffisante des suites de Farey. Cet opérateur divise le plan de l'image en plusieurs cônes, chaque cône étant délimité par deux directions principales. On fera l'étude dans le premier quadrant. Pour les autres quadrants, les formules sont similaires par symétrie par rapport à OX et OY. Dans la figure 4.3 nous montrons uniquement le premier et le dernier cônes du premier quadrant, les autres cônes se situent entre ceux-ci.

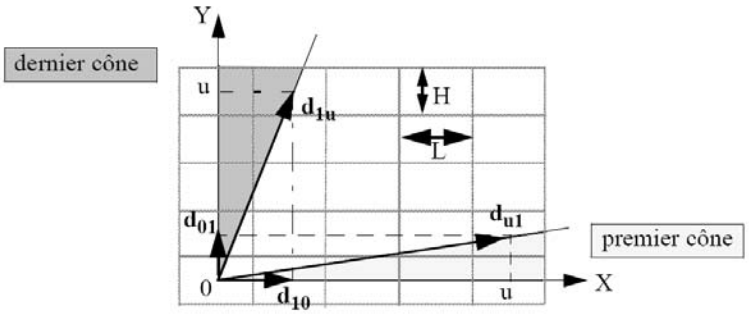


FIG. 4.3: Opérateur cubique $U \times U$ en maillage rectangulaire.

La distance $d_L(O, Q)$, entre l'origine $O(0, 0)$ et un pixel $Q(x, y)$ d'un cône, est constituée des déplacements suivant les deux directions délimitant ce cône. Verwer a montré que l'erreur maximale entre $d_E = R$ et d_L se produit dans le cône ayant le plus grand angle [Verwer 91]. Dans le cas du maillage rectangulaire avec $L \leq H$, le cône ayant le plus grand angle est le dernier, c'est-à-dire celui délimité par les directions d_{1u} et d_{01}. En utilisant la démarche de la section précédente nous obtenons les coefficients réels qui sont donnés par :

$$d_{01} = \frac{-2H + 2H\sqrt{1+\lambda}}{\lambda} \tag{4.4}$$

avec

$$\lambda = \frac{1}{L^2}\left(\sqrt{L^2 + u^2H^2} - uH\right)^2 \tag{4.5}$$

les autres coefficients sont donnés par

$$d_{ij} = T_{ij}\frac{d_{01}}{H} \tag{4.6}$$

avec

$$T_{ij} = \sqrt{(iL)^2 + (jH)^2} \tag{4.7}$$

L'erreur normalisée maximale a pour expression :

$$e_{max} = \frac{E_{max}}{R} = \left|1 - \frac{d_{01}}{H}\right| = \left|1 - \frac{-2 + 2\sqrt{1+\lambda}}{\lambda}\right| \tag{4.8}$$

Cette erreur ne dépend que du rapport H/L et de u (on rappelle que $u = (U-1)/2$) donc de la taille de l'opérateur U, puisque l'expression 4.5 est équivalente à

$$\lambda = \left(\sqrt{1 + u^2\frac{H^2}{L^2}} - u\frac{H}{L}\right)^2 \tag{4.9}$$

Le tableau 4.1 présente les valeurs de l'erreur normalisée maximale produite en utilisant des opérateurs cubiques de différentes tailles en 2D, lorsque L=H=1. Il est clair que e_{max} diminue lorsque la taille du masque augmente.

taille du masque	3	5	7	9	11
e_{max} (%)	3.9566	1.3557	0.6498	0.3760	0.2439

TAB. 4.1: Variations de l'erreur normalisée maximale avec la taille de l'opérateur, pour $L = H = 1$.

Le tableau 4.2 présente les valeurs de l'erreur normalisée maximale produite par l'opérateur 5×5 en fonction de différentes valeurs de la largeur du pixel L et de la hauteur H.

(L, H)	$(1, 1)$	$(1.41, 1)$	$(2, 1)$	$(10, 1)$
e_{max} (%)	1.3557	2.3679	3.9566	12.7814

TAB. 4.2: Variations de l'erreur normalisée maximale de l'opérateur 5×5 en fonction de L et H.

On remarque une augmentation de l'erreur lorsque la rapport L/H augmente.

L'étude de l'optimisation a montré que l'erreur produite avec des opérateurs cubiques n'est pas répartie uniformément dans les deux octants à cause de la forme rectangulaire du pixel. Ceci va entraîner une dissymétrie des boules de distance. L'erreur normalisée est maximale dans le dernier octant. D'où l'idée d'utiliser d'autres types d'opérateurs que l'on peut appeler ***opérateurs non-cubiques***. Il s'agit d'opérateurs dont la taille est différente en x et en y, donc de taille $U \times V$ avec $U \neq V$, comme par exemple les opérateurs 3×5 et 5×3. L'idée sous-jacente est de voir s'il est possible de limiter la complexité de calcul tout en maintenant de bonnes performances. Nous avons également montré dans le tableau 4.2 que l'erreur normalisée augmente en fonction de la largeur L du pixel. D'autre part, nous avons également montré que l'erreur normalisée diminue avec la taille du masque (tableau 4.1). C'est pour ces deux raisons que nous prendrons un masque 3×5 plutôt que 5×3. Il est préférable de diminuer l'erreur dans le dernier octant, là où elle est la plus importante. Nous avons donc proposé une généralisation des opérateurs non cubiques en 2D.

4.3.3 Généralisation : Opérateur non-cubique de type $U \times V$

Soient $U = 2u + 1$ et $V = 2v + 1$, les dimensions du masque. On suppose que $L \geq H \geq 1$. Dans ce cas, l'erreur absolue maximale se produit dans le dernier cône du premier quadrant (délimité par les directions de d_{01} et d_{1v}), voir figure 4.4, et augmente avec L (voir [Verwer 91], [Coquin 93] et [Coquin 94]). Donc l'erreur dans ce cône doit être réduite. D'autre part, l'erreur absolue maximale diminue avec la taille du masque. Ce qui nous conduit à choisir $U \leq V$.

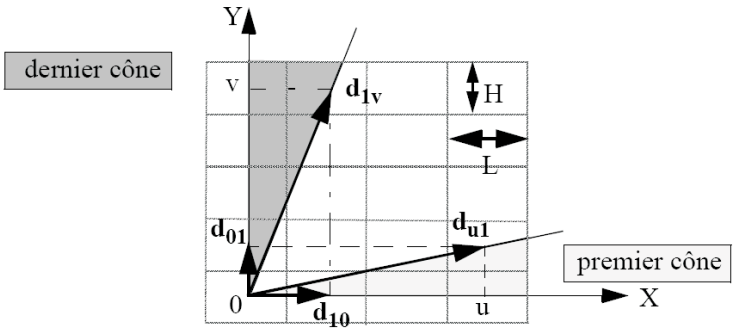


FIG. 4.4: Opérateur non-cubique $U \times V$ en maillage rectangulaire.

Optimisation des coefficients

Avec les masques non cubiques, deux procédures d'optimisation sont possibles (voir Fig. 4.4) :
- **(1)** Minimiser l'erreur dans le premier cône (délimité par les directions de d_{10} et d_{u1}, ce qui conduit à obtenir l'erreur e_1, qui est une fonction de u et du rapport L/H
- **(2)** Minimiser l'erreur dans le dernier cône (délimité par les directions de d_{1v} et d_{01}, ce qui conduit à obtenir l'erreur e_2, qui est donc une fonction de v et du rapport H/L

L'étude de e_1 et de e_2 en tant que fonction de L/H avec $L \geq H \geq 1$ montre que e_2 est une fonction strictement croissante de L/H, tandis que e_1 est une fonction strictement décroissante. Le point d'intersection de ces deux courbes (e_1=e_2) se produit pour :

$$R_{uv} = \frac{L}{H} = \sqrt{\frac{v}{u}} \tag{4.10}$$

Ainsi, connaissant la largeur du maillage L et sa hauteur H (ou bien le rapport L/H), pour avoir les meilleures performances avec un minimum de temps de traitement, on doit choisir un opérateur de taille $U \times V$ telle que la valeur de R_{uv} calculée par l'expression 4.10 soit égale ou soit la plus proche possible du rapport L/H. La figure 4.5 décrit l'algorithmique permettant l'optimisation des coefficients d'un opérateur non-cubique $U \times V$ en maillage rectangulaire.

Taille du masque $U \times V$	3×3	5×5	7×7	9×9	11×11	3×5	3×7	3×9	3×11	5×11
e_{max} (%)	8.070	3.957	2.192	1.356	0.910	3.957	2.192	1.356	1.356	0.910

TAB. 4.3: Erreur normalisée maximale produite avec des opérateurs de distance 2D cubiques et non-cubiques ($L = 2$ et $H = 1$).

Le tableau 4.3 présente les valeurs de l'erreur normalisée maximale produite par des opérateurs 2D de différentes tailles, pour un maillage tel que $L = 2$ et $H = 1$. Pour faciliter la comparaison, nous présentons des opérateurs cubiques et non-cubiques qui ont les mêmes performances que des opérateurs cubiques de tailles supérieures.

A partir de ce tableau, nous constatons qu'un opérateur non-cubique peut remplacer un opérateur cubique de taille plus grande tout en préservant les mêmes performances (même valeur de l'erreur maximale). Les opérateurs non-cubiques sont donc très appropriés aux maillages rectangulaires. Nous avons étudié l'anisotropie et la métricité de ces opérateurs dans [Coquin 95a] et [Chehadeh 95].

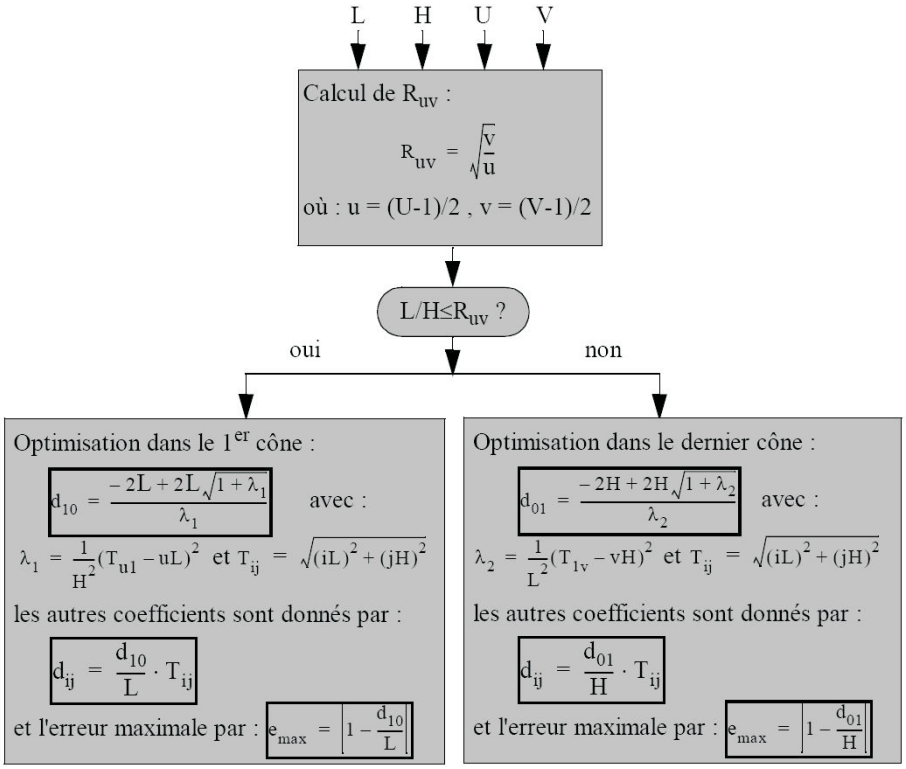


FIG. 4.5: Schéma de l'algorithme d'optimisation d'un opérateur non-cubique $U \times V$ en maillage rectangulaire.

4.3.4 Approximation entière

Pour des raisons de temps de calcul et d'espace mémoire, il est préférable d'utiliser des coefficients entiers. Ces coefficients sont obtenus en multipliant les coefficients réels par un facteur d'échelle [Borgefors 86], [Verwer 91], [Coquin 93], [Coquin 95a]. Nous utilisons cette approximation mais contrairement à Verwer qui utilise un facteur d'échelle réel nous considérons comme Borgefors, un facteur d'échelle entier, noté N, avec $N \geq 0$. Les coefficients résultants sont ensuite arrondis à l'entier le plus proche. Soit D_{ij} le coefficient entier correspondant au coefficient réel d_{ij}.

$$Dij = arrondi(N.d_{ij}) \tag{4.11}$$

Comme l'erreur maximale est une fonction continue des coefficients d_{ij}, l'erreur effective (avec les coefficients entiers) tend vers l'erreur théorique (avec les coefficients réels) lorsque N tend vers l'infini. Or le facteur N doit être limité pour des raisons pratiques. La valeur maximale N_{max} de N peut toutefois être donnée en fonction des dimensions de l'image $Dim \times Dim$, du nombre b de bits utilisés pour coder un pixel (dans le cas 2D) et de la valeur du coefficient réel d_{11}. Pour cela nous procédons comme suit :

soit Δ_{max} la distance maximale entre deux pixels de l'image, on peut écrire :

$$\Delta_{max} = Dim.D_{11} \tag{4.12}$$

On doit avoir :

$$Dim.D_{11} = Dim.round(d_{11}.N) \approx Dim.d_{11}.N \leq 2^b \tag{4.13}$$

donc le facteur d'échelle N doit vérifier la condition suivante :

$$N \leq N_{max} \approx \frac{2^b}{Dim.D_{11}} \tag{4.14}$$

Exemple : Soit $Dim = 512$, $L = H = 1$ (maillage carré), opérateur cubique 5×5 et $b = 16bits$: L'équation 4.14 donne $N_{max} = 90$. La meilleure approximation entière est obtenue pour $N = 73$. Les coefficients entiers ont alors les valeurs suivantes :

$$D_{10} = D_{01} = 72, \quad D_{21} = D_{12} = 161 \quad et \quad D_{11} = 102 \tag{4.15}$$

Avec ces coefficients et le facteur N, l'erreur maximale normalisée entre la distance locale et la distance euclidienne est de 0.013699. Il est clair que l'erreur produite par cet opérateur entier est très proche de celle obtenue par l'opérateur réel optimal qui est de 0.013557.

4.4 Opérateurs de distances discrètes en 3D

Pour les images 3D telles que les images confocales, les images sismiques ou les images médicales formées de coupes scanner successives, le pas d'échantillonnage varie selon la direction considérée. C'est également le cas des images à niveaux de gris pour lesquelles les distorsions géométriques (dans le domaine spatial) et les distorsions radiométriques (échelle des niveaux de gris) peuvent avoir des poids différents. Le maillage du volume est donc parallélépipédique, et un cas particulier est le maillage cubique. Comme nous l'avons déjà mentionné, le maillage parallélépipédique est défini sur une grille 3D dont l'élément est un parallélépipède donné par ses dimensions notées L (Largeur), H (Hauteur) et P (Profondeur). L, H et P sont des réels positifs non nuls et en général on a $L \neq H \neq P$. On considère dans la suite que $P \geq L \geq H$. Les autres cas peuvent se déduire à une permutation d'indices près.

Comme dans le cas des maillages rectangulaires (2D), en maillages parallélépipédiques il y a seulement une symétrie par rapport aux axes OX, OY et OZ. On peut donc diviser la sphère en 8 parties égales. L'étude de l'optimisation des opérateurs locaux de distances 3D peut être effectuée dans l'une de ces 8 parties. Nous choisissons de faire les calculs dans la partie pour laquelle x, y et z sont positifs. L'objectif est d'approximer la distance euclidienne d_E.

Le principe de la distance locale en 3D est le même qu'en 2D. La distance $d_L(O, Q)$ entre deux voxels $O(0, 0, 0)$ pris comme origine et un voxel $Q(x, y, z)$ est la longueur du plus court chemin entre O et Q. En 3D, le plus court chemin est constitué de 3 segments de droites au plus, suivant les trois directions délimitant la portion à l'intérieur de laquelle le voxel Q appartient. Cette distance $d_L(O, Q)$ est calculée par propagation d'opérateurs locaux de distance dont les coefficients sont choisis et calculés selon certains critères.

4.5 Généralisation des opérateurs locaux de distances discrètes

Comme dans le cas 2D, la technique d'optimisation peut être généralisée pour des opérateurs cubiques et non-cubiques de taille quelconque. Le critère d'optimisation est la minimisation de l'écart maximum entre la distance locale et la distance euclidienne $d_E(O, Q) = \sqrt{(Lx)^2 + (Hy)^2 + (Pz)^2} = R$, lorsque Q décrit une trajectoire sphérique, de rayon suffisamment grand par rapport aux dimensions L, H et P du voxel, pour assimiler le déplacement voxel à voxel à un déplacement continu, autorisant ainsi la dérivation. Dans la suite, nous considérons le cas de voxels ayant deux côtés égaux ($L = H$) et le troisième différent (P), avec $P \geq L$.

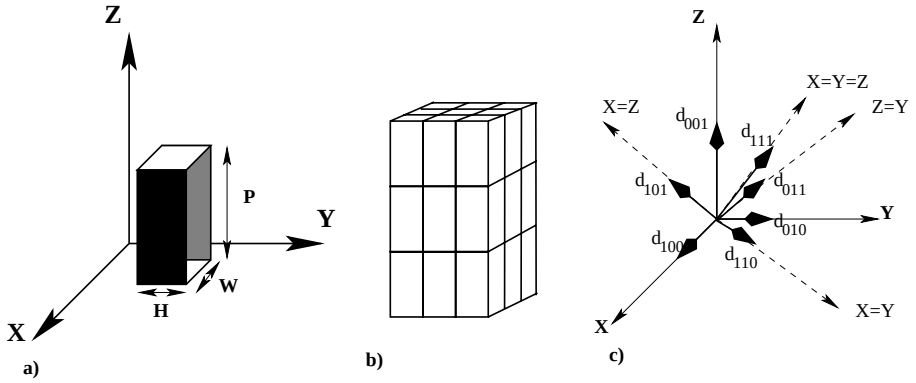


FIG. 4.6: a) caractéristiques du voxel, b) opérateur 3x3x3, c) déplacements élémentaires d_{ijk}

4.5.1 Optimisation des opérateurs cubiques $U \times U \times U$

Soient d_{ijk} les coefficients d'un opérateur de taille $U \times U \times U$ avec $(U = 2u + 1)$. Les indices i, j et k correspondent aux déplacements horizontaux en x, y et au déplacement vertical z. Ils vérifient :

$$Condition1 : i, j, k \in \{0, 1, ..., u\} \qquad Condition2 : PGCD(i, j, k) = 1 \tag{4.16}$$

Pour des raisons de symétrie, nous considérons seulement le cas $x \geq 0$, $y \geq 0$, et $z \geq 0$. Les déplacements élémentaires d_{ijk} définissent des portions de sphère dans le référentiel image. La figure 4.7 montre les 6 portions du premier octant de sphère

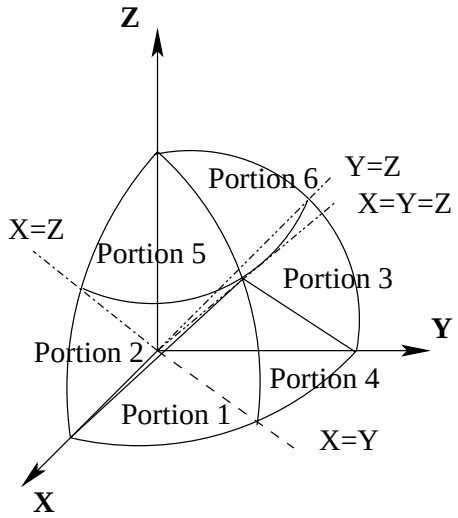


FIG. 4.7: Portion de sphère pour un opérateur local 3x3x3

Certaines contraintes doivent être appliquées aux coefficients d_{ijk} pour empêcher les chemins entre deux voxels qui violeraient l'inégalité triangulaire. Les coefficients de l'opérateur doivent satisfaire aux conditions de semi-régularité [Sintorn 02].

Pour un opérateur cubique $U \times U \times U$ avec $(U = 2u + 1)$, l'erreur maximale se situe dans la portion de sphère délimitée par les directions élémentaires d_{100}, d_{u10} et d_{u11} (première portion du premier octant de sphère, région de plus grand angle [Verwer 91]).

La distance locale dans cette portion peut s'exprimer par :

$$d_L(O, Q) = d_{100}x + (d_{u10} - ud_{100})y + (d_{u11} - d_{u10})z \tag{4.17}$$

ou bien en fonction de y et z :

$$d_L(y,z) = \frac{d_{100}}{L}\sqrt{R^2-(Ly)^2-(Pz)^2} + (d_{u10} - ud_{100})y + (d_{u11} - d_{u10})z \tag{4.18}$$

D'où l'expression de l'erreur $E = d_L - d_E$, qui peut s'exprimer en fonction de y et de z :

$$E(y,z) = \frac{d_{100}}{L}\sqrt{R^2-(Ly)^2-(Pz)^2} + (d_{u10} - ud_{100})y + (d_{u11} - d_{u10})z - R \tag{4.19}$$

En utilisant l'approche mentionnée dans [Borgefors 86] pour minimiser l'erreur, nous pouvons écrire que E est maximale sur les bords de la portion ou lorsque les dérivées partielles d'ordre 1 s'annulent. On note e_{max} l'erreur maximale normalisée, qui est donnée par : (voir [Chehadeh 95] pour plus de détails) :

$$e_{max} = \frac{E_{max}}{R} = \left|1 - \frac{d_{100}}{L}\right| \tag{4.20}$$

avec

$$d_{100} = \frac{-2L + 2L\sqrt{1+\lambda_u}}{\lambda_u} \tag{4.21}$$

et

$$\lambda_u = \frac{1}{L^2}(T_{u10} - uL)^2 + \frac{1}{P^2}(T_{u11} - T_{u10})^2 \tag{4.22}$$

Les autres coefficients d_{ijk} sont donnés par :

$$d_{ijk} = T_{ijk}\frac{d_{100}}{L} \tag{4.23}$$

avec

$$T_{ijk} = \sqrt{(iL)^2 + (jL)^2 + (kP)^2} \tag{4.24}$$

Le tableau 4.4 présente les valeurs de l'erreur maximale normalisée e_{max} produite en utilisant des opérateurs cubiques de différentes tailles en 3D, lorsque $L = H = P = 1$. Nous pouvons remarquer que l'erreur e_{max} diminue avec la taille du masque [Verwer 91, Borgefors 84].

Taille du masque	3	5	7	9	11
e_{max}	6.019	2.411	1.223	0.725	0.476

TAB. 4.4: Variation de l'erreur maximale normalisée avec la taille de l'opérateur cubique 3D, cas du maillage carré ($L = H = P = 1$).

Influence de la profondeur P du voxel et de la taille du masque U

Nous avons analysé les performances de l'opérateur de distance en fonction de l'erreur maximale normalisée. Nous étudions l'influence de la profondeur P du voxel, dans le cas $L = H = 1$ et $P \geq 1$, pour un opérateur de taille $3 \times 3 \times 3$. Comme $L = H$, les erreurs maximales normalisées de chaque portion du premier octant de la sphère sont égales pour des portions symétriques. De plus comme $L = H = 1$, l'erreur maximale normalisée dans les portions 1 et 4 est donnée par :

$$Err_1 = \frac{E_{max}}{R} = \left|1 - \frac{d_{100}}{L}\right| = 1 - \frac{-2 + 2\sqrt{1+\lambda_1}}{\lambda_1} \tag{4.25}$$

avec

$$\lambda_1 = (\sqrt{2}-1)^2 + \frac{(\sqrt{2+P^2} - \sqrt{2})^2}{P^2} \tag{4.26}$$

L'erreur maximale normalisée dans la portion 5 ($z \geq x \geq y$) et dans la portion 6 ($z \geq y \geq x$) est donnée par :

$$Err_2 = 1 - \frac{-2 + 2\sqrt{1+\lambda_2}}{\lambda_2} \tag{4.27}$$

avec

$$\lambda_2 = \left(\sqrt{1+P^2} - P\right)^2 + \left(\sqrt{2+P^2} - \sqrt{1+P^2}\right)^2 \tag{4.28}$$

L'erreur maximale normalisée dans la portion 2 ($x \geq z \geq y$) et la portion 3 ($y \geq z \geq x$) est donnée par :

$$Err_3 = 1 - \frac{-2 + 2\sqrt{1+\lambda_3}}{\lambda_3} \tag{4.29}$$

avec

$$\lambda_3 = \left(\sqrt{2+P^2} - \sqrt{1+P^2}\right)^2 + \frac{(\sqrt{1+P^2} - 1)^2}{P^2} \tag{4.30}$$

La figure 4.8 montre l'évolution de l'erreur maximale normalisée produite par un opérateur cubique de taille 3x3x3, pour différentes valeurs de $P \geq 1$.

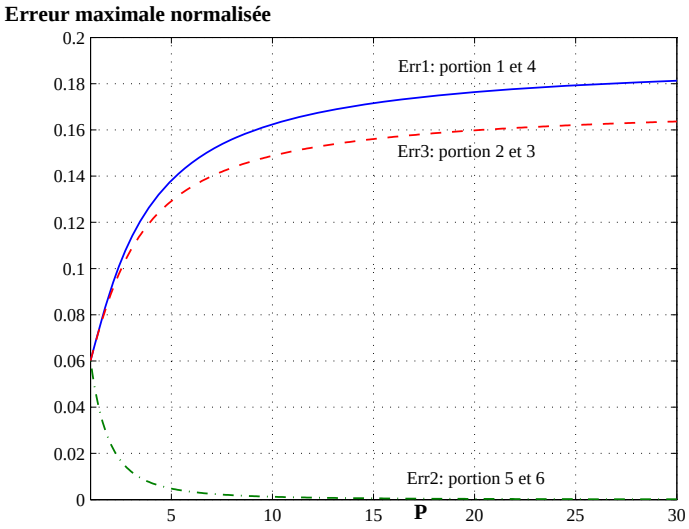


FIG. 4.8: Erreur maximale normalisée en fonction de la profondeur $P \geq 1$

Cette étude montre que plus la profondeur P augmente et plus l'erreur maximale normalisée est élevée dans les portions 1 et 4 du premier octant de la sphère (Fig. 4.8).

La figure 4.9 montre que l'erreur maximale normalisée décroît quand la taille du masque diminue, dans la première et la dernière portion du premier octant de sphère, conformément à ce qui est annoncé dans [Verwer 91].

4.5.2 Optimisation des opérateurs non-cubiques $U \times U \times V$

Comme nous l'avons montré en 2D, il est possible également de réduire le temps de calcul en utilisant des masques non-cubiques (opérateurs anisotropes 3D), et en conservant les mêmes performances en terme de minimisation de l'erreur maximale normalisée. Ces travaux sont détaillés dans [Chehadeh 97] et repris dans [Fouard 05].

Si nous considérons que les dimensions du voxel sont égales dans le plan horizontal ($L = H$), alors il faut considérer les deux cas suivants :

- **premier cas** $P \geq 1$: l'erreur maximale se produit dans la première portion du premier octant de la sphère (délimitée par les déplacements élémentaires d_{100}, d_{u10} et d_{u11}) et augmente si P augmente. L'erreur maximale décroît avec la taille du masque. Nous devons alors choisir un masque $U \times U \times V$ avec $U \geq V$ ($U = 2u+1$, et $V = 2v+1$). La figure 4.10 présente le masque d'un opérateur 5x5x3, et les différentes portions sur lesquelles doit se faire l'optimisation.

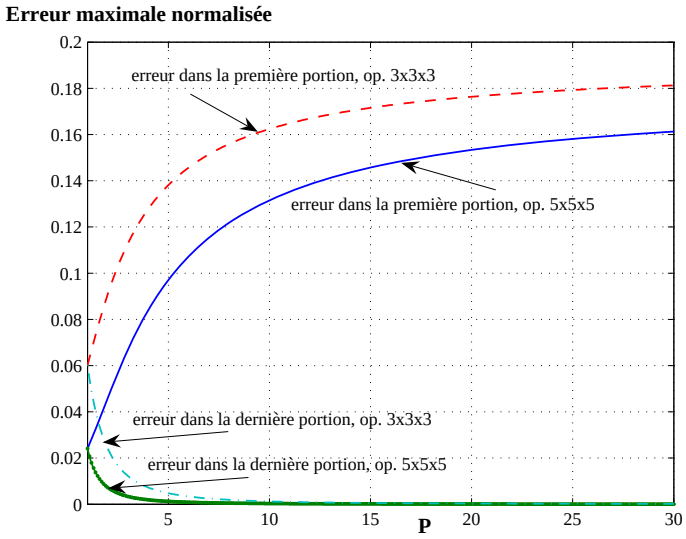


FIG. 4.9: Erreur maximale normalisée en fonction de $P \geq 1$, pour des opérateurs cubiques 3x3x3 et 5x5x5

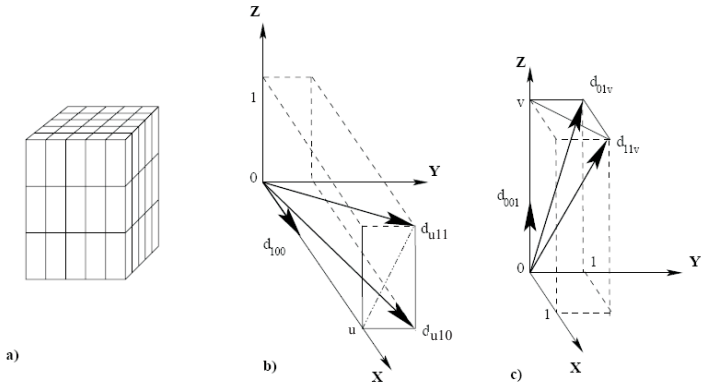


FIG. 4.10: Opérateur UxUxV en maillage parallélépipédique, a) opérateur 5x5x3, b) première portion du premier octant de la sphère, c) dernière portion du premier octant de la sphère

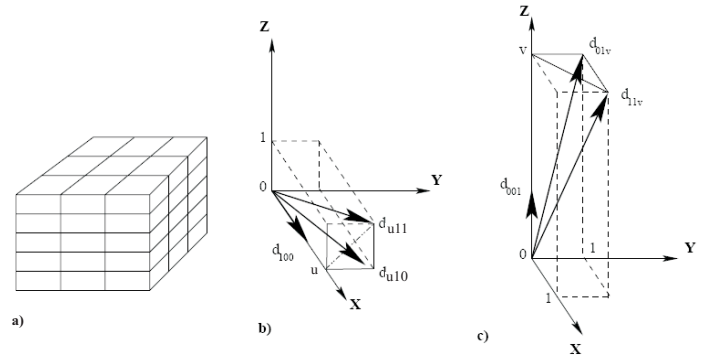


FIG. 4.11: Opérateur UxUxV en maillage parallélépipédique, a) opérateur 3x3x5, b) première portion du premier octant de la sphère, c) dernière portion du premier octant de la sphère

– **deuxième cas** $P \leq 1$: l'erreur maximale se produit dans la dernière portion du premier octant de la sphère (délimitée par les déplacements élémentaires d_{001}, d_{01v} et d_{11v}) et diminue

si P diminue. L'erreur maximale décroît avec la taille du masque. Nous devons alors choisir un masque $U \times U \times V$ avec $U \leq V$ ($U = 2u+1$, et $V = 2v+1$). La figure 4.11 présente le masque d'un opérateur 3x3x5, et les différentes portions sur lesquelles doit se faire l'optimisation.

Optimisation des coefficients :

L'optimisation des coefficients comporte deux procédures :
- **(1)** minimiser l'erreur dans la première portion délimitée par d_{100}, d_{u10} et d_{u11} (Fig. 4.10a) permet de calculer e_1
- **(2)** minimiser l'erreur dans la dernière portion délimitée par d_{001}, d_{01v} et d_{11v} (Fig. 4.11a) permet de calculer e_2

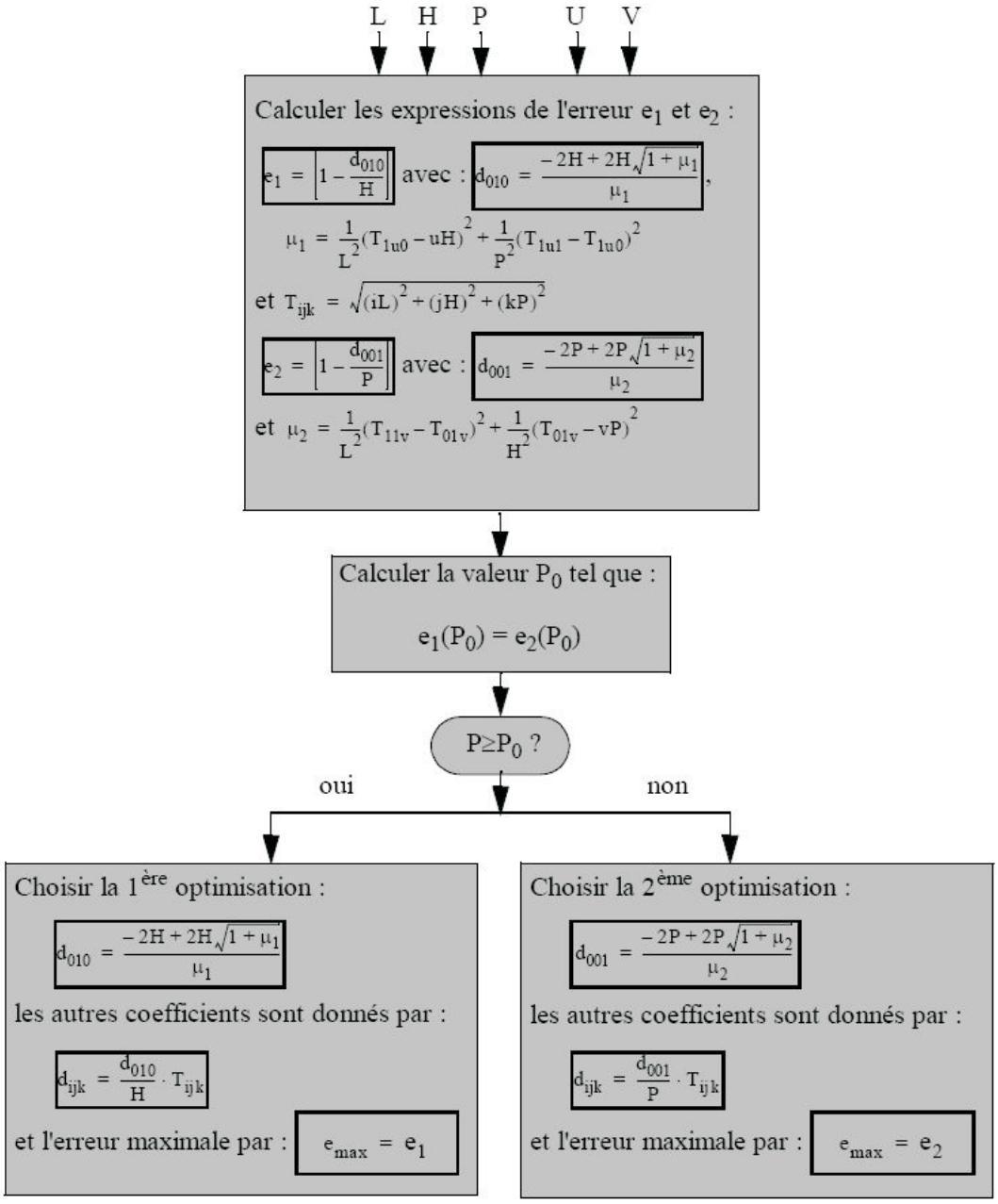


FIG. 4.12: Schéma de l'algorithme d'optimisation d'un opérateur non-cubique $U \times U \times V$ en maillage parallélépipédique, pour $P \geq 1$.

Pour $P \geq 1$, l'étude de e_1 et de e_2 en tant que fonctions de la profondeur P du voxel montre que e_1 est strictement croissante en fonction de P, tandis que e_2 est strictement décroissante. Nous avons $e_1 = e_2$ pour $P = P_0$, avec P_0 solution de l'équation $\mu_1 = \mu_2$.

- **si** $P \leq P_0$ **:** on doit appliquer la seconde procédure d'optimisation. L'erreur maximale normalisée vaut $e_{max} = e_2$.
- **si** $P \geq P_0$ **:** on doit appliquer la première procédure d'optimisation. L'erreur maximale normalisée vaut $e_{max} = e_1$.

La figure 4.12 décrit l'algorithmique permettant l'optimisation des coefficients d'un opérateur non-cubique $U \times U \times V$ en maillage parallélépipédique pour $P \geq 1$. C'est le cas le plus rencontré dans les applications.

Pour $P \leq 1$, l'étude de e_1 et de e_2 en tant que fonctions de la profondeur P du voxel montre que e_1 est strictement croissante en fonction de P, tandis que e_2 est strictement décroissante. Nous avons $e_1 = e_2$ pour $P = P_0^{'}$, avec $P_0^{'}$ solution de l'équation $\mu_1 = \mu_2$.

- **si** $P \leq P_0^{'}$ **:** on doit appliquer la seconde procédure d'optimisation. L'erreur maximale normalisée vaut $e_{max} = e_2$.
- **si** $P \geq P_0^{'}$ **:** on doit appliquer la première procédure d'optimisation. L'erreur maximale normalisée vaut $e_{max} = e_1$.

Etude des performances : Dans ce paragraphe, nous étudierons les performances de l'opérateur anisotrope 3D, et nous donnerons quelques exemples d'opérateurs 3D pour différentes tailles de masque.

Le tableau 4.5 présente les valeurs de l'erreur maximale normalisée produite par des opérateurs 3D de différentes tailles, pour un maillage ($L = H = 1$) et $P = 2$. Comme dans le cas 2D, nous remarquons que nous pouvons utiliser des opérateurs non-cubiques à la place des opérateurs cubiques de grandes tailles en conservant les mêmes performances.

taille du masque	3x3x3	5x5x5	7x7x7	9x9x9	5x5x3
$e_{max}\%$	9.08	4.59	2.61	1.64	4.59
taille du masque	7x7x3	9x9x3	11x11x3	9x9x5	11x11x5
$e_{max}\%$	2.61	2.41	2.41	1.64	1.11

TAB. 4.5: Erreur maximale normalisée produite par des opérateurs de distance 3D cubiques et non-cubiques, $L = H = 1$ et $P = 2$.

La figure 4.13 montre les fonctions d'erreur e_1 et e_2 produites par un opérateur de taille 5x5x3 en utilisant les deux procédures d'optimisation précédemment développées. Dans ce cas, $P_0 = 1.54$. C'est la solution de l'équation $\mu_1 = \mu_2$. La première optimisation est meilleure si $P \geq P_0$, et la seconde est meilleure si $P \leq P_0$. La courbe e_3 représente l'erreur maximale normalisée dans la dernière portion de sphère en utilisant la première optimisation. La courbe e_4 représente l'erreur maximale normalisée dans la première portion de sphère en utilisant la seconde optimisation. La courbe e_2, en pointillé vert, pour $P \leq P_0$ et la courbe e_1, en trait plein bleu, pour $P \geq P_0$ représentent la meilleure erreur maximale normalisée que l'on puisse obtenir.

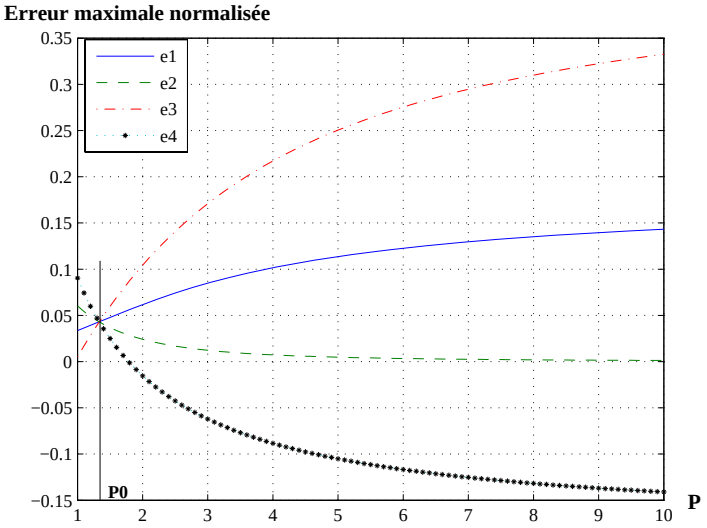


FIG. 4.13: Erreur maximale normalisée en fonction de $P \geq 1$ pour un opérateur 5x5x3

La figure 4.14 montre les fonctions d'erreurs e_1 et e_2 produites par un opérateur de taille 3x3x5 en utilisant les deux procédures d'optimisation précédemment développées. Dans ce cas, $P_0^{'} = 0.61$. C'est la solution de l'équation $\mu_1 = \mu_2$. La première optimisation est meilleure si $P \geq P_0^{'}$, et la seconde est meilleure si $P \leq P_0^{'}$. La courbe e_2, en pointillé vert, pour $P \leq P_0$ et la courbe e_1, en trait plein bleu, pour $P \geq P_0$ représentent la meilleure erreur maximale normalisée que l'on puisse obtenir.

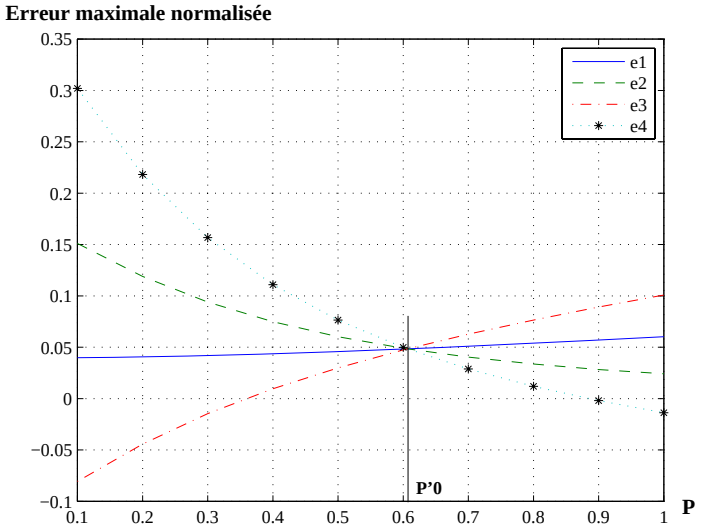


FIG. 4.14: Erreur maximale normalisée en fonction de $P \leq 1$ pour un opérateur 3x3x5

Approximation entière et implémentation : Nous présentons dans cette partie, une nouvelle approche permettant de choisir correctement le facteur d'échelle N utilisé pour l'approximation entière des coefficients du masque. Pour des raisons d'espace mémoire, il est préférable d'utiliser des opérateurs à coefficients entiers. L'approximation entière est réalisée en multipliant les coefficients réels par un entier N et en arrondissant à l'entier le plus proche [Coquin 94]. Nous proposons dans ce paragraphe une **borne inférieure** qui permet de préserver l'exactitude de la transformée de dis-

tance par rapport à la distance euclidienne, et une **borne supérieure** qui garantit qu'il n'y a aucun débordement numérique pour la représentation de la distance.

Soit i, le nombre de bits utilisé pour coder les valeurs de distance entre chaque voxel. Soit Dim, la dimension de l'image. La distance maximale qui doit être codée est :

$$D_{max} = Dim.round(N.d_{111}) = Dim.(N.d_{111} + q) \tag{4.31}$$

avec q l'erreur d'arrondi telle que $|q| \leq \frac{1}{2}$

Pour une grande valeur de N, l'erreur d'arrondi peut être négligée, ainsi nous avons :

$$N_{max} < \frac{2^i}{Dim.d_{111}} \tag{4.32}$$

avec

$$d_{111} = \left(\sqrt{2+P^2}\right)\left[\frac{-2+2\sqrt{1+\lambda}}{\lambda}\right] \tag{4.33}$$

et

$$\lambda = \left(\sqrt{(u^2+1)} - m\right)^2 + \frac{1}{P^2}\left[\sqrt{(u^2+1+P^2)} - \sqrt{(u^2+1)}\right]^2 \tag{4.34}$$

où $U = 2u + 1$ est la taille de l'opérateur.

La valeur minimale est choisie de sorte que l'erreur d'arrondi soit du même ordre de grandeur que la transformée de distance.

Soit E_{max} l'erreur maximale de la transformée de distance. Comme il est montré dans [Verwer 91], l'erreur maximale est proportionnelle au rayon R. Elle est obtenue au centre du premier cône dans la direction d_{100}. Soit D_{100}, l'approximation entière du déplacement élémentaire d_{100}. Nous avons :

$$D_{100} = round(N.d_{100}) \tag{4.35}$$

et l'erreur relative induite par l'approximation entière est

$$|\epsilon| = \frac{|round(N.d_{100}) - N.d_{100}|}{N.d_{100}} = \frac{|q|}{N.d_{100}} \tag{4.36}$$

avec $|q| \leq \frac{1}{2}$

Pour $L = H = 1$, la valeur absolue de l'erreur maximale produite par l'opérateur local de distance est :

$$|e_{max}| = \frac{|E_{max}|}{R} = |1 - d_{100}| \tag{4.37}$$

En utilisant les équations 4.37 et 4.21, on en déduit la valeur minimale du facteur d'échelle qui doit satisfaire

$$N_{min} > \frac{1}{2.d_{100}.e_{max}} \tag{4.38}$$

avec

$$d_{100} = \frac{-2+2\sqrt{1+\lambda}}{\lambda} \tag{4.39}$$

et

$$\lambda = \left(\sqrt{(u^2+1)} - u\right)^2 + \frac{1}{P^2}\left[\sqrt{(u^2+1+P^2)} - \sqrt{(u^2+1)}\right]^2 \tag{4.40}$$

où $U = 2u + 1$ est la taille du masque de l'opérateur.

Nous pouvons remarquer que N_{min} dépend de la profondeur P du voxel.

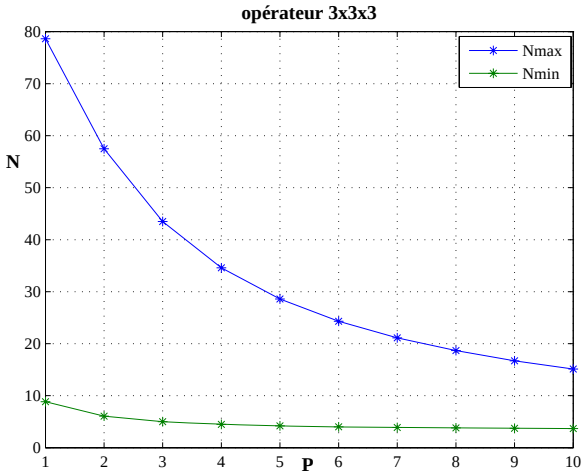


FIG. 4.15: Bornes N_{max} et N_{min} du facteur d'échelle pour un opérateur 3x3x3 en fonction de la profondeur P du voxel

Nous pouvons remarquer que l'équation 4.32 est valable pour des volumes ayant le même nombre de lignes que de colonnes. Pour l'équation 4.35, nous avons considéré que les coefficients réels sont optimaux. Dans la pratique, en utilisant $d_{100} = 1$ et $d_{111} = \sqrt{2 + P^2}$, on obtient le meilleur estimateur qui soit pour N_{min} et N_{max}, que nous nommerons $N_{min-approx}$ et $N_{max-approx}$.

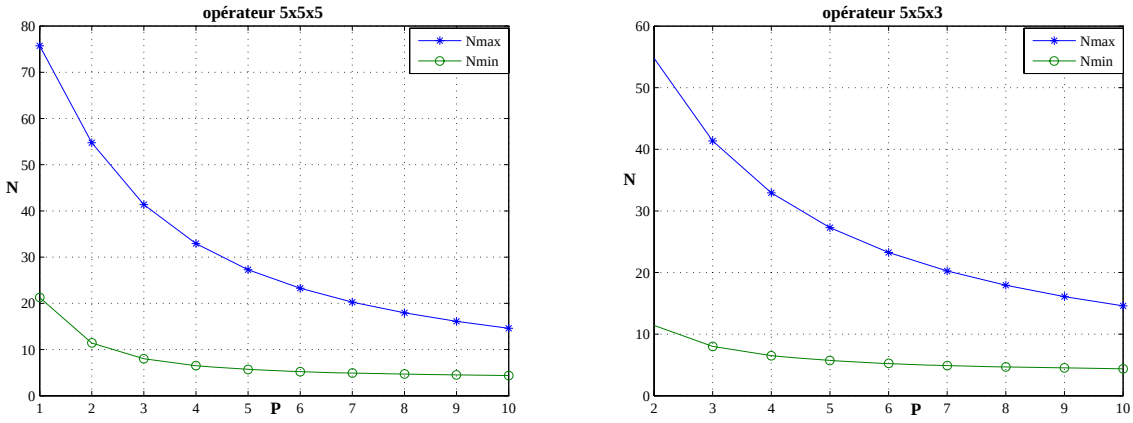


FIG. 4.16: Bornes N_{max} et N_{min} du facteur d'échelle pour un opérateur 5x5x5 et 5x5x3 en fonction de la profondeur P du voxel

Les figures Fig. 4.15 et Fig. 4.16 montrent les variations des bornes supérieure N_{max} et inférieure N_{min} du facteur d'échelle en fonction de la profondeur P du voxel, pour les opérateurs de distances 3x3x3, 5x5x5 et 5x5x3.

Le tableau 4.6 nous donne la meilleure approximation entière pour quelques opérateurs 3D cubiques pour $L = H = P = 1$ et non-cubiques pour $L = H = 1$ et $P = 2$. Nous avons considéré une dimension d'image $Dim = 512$ et un codage sur $i = 16$ bits. Le facteur d'échelle N et l'erreur relative maximale e_{max} sont donnés pour chaque opérateur.

L,H,P	(1,1,1)	(1,1,1)	(1,1,2)	(1,1,2)
taille	3x3x3	5x5x5	5x5x5	5x5x3
N_{min}	8.84	21.25	11.42	11.42
N_{max}	78.63	75.72	54.77	54.77
$N_{min-approx}$	8.31	20.74	10.89	10.89
$N_{max-approx}$	73.90	73.90	52.25	52.25
N	67	43	22	44
D_{100}, D_{010}	63	42	21	42
D_{001}	63	42	42	84
D_{110}	89	59	30	59
D_{101}, D_{011}	89	59	47	94
D_{111}	109	73	51	103
D_{210}, D_{120}		94	47	94
D_{201}, D_{021}		94	59	119
D_{102}, D_{012}		94	87	
D_{211}, D_{121}		103	63	126
D_{112}		103	89	
D_{221}		126	73	145
D_{212}, D_{122}		126	96	
e_{max}	6.073	2.563	4.647	4.644

TAB. 4.6: La meilleure approximation entière pour quelques opérateurs 3D isotropes $L = H = P = 1$ et opérateurs 3D anisotropes avec $L = H = 1$, $P = 2$, $i = 16$ bits, $Dim = 512$.

A partir du tableau 4.6, nous pouvons voir que l'erreur maximale obtenue par ces opérateurs à coefficients entiers est proche de la valeur théorique (obtenue pour des coefficients réels avec la même taille d'opérateur). Nous remarquons que nous obtenons la même erreur maximale pour un opérateur cubique de taille 5x5x5 que pour un opérateur non-cubique de taille 5x5x3, dans le cas où $L = H = 1$ et $P = 2$. Comme en 2D, ces opérateurs non-cubiques sont très intéressants car nous conservons les mêmes performances en terme d'erreur maximale normalisée tout en gagnant en temps de traitement.

Discussion : A ce stade, nous pourrions nous demander pourquoi nous n'avons pas traité l'optimisation des opérateurs $U \times V \times W$ avec $U \neq V \neq W$? Nous avons effectivement pensé à généraliser notre démarche, mais nous ne l'avons pas faite car, la plupart, pour ne pas dire la majorité, des systèmes d'acquisition actuels, utilise le même pas d'échantillonnage dans deux directions, et un pas différents dans la troisième direction.

4.5.3 Optimisation des opérateurs non-stationnaires 3D

Nous avons introduit, en 1999, l'étude des opérateurs non-stationnaires en 3D, par le stage de DEA d'Onéa Alexandru. L'objectif de ces opérateurs est de pouvoir s'adapter aux grilles éventuellement non-uniformes, selon une direction (acquisition de coupes scanner successives avec un pas variable entre ces coupes). Nous présentons la démarche que nous avons développée dans [Coquin 00a]. C'est une extension des opérateurs non-cubiques.

Nous nous plaçons dans le cas $L = H = 1$ et $P > 1$. C'est le cas de la plupart des applications. Nous avons montré, dans la section précédente, que pour $P > 1$, l'erreur maximale normalisée se situait dans la portion 1 (ou la portion 4) du premier octant de la sphère, délimité par les directions élémentaires d_{100}, d_{u10} et d_{u11}. L'étude de l'évolution de cette erreur montre que plus la profondeur P du voxel augmente et plus l'erreur maximale normalisée augmente dans cette portion de sphère.

La figure 4.17 montre une coupe du volume image 3D, avec un pas d'échantillonnage variable dans la direction Z. Soit P_r la variation entre les couches C_r et C_{r+1} (deux coupes scanner, ...) dans la direction Z. P_r est défini par : $P_r = C_{r+1} - C_r$.

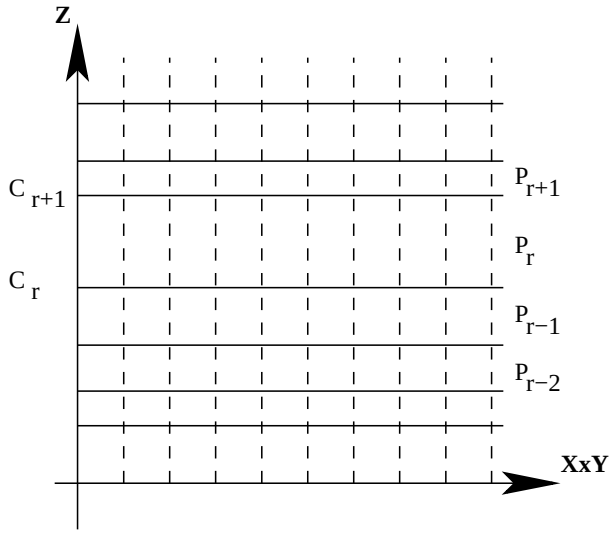


FIG. 4.17: Coupe du volume image 3D

Soit $U = 2u+1$, la taille du masque. Les coefficients du masque doivent être calculés pour chaque couche C_r. Les coefficients réels sont donnés par :

$$d^r_{100} = \frac{-2L + 2L\sqrt{1+\lambda_r}}{\lambda_r} \tag{4.41}$$

avec

$$\lambda_r = \frac{1}{L^2}(T^r_{u10} - uL)^2 + \frac{1}{P_r^2}(T^r_{u11} - T^r_{u10})^2 \tag{4.42}$$

Les autres coefficients d^r_{ijk} sont donnés par :

$$d^r_{ijk} = T^r_{ijk}\frac{d^r_{100}}{L} \tag{4.43}$$

avec

$$T^r_{ijk} = \sqrt{(iL)^2 + (jH)^2 + (kP_r)^2} \tag{4.44}$$

La figure 4.18 montre un masque de chanfrein 3x3x3 non-stationnaire. Nous avons autant de masques différents que de couches de profondeur P_r différentes. Les coefficients entiers D^r_{ijk} utilisés dans le masque de chanfrein sont calculés comme précédemment, en multipliant les coefficients réels d^r_{ijk} par un entier N et en arrondissant à l'entier le plus proche conformément à :

$$D^r_{ijk} = round[N.d^r_{ijk}] \tag{4.45}$$

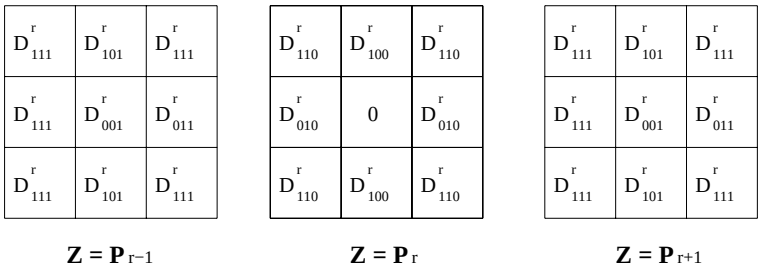


FIG. 4.18: Masque de chanfrein 3D non-stationnaire

Cette manière de faire soulève plusieurs remarques :
- les propriétés d'une distance sont-elles respectées ?
- pour l'approximation entière, il faut s'assurer que dans le plan $G = P_r$, les déplacements élémentaires D^r_{110}, D^r_{100} et D^r_{010} restent constants d'un plan à l'autre. Donc ces coefficients doivent satisfaire à la condition suivante :

$$D^r_{ij0} = D^{r+1}_{ij0} \quad r \in 1, 2, ..., U-1 \tag{4.46}$$

- les coefficients entiers D^r_{ijk} utilisés dans le masque de chanfrein ne sont pas optimaux en raison de l'arrondi et du fait que le calcul réel devrait tenir compte de la minimisation de l'erreur entre la distance locale et la distance euclidienne. Cela devrait se faire pour chaque couche. Cependant, le masque de chanfrein que nous avons utilisé, s'adapte à la profondeur de chaque couche et les déplacements élémentaires sont très proches des valeurs réelles,
- La plupart des couches ont une profondeur supérieure à l'unité. Bien que le résultat soit sous-optimal, les couches ayant une profondeur $P \leq 1$ sont traitées comme les autres,
- à chaque couche, seuls les 3 nouveaux coefficients D^r_{001}, D^r_{101} et D^r_{111} sont calculés.

Discussion : Il est évident que ces travaux sur les opérateurs locaux de distances non-stationnaires 3D sont une première réflexion sur ce qui devrait être fait. Les différents points que nous avons soulevés précédemment mériteraient d'être approfondis et une étude plus poussée devrait être réalisée. Nous verrons dans le chapitre 5 l'utilisation de cet opérateur local de distance non-stationnaire 3D appliquée à la **comparaison** des images couleur avec, notamment, des résultats encourageants.

4.6 Conclusion

Nous avons montré, dans ce chapitre, les différents travaux qui nous ont préoccupés pendant près de 10 années. Nous avons développé une panoplie d'opérateurs locaux de distances en 2D puis en 3D s'adaptant aux maillages rectangulaires et parallélépipédiques. Le but était de travailler directement à partir des données qui nous étaient fournies.

Nous avons proposé des opérateurs de distance 2D et 3D, cubiques et non-cubiques, voire non-stationnaires en 3D. Ces opérateurs sont optimisés en suivant les approches de Borgefors et de Verwer. Pour le 3D, nous nous sommes limités au cas où le pas d'échantillonnage en x et en y était identique, tout en étant différents du pas d'échantillonnage en z. Cependant, la méthode développée s'applique de la même manière dans le cas général.

Nous avons montré, qu'en 2D tout comme en 3D, il était préférable de travailler avec des opérateurs non-cubiques, lorsque cela était possible, puisqu'ils possèdent les mêmes propriétés en terme d'anisotropie que les opérateurs cubiques pour un temps de traitement plus faible.

Nous avons côtoyé la communauté de la **Géométrie Discrète** avec qui nous avons beaucoup échangé. Comme nous l'avons vu au début de ce chapitre, les opérateurs de distances ont fait coulé beaucoup d'encre, avec les défenseurs des distances exactes, et ceux des distances de chanfrein. Les moyens informatiques permettant maintenant de faire les calculs plus rapidemment et permettant de stocker d'avantage d'information, les transformées de distances exactes semblent, à l'heure actuelle, avoir la préférence, sauf lorsque les données à traiter sont gigantesques. Mais je dirais que chaque transformée de distance a son utilité pour l'application pour laquelle elle est destinée.

Nous avons dans la suite orienté nos travaux sur les applications de ces opérateurs en vue :
- de comparer les images binaires, en niveaux de gris et en couleur,
- de caractériser et évaluer les traitements,
- de reconnaître les objets.

C'est ce que nous allons développer dans les prochains chapitres de ce mémoire.

Chapitre 5

Les mesures de dissimilarité

Résumé : Dans ce chapitre, nous allons décrire la mesure de dissimilarité que nous proposons pour comparer de manière directe deux images binaires, deux images à niveaux de gris ou deux images couleurs. La comparaison de séquences d'images sera détaillée au chapitre 6. Nous montrons qu'il est possible d'avoir une information globale, correspondant à la distance entre deux images, mais également une information locale, en utilisant les cartes de distances. Ces informations sont d'ordre géométrique et/ou radiométrique. Elles permettent de caractériser certaines transformations d'images.

5.1 Introduction

Notre approche de la **comparaison directe des images** porte sur la comparaison globale de l'image par une mesure de dissimilarité. La mesure de dissimilarité que nous proposons nécessite le calcul de distances entre un point et un ensemble de points. Nous avons donc utilisé pour cela les opérateurs locaux de distance 2D et 3D développés dans le chapitre précédent. La mesure de dissimilarité décrite ici peut être calculée sur n'importe quelle paire d'images binaires, à niveaux de gris ou en couleurs. Cette mesure de dissimilarité donne également des renseignements utiles sur la localisation des classes de distorsions.

5.2 Mesure de dissimilarité entre images

Les travaux que nous avons entrepris dans ce domaine se situent entre 1995 et 2002. A l'époque, quelques distances avaient été proposées pour mesurer la dissimilarité entre des objets dans les images binaires [Baddeley 92], [Huttenlocher 93], et [Dubuisson 94]. Ces mesures de dissimilarité fournissaient une valeur globale correspondant à la valeur moyenne entre les deux images binaires. D'autres critères objectifs avaient été proposés, par la suite, pour mesurer la dissimilarité entre images à niveaux de gris [Coquin 95b], [Zamperoni 96], [Coquin 97], [Wilson 97], [Di Gesù 99] et dans le but de retrouver des images [Jacobs 00]. Zamperoni et Starovoitov (1996) ont proposé une mesure de dissimilarité multi-échelle dans laquelle chaque échelle (pixel à pixel, pixel à fenêtre, fenêtre à fenêtre, et image à image) peut être basée sur différentes mesures, ce qui entraîne différentes variantes de dissimilarités. Wilson et al. (1997) ont proposé une extension aux images en niveaux de gris de la distance de Baddeley. Di Gesù et Starovoitov (1999) ont proposé trois fonctions de distance entre images qui peuvent être utilisées pour la comparaison d'images. Les résultats expérimentaux montrent que la fonction qui a la meilleure sensibilité est celle qui combine l'intensité des pixels à la structure locale des caractéristiques de l'image. Contrairement au critère du RMS (Root Mean Square = Racine carrée de la moyenne du carré des valeurs), qui ne tient compte que de la différence des intensités entre les images, ces approches combinent des comparaisons en intensité (radiométrique) et dans le domaine spatial (géométrique).

Ces différents auteurs ont montré l'intérêt de choisir une distance basée sur la combinaison de

déplacements spatiaux et en niveaux de gris. Dans cette section, nous introduisons la mesure de dissimilarité que nous avons développée, qui est une extension de la distance de Baddeley aux images en niveaux de gris. Cette mesure est basée sur l'accumulation d'informations locales de distance. Le principal avantage de cette mesure de dissimilarité réside dans le contrôle de l'importance de chaque classe de distorsions en ajustant un paramètre de l'opérateur local de distance (*le rapport P/H entre la profondeur et la largeur du voxel*). Pour chaque classe, une carte de distances peut être calculée [Coquin 95b] et [Chehadeh 96].

Nous allons maintenant décrire la démarche qui nous a conduits à cette mesure de dissimilarité.

5.2.1 Conditions auxquelles doit satisfaire une mesure de dissimilarité entre images

Pour que la mesure de dissimilarité D satisfasse aux critères d'une distance, elle doit vérifier les conditions suivantes :

- **1.** $D(A, B)$ doit être égale à 0 si $A = B$ (identité) et égale à une grande valeur pour la dissimilarité extrêmale. Si on considère une échelle de niveaux de gris entre 0 et 255, la plus grande valeur de la dissimilarité doit avoir lieu entre une image uniformément blanche et une image uniformément noire. La mesure de dissimilarité doit avoir un grand pouvoir discriminant, sa valeur doit croître avec la différence entre deux objets.
- **2.** D doit satisfaire aux propriétés d'une distance à savoir :
  - (i) $D(A, B) \geq 0$, et $D(A, B) = 0$ si et seulement si $A = B$
  - (ii) $D(A, B) = D(B, A)$
  - (iii) $D(A, C) \leq D(A, B) + D(B, C)$, (inégalité triangulaire). Il arrive parfois que nous soyons obligés de travailler avec une mesure qui ne soit pas forcément une distance. Nous considérerons alors la propriété 2-(iii) comme une condition souhaitée mais non obligatoire.
- **3.** D doit être sensible à la déformation des objets, au déplacement des contours, à une variation du niveau de gris moyen .....
- **4.** Au moins pour de petites translations entre deux images, D devrait augmenter.
- **5.** Si l'image B était obtenue par addition d'un bruit blanc à l'image A, et si C est la version restaurée de l'image B, alors on devrait avoir $D(A, B) \geq D(A, C)$.

Les méthodes basées sur l'***erreur quadratique moyenne*** ne sont pas adaptées puisque ce critère dépend uniquement de la différence des niveaux de gris, ne prenant en compte l'information géométrique que *via* la fonction d'autocorrélation de l'image. Les méthodes basées sur le ***flux optique*** ne conviennent pas lorsque le déplacement est trop important ou lorsqu'il y a des fortes variations de luminosité d'une image à l'autre. Les méthodes basées sur l'***intercorrélation*** numérique entre images à niveaux de gris sont quant-à-elles performantes du point de vue géométrique mais sont biaisées par les variations de niveaux de gris. Elles nécessitent la connaissance de la taille du motif ayant subi une déformation et sont de ce fait, sensibles aux effets de bords. Nous allons donc proposer une dissimilarité permettant de déceler aussi bien une variation de niveau de gris qu'une variation de forme ou un déplacement spatial.

La plupart des distances et dissimilarités citées dans le paragraphe précédent vérifient ces conditions. Mais le choix de la mesure de dissimilarité à adopter sera guidé par la façon dont nous allons représenter l'image.

5.2.2 Représentation d'une image

Les deux approches classiques de **comparaison quantitative** d'images sont basées sur les deux modèles classiques d'image : considérer une image en tant que fonction ou en tant qu'ensemble de points.

Soit $S \subset \mathbb{Z}^2$, tel que $S = N \times M$ le domaine sur lequel les images sont définies. Soit $G = \{0, 1, 2, ..., 255\}$ l'ensemble des niveaux de gris, ($G = \{0, 1\}$ *pour une image binaire*). Soient A et B deux images en niveaux de gris. A étant l'image de référence.

L'image A peut être représentée de deux façons différentes :

– **en tant que fonction** : sous la forme d'une fonction 2D notée f_A définie de la façon suivante :

$$f_A : S \rightarrow G \tag{5.1}$$

telle que $f_A(x, y) = g$ où g est la valeur du niveau de gris du pixel (x, y) dans l'image A. Avec ce modèle, l'opérateur de comparaison peut être défini par :

$$\mathcal{D}(A, B) = \Psi\{\sum_{s \in S} \Phi[f_A(s) - f_B(s)]\} \tag{5.2}$$

avec Ψ une fonction positive et croissante avec $\Psi(0) = 0$, et Φ une fonction positive de préférence paire. Alors, la dissimilarité entre deux images est obtenue par accumulation d'informations locales, sans tenir compte de la position spatiale des objets de l'image de référence par rapport à l'image de test.

Exemple : avec Φ la valeur absolue et Ψ la moyenne, on obtient le critère de la moyenne des différences absolues. Avec Φ la fonction élévation au carré et Ψ la racine carrée de la moyenne, on obtient le critère RMS. Ceci peut être généralisé aux distances de Minkowski en utilisant des fonctions puissance.

– **en tant qu'ensemble de points** : sous la forme d'une fonction 3D notée F_A définie de la façon suivante :

$$F_A : S \times G \rightarrow \{0, 1\} \tag{5.3}$$

telle que :

$$F_A(s, g) = \begin{cases} 1 & si \quad f_A(s) = g \\ 0 & sinon \end{cases} \tag{5.4}$$

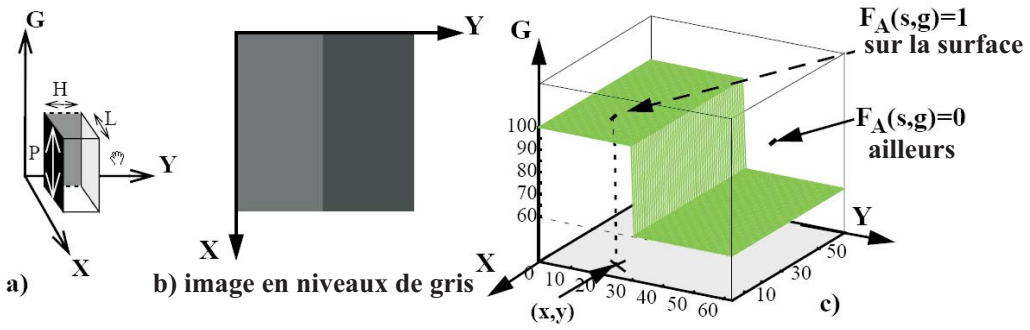


FIG. 5.1: Représentation d'une image comme un ensemble de voxels.

Soit A l'image de référence ayant toujours comme support $S = N \times M$ (N, M sont les dimensions de l'image) et comme échelle des niveaux de gris G. L'image A peut être représentée par l'ensemble des voxels $X_A \subset S \times G$, où $\times$ désigne le produit cartésien (voir Fig. 5.1).

$$X_A = \{v(s, g) \in S \times G, f_A(s) = g\} \tag{5.5}$$

A partir de cette définition, les mesures de dissimilarité peuvent être définies comme étant la distance entre deux ensembles de points.

$$\mathcal{D}(A, B) = d(X_A, X_B) \tag{5.6}$$

Dans cette catégorie, on trouve la distance de Hausdorff [Huttenlocher 93], [Dubuisson 94] et la distance de Baddeley [Baddeley 92].

– **la distance de Hausdorff** entre deux ensembles de points X et Y est définie par :

$$H(X,Y) = max\{sup_{x\in X}[d(x,Y)], sup_{y\in Y}[d(y,X)]\} \tag{5.7}$$

avec $d(x,Y) = \min_{y\in Y}\{d(x,y)\}$. Si on adapte cette définition aux images en niveaux de gris, le critère devient :

$$D_H(X_A,X_B) = max\{sup_{a\in X_A}[d(a,X_B)], sup_{b\in X_B}[d(b,X_A)]\} \tag{5.8}$$

– **la distance de Baddeley** entre deux ensembles X et Y inclus dans le référentiel $S = N\times M$ est définie par :

$$B(X,Y) = [\frac{1}{N\times M}\sum_{s\in S}|d(s,X) - d(s,Y)|^q]^{\frac{1}{q}} \tag{5.9}$$

avec l'exposant q tel que $1 \leq q \leq \infty$. Cette distance a été proposée en 1992, pour calculer l'erreur entre deux images binaires [Baddeley 92]. La distance de Baddeley est une distance moyenne. Elle est moins sensible que la distance de Hausdorff pour de petites distorsions localisées, et elle résiste bien à un bruit additif.

Dans la suite, notre choix s'est porté sur la distance de Baddeley, pour les raisons évoquées précédemment. Nous l'utilisons pour comparer des images binaires puis, nous en proposerons une extension pour comparer des images en niveaux de gris, et enfin en couleur.

5.3 Comparaison d'images binaires

Ce travail se situe dans le contexte plus général du contrôle d'un poste de travail par reconnaissance des gestes de la main d'un opérateur[1]. Les mesures primaires des angles des articulations des doigts de la main sont effectuées à l'aide d'un gant numérique. Au même niveau, un second module de traitement basé sur un système de reconnaissance par vision vient enrichir la prise de décision, permettant en particulier la prise en compte du mouvement global de la main. Dans cette partie, nous nous intéressons à ce second module dont l'objectif est la reconnaissance d'un geste dynamique de la main. L'acquisition des séquences du geste est réalisée dans un environnement presque idéal puisque la main et l'avant-bras se détachent facilement d'un fond homogène et sombre (voir Fig. 5.2). Dans un environnement moins favorable il faudrait mettre en œuvre une méthode de segmentation basée sur la détection de la peau et une méthode de suivi de gestes.

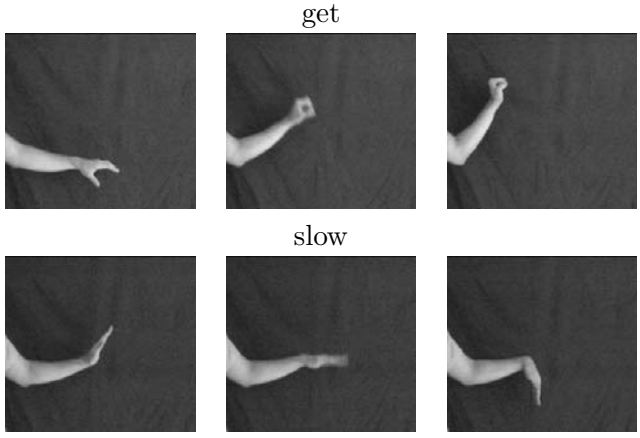


FIG. 5.2: Les gestes "GET" et "SLOW".

[1]Projet BQR : 2002-2004

Comme nous l'avons évoqué au chapitre 3, la **comparaison** peut se faire à partir des signatures des images. Nous avons proposé une méthode de reconnaissance de gestes de la main basée sur l'utilisation conjointe de signatures statiques et d'une signature dynamique. Les signatures statiques, composées de l'histogramme des orientations du gradient des pixels de la boite englobante de la main et de l'avant-bras, permettent de détecter la partie dynamique utile. La signature dynamique proposée est une solution originale qui consiste à superposer les squelettes (**images binaires**) des différentes images composant le geste. Nous avons utilisé la distance de Baddeley pour **comparer** les signatures dynamiques des gestes de la main.

5.3.1 Signature statique et signature dynamique

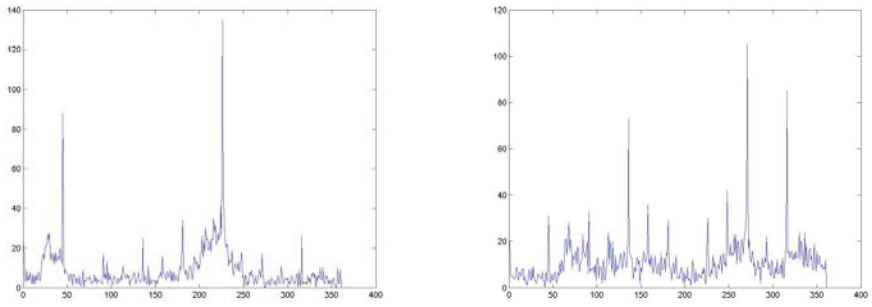
L'analyse des gestes suppose que l'on fasse de l'analyse spatiale et de l'analyse temporelle. De nombreuses méthodes de reconnaissance de gestes existent dans la littérature. Nous avons retenu deux approches : une première basée sur l'histogramme des orientations du gradient et une seconde utilisant le squelette de la main et de l'avant-bras. L'utilisation conjointe de ces deux approches permet d'obtenir une meilleure robustesse. Dans notre application, un geste dynamique est caractérisé par une séquence comportant de 30 à 50 images, le début et la fin de la séquence étant supposés connus.

L'histogramme des orientations du gradient est utilisé comme une signature statique calculée sur la première et la dernière image de la séquence. Par contre le squelette est utilisé comme une signature dynamique calculée sur chaque image de la séquence.

Signature statique : L'histogramme des orientations du gradient n'est calculé que sur l'ensemble des pixels de la boite englobante de l'avant-bras et la main, le fond de l'image n'étant pas forcément homogène. Ce calcul comprend quatre étapes :

- élimination automatique du fond (binarisation et enchaînement de quelques opérateurs morphologiques),
- calcul du gradient sur chaque pixel de la main et de l'avant bras,
- élimination des pixels dont le module du gradient est inférieur à K fois le module moyen. Des tests ont conduit à prendre $K = 1.2$,
- calcul de l'histogramme des orientations du gradient.

La figure 5.3 montre les histogrammes des orientations du gradient des gestes "GET" et "SLOW".



(a) (b)

FIG. 5.3: Histogrammes des orientations du gradient pour (a) le geste "GET" et (b) le geste "SLOW"

Signature dynamique : Les squelettes sont calculés à partir des images binaires de la séquence du geste dynamique. Nous avons utilisé la méthode développée dans [Chehadeh 96] pour extraire le squelette simplifié pour chaque image de la séquence. Les détails sur le calcul du squelette sont donnés

dans [Ionescu 03], il est basé sur le suivi et l'extraction de la ligne de crête maximale calculée sur les images de distance de la main et l'avant-bras (Fig. 5.4).

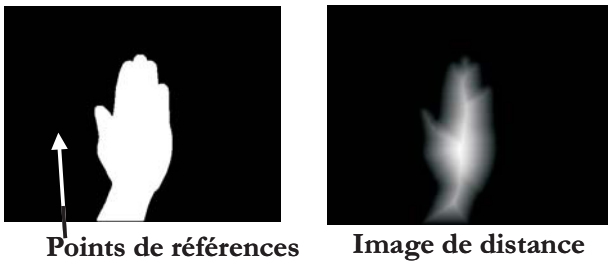


FIG. 5.4: Image de distance de la main

La superposition de l'ensemble de ces squelettes va constituer la signature dynamique du geste. L'image de droite de figure 5.5 représente la signature dynamique de la séquence "GET".

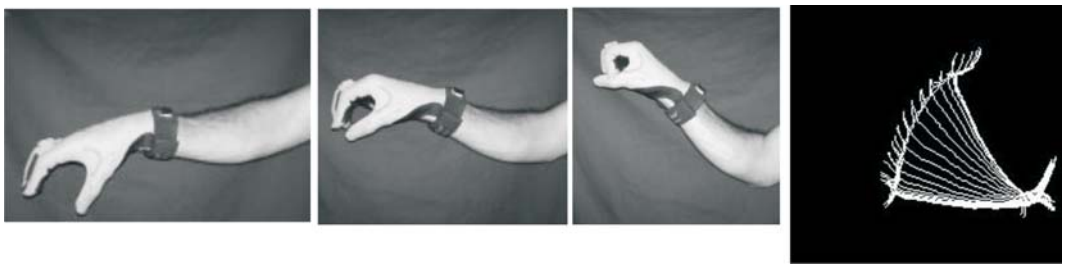


FIG. 5.5: Geste "GET" et signature dynamique

La figure 5.6 nous montre la signature dynamique de deux séquences différentes : le geste "GET" et le geste "SLOW". Nous avons implémenté cette méthode dans un PC muni d'une carte d'acquisition et de traitement Matrox. Nous avons testé notre méthode pour la commande d'un mini robot.

(a) (b)

FIG. 5.6: Signature dynamique de la séquence (a) "GET" et (b) "SLOW"

Mesures de dissimilarité entre images binaires

Nous avons utilisé la distance de Baddeley entre deux ensembles X et Y (*ici les squelettes*) inclus

dans le référentiel $S = N \times M$ (*dimension de l'image*) qui est définie par :

$$B(X,Y) = [\frac{1}{N \times M} \sum_{s \in S} |d(s,X) - d(s,Y)|^q]^{\frac{1}{q}} \tag{5.10}$$

avec l'exposant q tel que $1 \leq q \leq \infty$.

Pour l'implémentation de cette distance, il faut définir la distance $d(s,X)$ entre un pixel s de l'image et un ensemble X de référence ici le squelette. Nous avons tout naturellement utilisé nos opérateurs locaux de distance en 2D pour calculer ces distances en prenant un masque 5×5, et $L = H = 1$. Dans ces images de distances, chaque pixel du fond est étiqueté à la distance la plus courte au squelette. Nous avons également choisi de prendre comme valeur de l'exposant $q = 2$.

5.3.2 Résultats

La reconnaissance du geste s'effectue en deux temps. D'abord, on cherche à reconnaître les signatures statiques de début et de fin de geste. Ces signatures, une fois reconnues, permettent de délimiter le calcul de la signature dynamique. Dans un deuxième temps, on procède à la reconnaissance de cette signature dynamique. Le principe de la reconnaissance d'un geste repose sur la **comparaison** entre les signatures du geste courant et les signatures des séquences d'apprentissage correspondant à un alphabet de gestes connus. Le geste reconnu est associé à la distance la plus petite. Pour la signature statique, nous utilisons simplement la distance euclidienne entre les histogrammes des orientations du gradient. Pour la signature dynamique, nous devons comparer deux **images binaires** représentant la superposition des squelettes de la séquence courante et des séquences d'apprentissage. La **comparaison** se fait en calculant la distance de Baddeley [Baddeley 92]. Cette méthode a été testée sur 40 séquences de geste, pour un alphabet de 10 gestes. Les performances obtenues dépendent de la nature du geste (dynamique ou statique) et du nombre de séquences d'apprentissage utilisé pour un même geste. En effet, pour rendre la méthode robuste à la variabilité pouvant apparaître sur un même geste, la base d'apprentissage peut éventuellement comporter plusieurs modèles pour un même geste [Ionescu 05].

Tab. 5.1: Evolution du taux de reconnaissance (en %) en fonction du nombre de séquences d'apprentissage.

Nb séq. apprentissage	geste dynamique
1	48
2	72
3	82
4	84
5	92
6	94

Nous avons poursuivi ce travail dans les deux directions suivantes :

- d'une part, vers l'amélioration des performances en terme de temps de calcul. Si la reconnaissance des gestes statiques est rapide et permet d'envisager une implantation temps réel, la reconnaissance du geste dynamique a nécessité une phase d'optimisation et d'accélération des traitements, par implémentation des algorithmes sur un PC utilisant une carte de traitement d'images (Matrox).
- d'autre part, les résultats fournis par l'analyse des images ont été fusionnés avec ceux fournis par le gant numérique. Le gant numérique donne la position de la main et plus particulièrement des doigts de la main. La caméra vidéo donne l'allure du geste. Les mesures peuvent être partiellement complémentaires et/ou partiellement redondantes. Nous avons développé une

stratégie basée sur le degré de confiance associé à chacun des capteurs afin de prendre en compte, soit le gant numérique, soit la caméra dans le processus de décision finale pour la reconnaissance du geste [Coquin 06].

Pour utiliser la distance de Baddeley avec des images en niveaux de gris, il faut proposer une extension. Nous allons dans la section suivante exposer une solution que nous avons retenue.

5.4 Mesure de dissimilarité proposée

Wilson et al. [Wilson 97] ont proposé une extension de la distance de Baddeley aux images en niveaux de gris. Le calcul se fait pour tous les voxels appartenant à un sous-graphe (uniquement les voxels situés à proximité de la surface ou image contribuent à la mesure de cette dissimilarité). Cette restriction introduit une certaine asymétrie dans le processus de calcul. Par exemple, la dissimilarité entre deux images est différente de celle entre leurs inverses vidéo. Des détails sur les propriétés de cette dissimilarité sont donnés dans [Coquin 00a].

Nous avons donc proposé une ***mesure de dissimilarité*** qui est une extension de la distance de Wilson aux cas des images en niveaux de gris, sans la restreindre au sous-graphe. Cette dissimilarité peut également être vue comme une extension aux images en niveaux de gris de la distance de Baddeley. Elle nécessite la distance d'un voxel à un ensemble de voxels (ici la surface de référence donc l'image). Nous utilisons un opérateur local de distance 3D dépendant des dimensions L, H et P caractérisant le voxel pour calculer ces distances entre points et ensemble de points. L'élément de base est le calcul de la distance entre voxels en maillage parallélépipédique. L'ensemble des points constituant l'image forme une surface dans l'espace 3D. Tous les points de cette surface sont étiquetés à 1. C'est donc un ensemble binaire dans l'espace 3D. Soient A et B deux ensembles binaires (ici les surfaces) à comparer dans un volume V, contenant $Card(V)$ voxels. Notre dissimilarité entre A et B est définie par :

$$\mathcal{D}(A,B) = [\frac{1}{Card(V)} \sum_{v \in V} |d_A(v) - d_B(v)|^q]^{\frac{1}{q}} \tag{5.11}$$

avec l'exposant q tel que $1 \leq q \leq \infty$, $V = S \times G$ est le volume sur lequel la dissimilarité est calculée, $d_A(v)$ est la distance entre le voxel v et l'ensemble binaire caractérisant l'image A (voir Fig. 5.7).

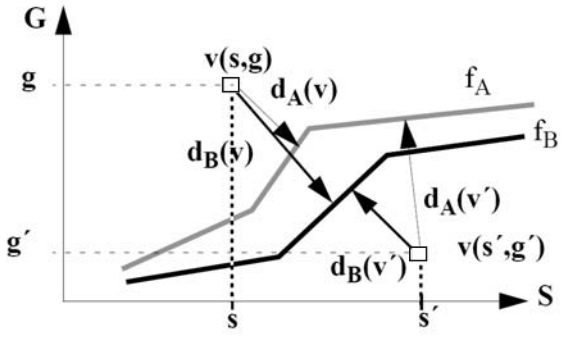


FIG. 5.7: Dissimilarité entre deux images.

Nous avons donc choisi d'appliquer le calcul de cette dissimilarité à tous les voxels du volume V.

5.4.1 Implémentation

Afin de réduire le temps de calcul et l'espace mémoire, les distances euclidiennes $d_A(v)$ et $d_B(v)$ sont approximées au moyen d'un opérateur local de distance 3D en maillage parallélépipédique. Nous utilisons une transformation de distances permettant d'avoir rapidement une image de distance 3D composée d'iso-surfaces constituées de voxels situés à une distance égale de la surface de référence (donc l'image elle-même). Les distances globales sont calculées par propagation des distances locales c'est-à-dire, à partir de la connaissance des distances au voisinage d'un voxel.

5.4.2 Paramétrisation

Les dimensions du voxel sont fonction des paramètres L, H et P. Les valeurs de L et H dépendent de la période d'échantillonnage spatial des données. Le paramètre P dépend du poids des niveaux de gris par rapport aux distorsions spatiales (Fig. 5.8).

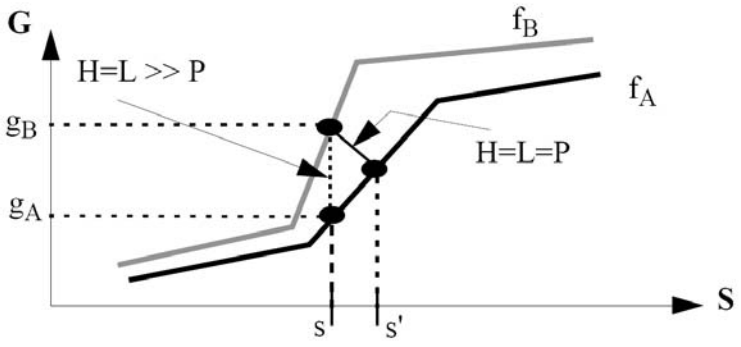


FIG. 5.8: Distance entre deux images.

Premier cas : $L, H \gg P$

Le poids des déplacements spatiaux est très supérieur aux déplacements selon l'axe des niveaux de gris. Le chemin minimal sera alors vertical. Nous avons, alors :

$$d_{AB}(v) \approx |f_B(s) - f_A(s)| \tag{5.12}$$

Seule la différence entre les niveaux de gris sera prise en compte dans le calcul de la dissimilarité locale.

Deuxième cas : $L, H \ll P$

A l'opposé, si le poids des déplacements selon l'axe des niveaux de gris est très grand devant les déplacements spatiaux alors, le chemin minimal sera horizontal et égal au déplacement entre le point $s = (x, y)$ et le point $s^{'} = (x^{'}, y^{'})$ dans le plan de niveau de gris g. Nous avons, alors :

$$d_{AB}(v) \approx \frac{1}{2}[d_A(s, f_B(s)) - d_B(s^{'}, f_A(s^{'}))] \tag{5.13}$$

Seules les distorsions géométriques seront prises en compte dans le calcul de la dissimilarité locale. Ces distorsions seront visibles dans l'image des distances. La démonstration de ces résultats se trouvent dans [Coquin 01a]. Par rapport aux méthodes de Wilson [Wilson 97] ou de Zamperoni [Zamperoni 96], le fait de considérer des voxels parallélépipédiques permet ***d'ajuster les poids respectifs des écarts spatiaux et en amplitude***. Cette technique permet de dissocier les déformations de structure par rapport aux déformations en niveaux de gris. Cette remarque est vraiment essentielle et va nous permettre d'analyser les traitements d'images.

5.5 Comparaison d'images en niveaux de gris

Pour bien comprendre ce que nous venons de dire, nous l'illustrerons sur des exemples simples. Nous comparerons notre dissimilarité par rapport à d'autres mesures en étudiant l'effet d'une augmentation uniforme du niveau de gris sur l'image de référence, puis l'effet d'un déplacement spatial d'un objet présent dans l'image. Un deuxième exemple permettra de voir les déformations géométriques dues à un cycle de compression/décompression sur une image. Enfin, dans un troisième exemple, nous comparerons des filtres appliqués sur une image synthétique bruitée par un bruit exponentiel. La comparaison est effectuée à l'aide des cartes de distances, entre les images filtrées et l'image initiale et ensuite, quantitativement, avec les valeurs de la dissimilarité entre les images filtrées et l'image initiale.

5.5.1 Augmentation uniforme du niveau de gris

Dans cette partie, nous présentons des résultats expérimentaux appliqués à des images en niveaux de gris permettant de comparer :

- l'erreur quadratique moyenne RMS,
- notre extension aux images en niveaux de gris de la dissimilarité de Baddeley : $\mathcal{D}$,
- la mesure de dissimilarité développée par Zamperoni et Starovoitov : D_z [Zamperoni 96],
- la mesure de dissimilarité développée par Wilson et al. Δ_g [Wilson 97],
- la mesure MHD (Modified Hausdorff Distance) développée par Dubuisson et Jain [Dubuisson 94], étendue aux images en niveaux de gris.

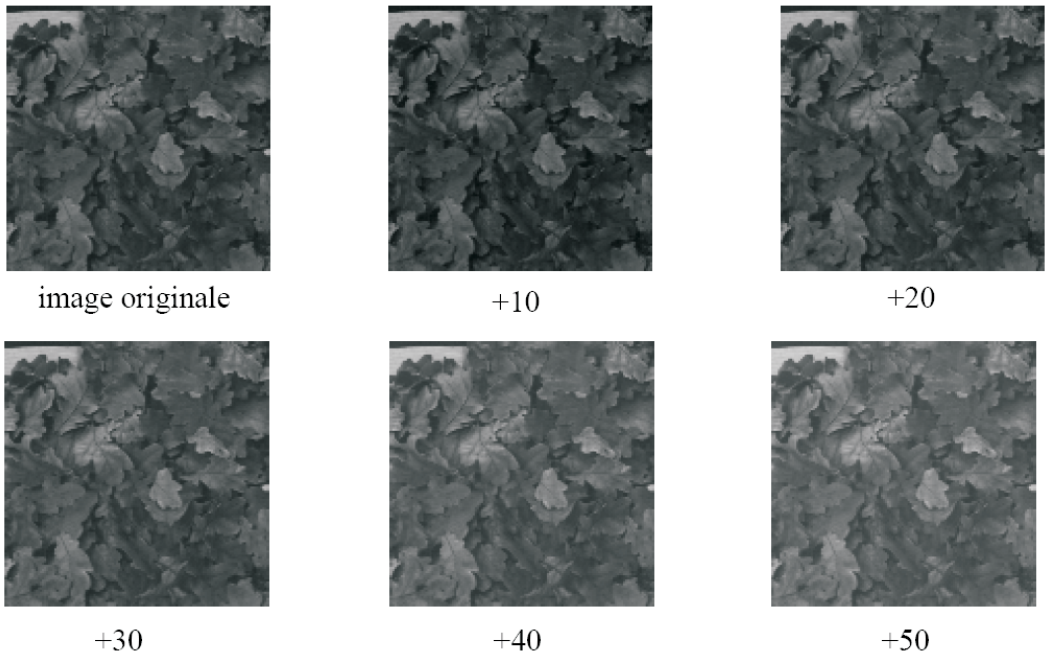


FIG. 5.9: Image originale et images translatées de 10 niveaux de gris en 10 niveau de gris

Nous comparons les réponses de ces 5 dissimilarités. Dans tous les tests qui suivent, nous avons utilisé le rapport $P/H = 1$ pour le calcul de la dissimilarité $\mathcal{D}$ avec un masque de chanfrein 3x3x3 et avec la distance $d_{16-23-28}$ pour un facteur de normalisation $N = 17$. D_z est calculée avec une taille de fenêtre $W = 21 \times 21$ et la distance d_{city}. Pour toutes les dissimilarités, nous avons pris l'exposant $q = 2$.

Le premier exemple considéré ici est une augmentation de 10 niveaux de gris par pas de 10, d'une image. Les résultats sont obtenus en comparant chaque image modifiée par rapport à l'image originale (Fig. 5.9).

On peut remarquer dans la figure 5.10 que la réponse RMS est linéaire, tout comme la réponse de notre extension de la distance de Baddeley $\mathcal{D}$. Par contre, la réponse de Wilson-Baddeley Δ_g ainsi que les réponses de Zamperoni-Starovoitov D_z et la méthode de Hausdorff modifiée MHD ne sont pas linéaires.

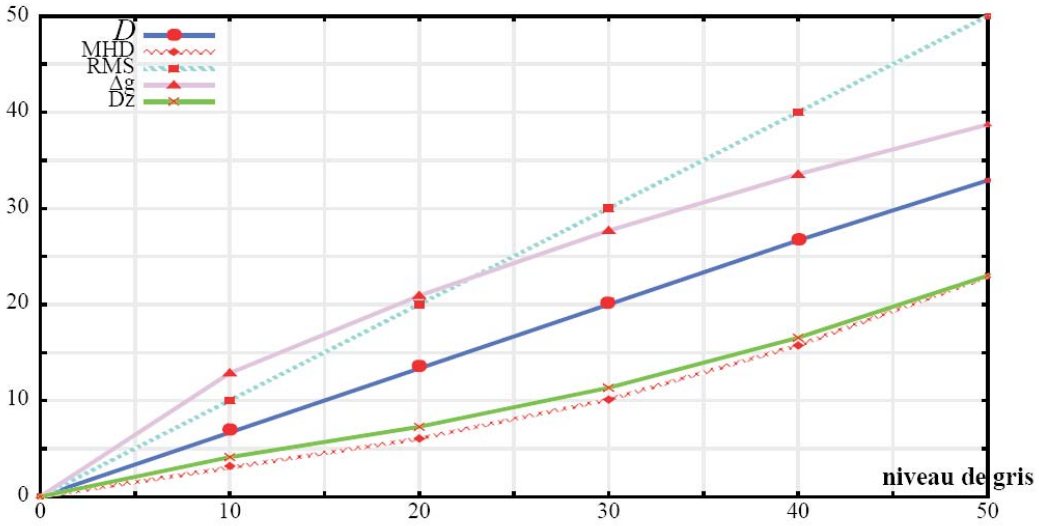


FIG. 5.10: Mesures de dissimilarité par rapport à une augmentation uniforme du niveau de gris

On notera que pour une augmentation uniforme du niveau de gris, il est préférable que la réponse de l'opérateur de dissimilarité soit linéaire. En effet, les conditions de luminosité lors de l'acquisition d'une image peuvent changer entre deux prises de vues. Avec un opérateur qui possède une réponse linéaire, la comparaison entre les images sera plus facile à étudier.

5.5.2 Effet d'un déplacement spatial

La figure 5.11 nous montre l'image originale qui contient un objet. Celui-ci a été translatée horizontalement sur la gauche, de 4 colonnes en 4 colonnes, provoquant ainsi un déplacement spatial. L'image originale est issue de la base élaborée par P. Gros [2].

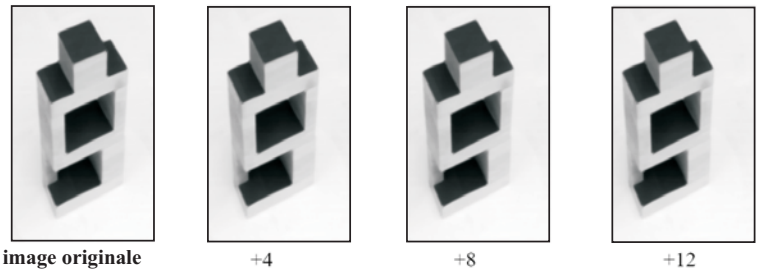


FIG. 5.11: Image originale, et images translatées

Nous avons étudié l'effet de ce déplacement spatial linéaire sur notre dissimilarité $\mathcal{D}$, la mesure

[2]images disponibles sur http ://www.irisa.fr/texmex/ressources/bases/base_images_movi/index.html

du RMS et la dissimilarité de Wilson Δ_g.

Nous pouvons remarquer que la distance $\mathcal{D}(A, B)$ est un critère global qui ne dépend pas seulement des caractéristiques des 2 images A et B. En effet, $\mathcal{D}(A, B)$ dépend également du nombre de niveaux de gris sur lequel elle est calculée puisque $Card(V) = Card(G).Card(S)$. Pour faciliter la comparaison, nous allons utiliser la dissimilarité normalisée définie par :

$$\mathcal{D}_N(A, B) = \frac{\mathcal{D}(A, B)}{\mathcal{D}(Blanc, Noir)} \tag{5.14}$$

où $\mathcal{D}(Blanc, Noir)$ est la dissimilarité entre une image de niveau de gris 255 (blanc) et d'une image de niveau de gris 0 (noir).

La figure 5.12 nous montre la dissimilarité normalisée entre l'image de référence et l'image résultante (après le déplacement horizontal) en fonction du déplacement.

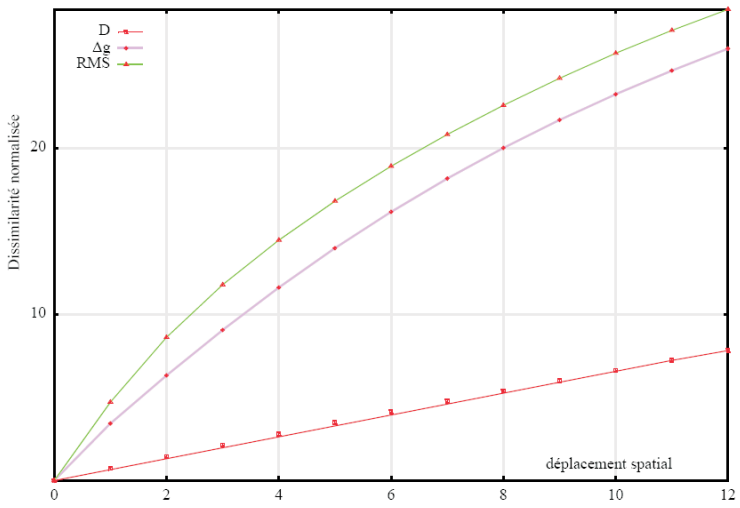


FIG. 5.12: Mesures de dissimilarité par rapport au déplacement spatial

Nous pouvons voir que seules les variations de notre dissimilarité $\mathcal{D}$ sont linéaires. Ce n'est pas le cas de la mesure RMS et de la dissimilarité de Wilson Δ_g.

5.5.3 Compression/décompression

Dans ce paragraphe, nous allons étudier l'influence du taux de compression sur les mesures de dissimilarité. L'algorithme de compression utilisé pour réaliser ces expériences est celui du programme XV[3]. Il permet de passer d'une image originale à l'image compressée au format JPEG. L'image originale est compressée avec des taux de compression croissants. Plus le taux de compression augmente et plus les distorsions sont visibles à l'œil sous forme de pavé. L'image originale est comparée à chaque version de l'image subissant un cycle de compression/décompression avec un taux de compression qui augmente (Fig. 5.13).

La figure 5.14 illustre le résultat de l'image "muscle" pour un facteur de qualité de 80%. La figure est composée de 6 images :

- l'image (a) est l'image compressée/décompressée avec un facteur de qualité de 100% (correspond à un facteur de compression de 94%). Ce sera l'image de référence.
- l'image (b) montre l'image de test. C'est l'image (a) après compression et décompression avec un facteur de qualité de 80%.

[3]Logiciel de visualisation d'image, J. Bradley, version 3.0, 1994

- l'image (c) montre la **carte de différence** d'intensité (pixel à pixel) entre l'image (a) et l'image (b) $L = H = 1$ et $P = 0$.
- l'image (d) montre la carte de distance $L = H = 1$ et $P = 0.1$ entre l'image (a) et l'image (b).
- l'image (e) montre la carte de distance $L = H = P = 1$ entre l'image (a) et l'image (b).
- l'image (f) montre la carte de distance $L = H = 1$ et $P = 10$ entre l'image (a) et l'image (b).

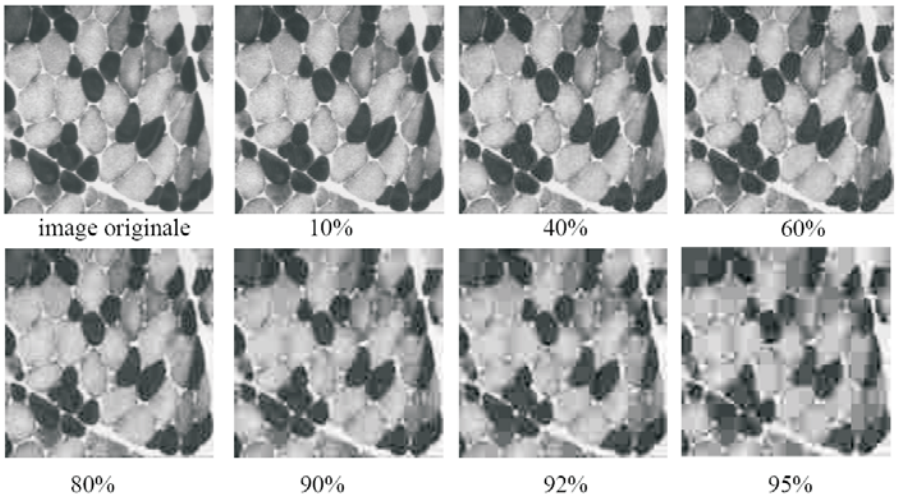


FIG. 5.13: Image originale et images ayant subi un cycle de compression/décompression, avec un taux de compression croissant.

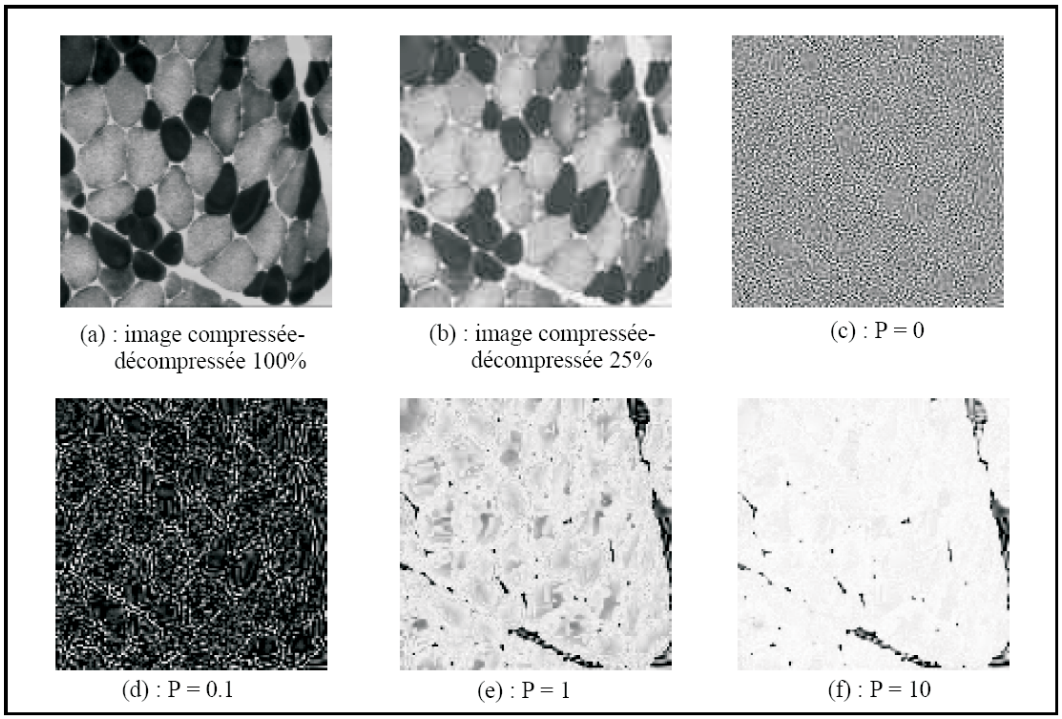


FIG. 5.14: Test pour une qualité de compression de 80%. Les images (c) à (f) sont présentées en inverse vidéo.

La figure 5.14c montre que les différences d'intensité sont distribuées de façon quasi-stationnaire. Par contre, les cartes de distances données pour les mêmes figures montrent que la distance est plus

forte dans les zones claires. D'autre part, l'influence de la dimension du voxel et, plus précisément la valeur de P par rapport à L et H, se voit nettement sur les cartes de distances (d), (e) et (f). Quand P devient très faible devant L et H, on retrouve l'information de différence d'intensité.

En faisant varier P,on peut changer la résolution de l'analyse des différences entre deux images et passer d'un analyse ponctuelle à une analyse plus large. La valeur de la dissimilarité $\mathcal{D}$ constitue un critère qualitatif global. Avec les cartes de distances, l'analyse devient plus fine voire locale.

5.5.4 Comparaison de filtrages

La figure 5.15a montre l'image de référence. C'est une image synthétique contenant deux niveaux de gris. Elle représente un contour net orienté suivant l'angle de 25 degrés. L'image bruitée est représentée par la figure 5.15b, le bruit étant de type exponentiel avec un écart-type de 10 par rapport au $Signal/Bruit = 3$.

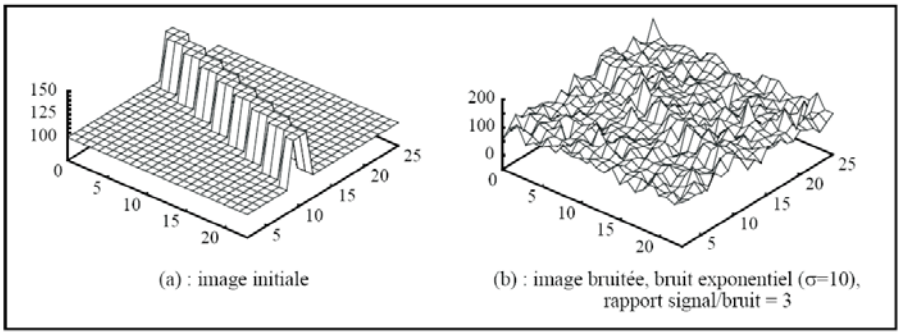


FIG. 5.15: Test sur une image synthétique

L'image bruitée est ensuite filtrée par 4 filtres :
- linéaire gaussien de taille 7×7 ($\sigma = 2.25$) (Fig. 5.16a),
- moyenneur de taille 7×7 (Fig. 5.16b),
- médian de taille 7×7 (Fig. 5.16c),
- adaptatif directionnel pondéré (d_α) de taille 7×7 (Fig. 5.16d).

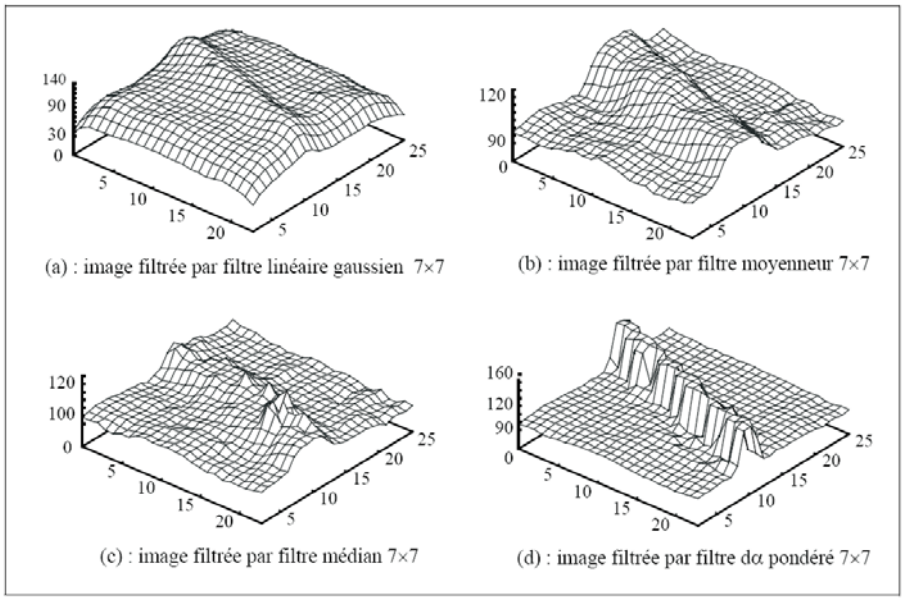


FIG. 5.16: L'image 5.15b filtrée par 4 filtres

Cartes de distances : Les cartes de distances entre l'image filtrée et l'image de référence sont aussi calculées pour ces quatre images (voir figure : 5.17). A partir des cartes de distances, on remarque que les quatre filtres produisent une erreur maximale sur le contour. Le filtre linéaire produit aussi une erreur importante sur les bords de l'image. Cela provient de l'effet de bord. De plus, la distance n'est pas la même sur tout le contour alors que celui-ci est identique dans l'image de référence (amplitude constante). On peut donc constater que cette différence est due aux impulsions de bruit distribuées arbitrairement dans l'image et bien sûr, sur les contours. L'étude des cartes de distance met en évidence un autre phénomène : après un traitement par filtre moyenneur, la distance est forte au niveau de la transition puis, elle diminue régulièrement lorsqu'on s'écarte du contour.

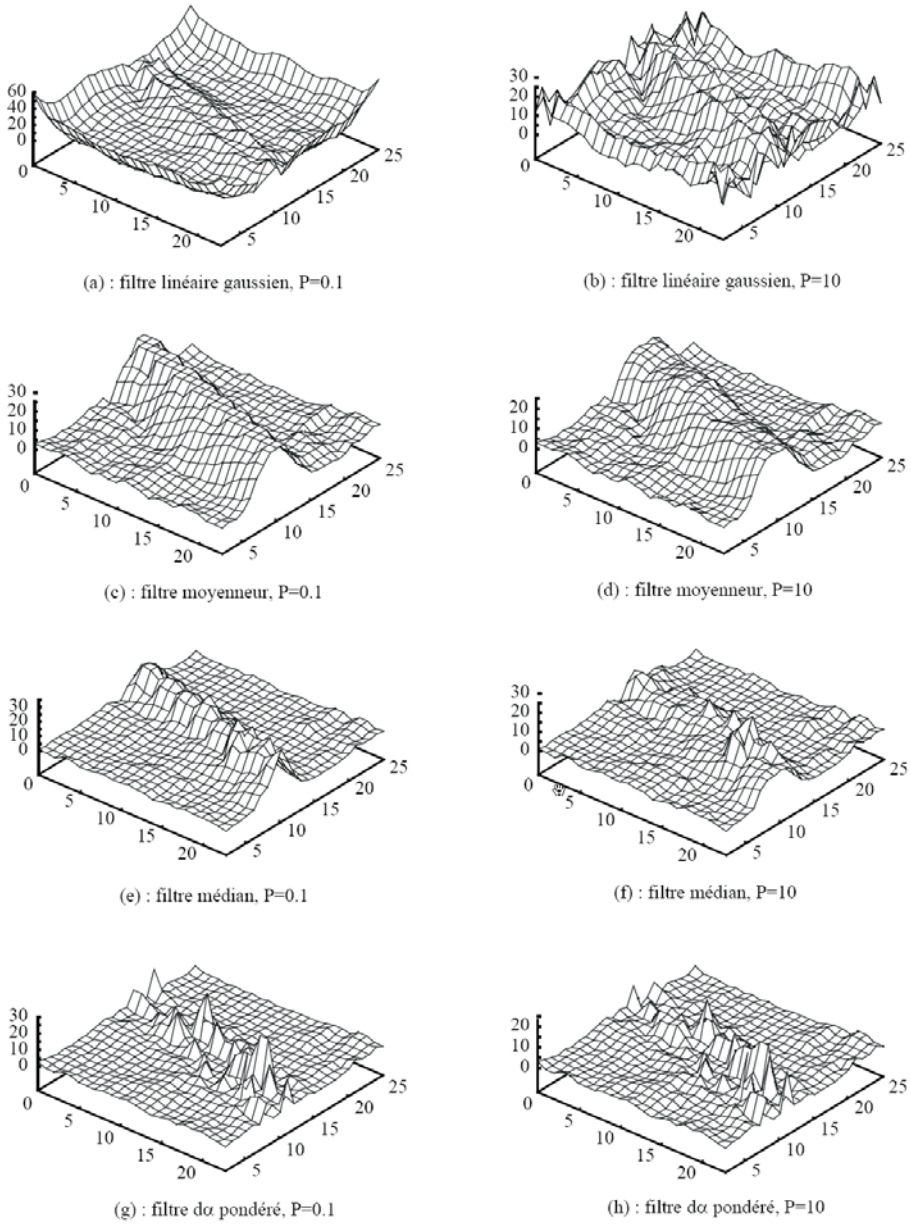


FIG. 5.17: Cartes de distances entre les images filtrées (Fig. 5.16) et l'image de référence (Fig. 5.15a), dans les cas $P = 0.1$ et $P = 10$ avec $L = H = 1$.

Ces résultats montrent que le filtre d_α pondéré [Issa 96] préserve mieux la transition. Par contre, dans les zones stationnaires, les filtres moyenneur et médian filtrent mieux les impulsions de bruit.

L'exemple du filtre médian montre bien que pour $P = 0.1$, la distance la plus grande est sur tout le contour. Par contre, lorsque l'on prend $P = 10$, la distance diminue sur la plupart du contour et seules quelques impulsions restent. Ce qui signifie que le filtre médian préserve bien la géométrie des contours, tout comme le filtre d_α pondéré.

Dissimilarité : Le tableau 5.2 donne les valeurs numériques de la dissimilarité $\mathcal{D}$ appliquée à l'exemple illustré par les figures 5.15 et 5.16, pour différentes valeurs de P. Trois cas sont considérés : $P = 0.1$, $P = 1$, et $P = 10$, tout en conservant $L = H = 1$.

Image testée	$P = 0.1$	$P = 1$	$P = 10$
image bruitée	21.3	37.0	41.9
image filtrée par un filtre linéaire gaussien 7×7	17.0	28.0	33.0
image filtrée par un filtre moyenneur 7×7	11.1	18.5	19.6
image filtrée par un filtre médian 7×7	13.0	18.6	18.6
image filtrée par un filtre d_α pondéré 7×7	3.4	4.2	7.0

TAB. 5.2: Evaluation quantitative des filtres, mesure de dissimilarité $\mathcal{D}$ dans les cas $L = H \ll P$, $L = H = P$ et $L = H \gg P$.

Nous remarquons que c'est le filtre adaptatif directionnel pondéré qui a la plus petite valeur de $\mathcal{D}$, aussi bien lorsque $L = H \ll P$ que lorsque $L = H \gg P$. Lorsque $L = H \gg P$ (prépondérance des erreurs d'intensité), le filtre moyenneur devient meilleur que le filtre médian. Ce résultat s'inverse lorsque $L = H \ll P$. Ceci peut s'interpréter par la meilleure préservation des formes par le filtre médian. Il faut noter que les quatre images filtrées ont une mesure de dissimilarité plus faible que celle de l'image bruitée et ceci, quel que soit le maillage considéré.

Dans [Coquin 01b], nous analysons le comportement de notre dissimilarité $\mathcal{D}$ face à 4 dissimilarités vis-à-vis de la sensibilité au bruit et de la variation de la forme. Nous montrons que notre dissimilarité est stable vis-à-vis du bruit dans l'image, tolérant vis-à-vis de la variation des formes et qu'elle permet de discriminer les performances des différents filtres contrairement aux autres dissimilarités qui pèchent dans certaines situations.

Des travaux récents sur des transformées de distances dans un espace courbe sont présentés dans [Ikonen 05, Ikonen 07], utilisant des distances discrètes 2D et 3D. Ces transformées permettent de calculer des distances entre pixels sur des images en niveaux de gris, et sont utilisées pour mesurer la rugosité d'une surface, par exemple. Ces transformées sont proches de la distance géodésique ou de la ligne de partage des eaux [Soille 99].

5.6 Comparaison de signatures d'images

Nous avons étudié le comportement de la dissimilarité $\mathcal{D}$ vis-à-vis de différentes déformations que peut subir une image comme :

- **(i)** une augmentation du niveau de gris moyen de l'image,
- **(ii)** un déplacement spatial de l'objet,
- **(iii)** la combinaison de ces deux déformations (i) + (ii).

Nous avons montré que la dissimilarité $\mathcal{D}$ a un comportement linéaire par rapport à l'amplitude de chacune de ces déformations et ce, pour une plage de variation relativement large [Coquin 97].

Nous avons mentionné que l'opérateur local de distance 3D dépendait des paramètres liés aux dimensions L, H et P du voxel. L'allure de la variation de $\mathcal{D}$ par rapport au paramètre P/H peut

être considérée comme la signature de la déformation entre les 2 images et peut donc servir à la **comparaison** des images en niveaux de gris.

5.6.1 Implémentation

Nous donnons la valeur des distances élémentaires D_{ijk} qui seront utilisées lors du calcul des signatures (Tab. 5.3) pour différentes valeurs du rapport P/H. Les distances sont calculées pour un coefficient de normalisation $N = 17$.

P/H	D_{100}	D_{110}	D_{001}	D_{011}	D_{111}
0.1	141	199	14	142	200
0.5	30	43	15	34	46
1	16	23	16	23	28
2	15	22	31	35	38
3	15	21	45	48	50
4	15	21	59	61	63
5	15	21	73	75	76

TAB. 5.3: Distances élémentaires en fonction de P/H.

5.6.2 Augmentation du niveau de gris moyen et déplacement spatial

Nous nous limitons ici à des déformations simples telles que l'augmentation du niveau de gris moyen de l'image et la translation spatiale.

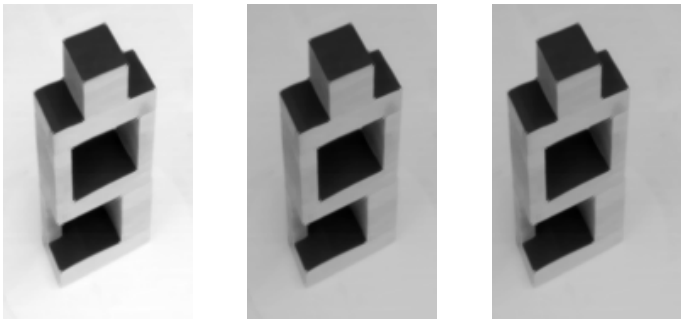


FIG. 5.18: Image de référence et images ayant subi des déformations

La figure 5.18 montre l'image de référence A et 2 images déformées $B1$ et $B2$ correspondant successivement à une augmentation du niveau de gris (+20) puis à un déplacement horizontal vers la gauche de 12 pixels. Les figures 5.19 et 5.21 montrent les signatures des déformations.

La figure 5.19 correspond à une augmentation du niveau de gris moyen de l'image A (+10,+20 ou +30).

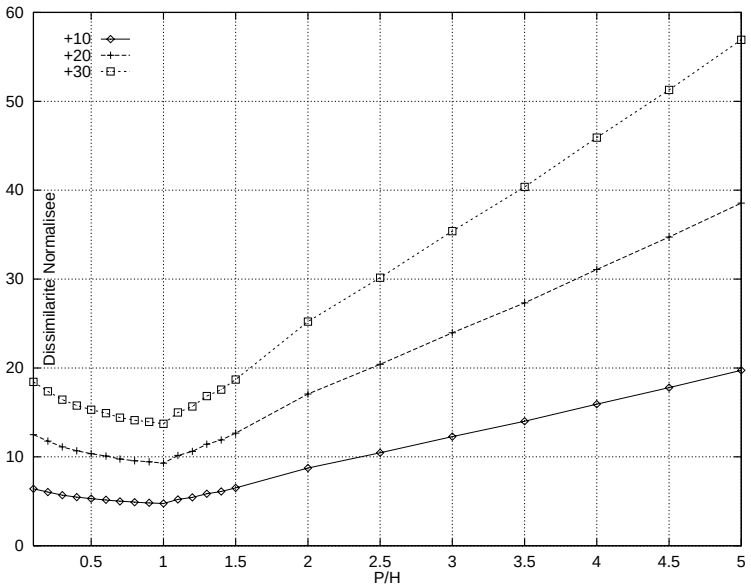


FIG. 5.19: Signature d'une variation du niveau de gris moyen

La signature de la distance normalisée est une courbe en forme de L inversé (Fig. 5.19). Pour les valeurs faibles de P/H, la distance $\mathcal{D}_{\mathcal{N}}$ tend vers la valeur efficace de la différence entre les 2 images (RMS) [Coquin 97]. Inversement, pour P/H élevé, la distance normalisée augmente linéairement avec P/H. En effet, pour la plupart des voxels, la différence des distances est égale à la variation d'intensité (notée h en Fig. 5.20) dont le poids est proportionnel à P.

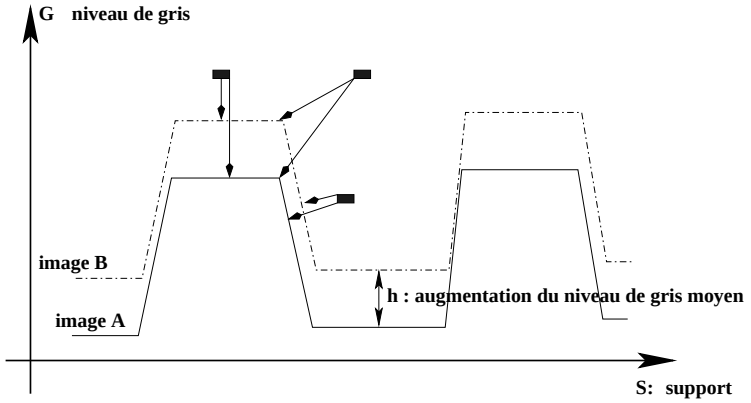


FIG. 5.20: Variation d'intensité entre 2 images

La figure 5.21 correspond à la signature d'un déplacement spatial horizontal de l'objet de 5 et de 10 pixels. Une translation se traduit par une signature en forme de L (Fig. 5.21).

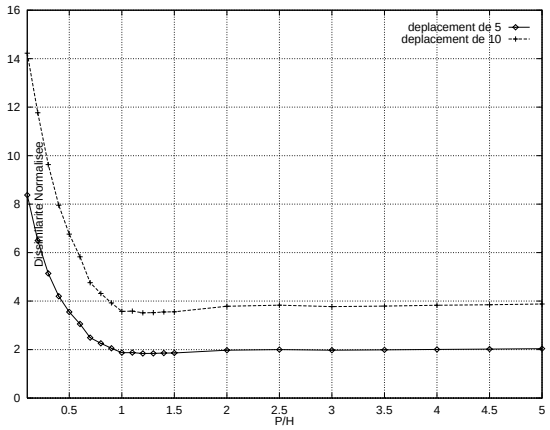


FIG. 5.21: Signature d'un déplacement spatial

Pour un rapport P/H supérieur à 1, les courbes de dissimilarité sont pratiquement constantes. En effet, pour les fortes valeurs de P/H, la distance entre un voxel et les surfaces de référence correspond au plus court chemin horizontal (Fig. 5.22) [Coquin 01a]. La différence des distances aux surfaces A et B tend alors vers l'amplitude du déplacement spatial d entre les courbes. Ce résultat reste valable si le contraste de l'objet est supérieur à l'amplitude du déplacement.

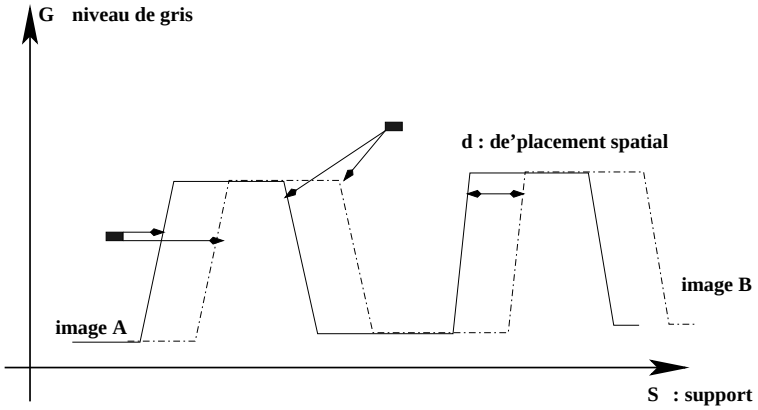


FIG. 5.22: Déplacement spatial entre 2 images

La figure 5.23 montre les signatures correspondant à la combinaison des deux types précédents de déformations.

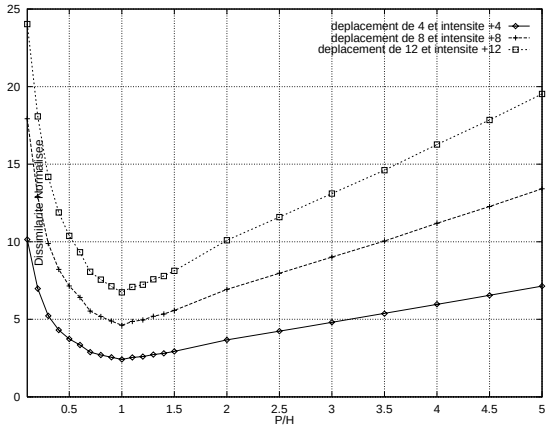


FIG. 5.23: Signature de la combinaison d'un déplacement et d'une augmentation de niveaux de gris

La combinaison d'une variation du niveau de gris et d'un déplacement horizontal se traduit par une signature en forme de **V** (Fig. 5.23). Ceci est dû au comportement linéaire de l'opérateur de distances entre images, étudié dans [Coquin 01a]. Il est à noter qu'un comportement similaire est obtenu pour des images de nature très différentes (scènes d'extérieur, objets polyédriques, ...).

5.6.3 Estimation des paramètres de la déformation

L'analyse de l'histogramme de la différence des images de distances $\mid d_A(v) - d_B(v) \mid$ fournit des informations sur l'amplitude de la déformation.

La figure 5.24 présente l'histogramme de la différence des images de distances lors d'une variation d'intensité.

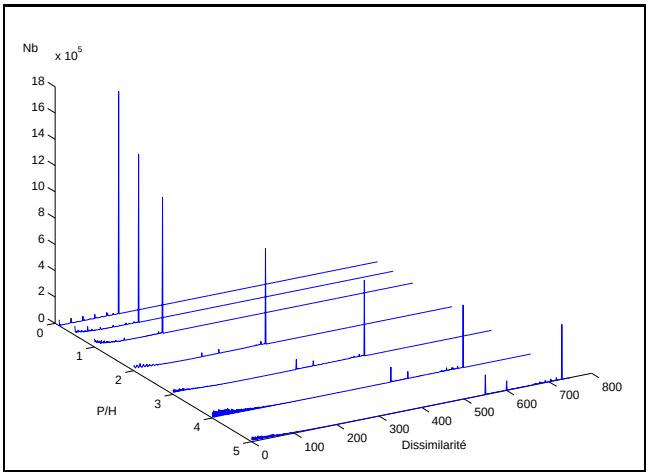


FIG. 5.24: Histogramme des distances pour une variation d'intensité

Pour les valeurs faibles de P/H, la distance entre un voxel et la surface de référence tend vers la différence des niveaux de gris. On retrouve donc dans l'histogramme, un pic dominant correspondant à l'amplitude de la variation d'intensité. Quand P/H augmente, les trajets optimaux entre voxels et surface de référence tendent à devenir horizontaux. Il y a donc dispersion de la différence des distances, essentiellement en fonction de la forme de l'objet.

La figure 5.25 présente l'histogramme de la différence des images de distance lors d'un déplacement horizontal. Pour P/H faible, il y a une faible dispersion liée au contraste d'intensité à l'intérieur de l'objet. Pour P/H fort, on retrouve, pour les mêmes raisons que ci-dessus, un pic dominant correspondant à l'amplitude du déplacement.

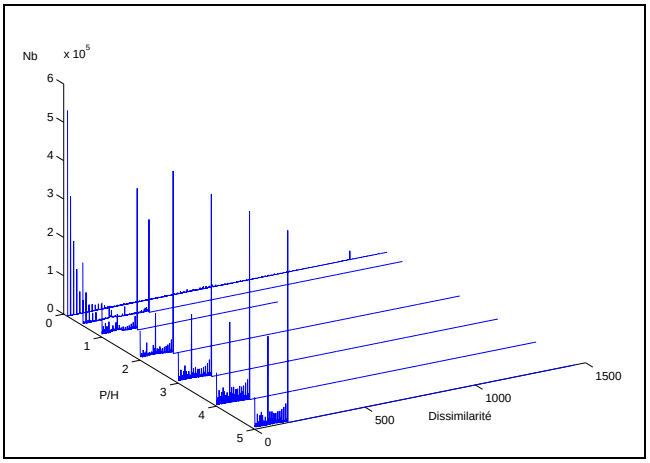


FIG. 5.25: Histogramme des distances pour un déplacement horizontal

La figure 5.26 correspond à la combinaison des 2 types de déformations.

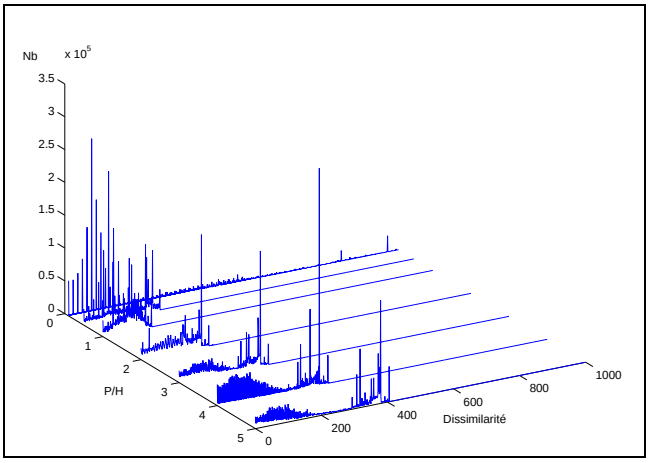
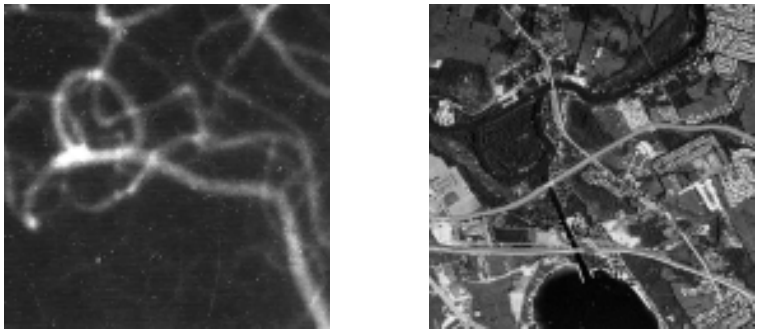


FIG. 5.26: Histogramme des distances pour un déplacement horizontal combiné à une augmentation du niveau de gris

La figure 5.26 présente l'histogramme de la différence des images de distances lors de la combinaison des 2 types de déformations. Pour P/H faible, on retrouve un pic correspondant à la variation d'intensité à condition que les variations d'intensité de l'objet soient suffisamment basses fréquences. Pour P/H élevé, le pic significatif le plus éloigné de l'origine correspond à la combinaison du déplacement spatial et de la variation d'intensité.

La figure 5.28 présente l'histogramme de la différence des images de distances pour les deux images de la figure 5.27. Lorsque les deux images sont totalement différentes, c'est à dire que la deuxième ne résulte pas d'une déformation simple de la première (Fig. 5.27), l'histogramme des différences de distances est très dispersé et ne présente pas de pics caractéristiques.



a) b)

FIG. 5.27: a) angiographie et b) vue aérienne

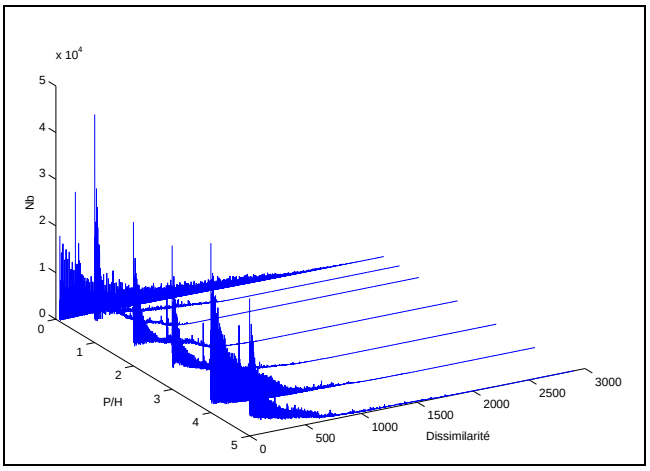


FIG. 5.28: Histogramme des distances pour deux images différentes

Discussion : Nous avons proposé une nouvelle méthode permettant de caractériser par leur signature, différentes déformations que peut subir une image en niveaux de gris. L'analyse de l'histogramme de la différence des images de distances fournit des informations sur l'amplitude de la déformation, tant en niveau de gris qu'en déplacement spatial de l'image. Par rapport aux méthodes de corrélation, cette nouvelle technique est moins perturbée par les variations d'intensité et s'avère moins limitée par la taille des fenêtres ou l'amplitude des vecteurs déplacements.

5.7 Comparaison des images couleurs

Nous nous sommes également intéressés à la comparaison des images couleurs. Étant donné la représentation des images couleur, la première méthode consiste à étendre en 5D la méthode développée pour les images en niveaux de gris (utilisation d'un opérateur local de distances à 5 dimensions (x,y,R,V,B)). C'est la meilleure façon de tenir compte de la nature vectorielle des pixels des images couleur. L'inconvénient de cette méthode est la taille des données et le nombre d'opérations lié à chaque point de l'image pour le calcul de l'image de distances. Nous avons préféré indexer les couleurs pour construire une palette de 256 couleurs, et utiliser la méthode des images en niveaux de gris sur les images indexées.

Pour indexer les couleurs, nous avons utilisé un réseau de neurones de Kohonen connu également sous le non de SOM (Self Organizing Map). La figure 5.29 nous présente une image ayant subi cette indexation des couleurs.

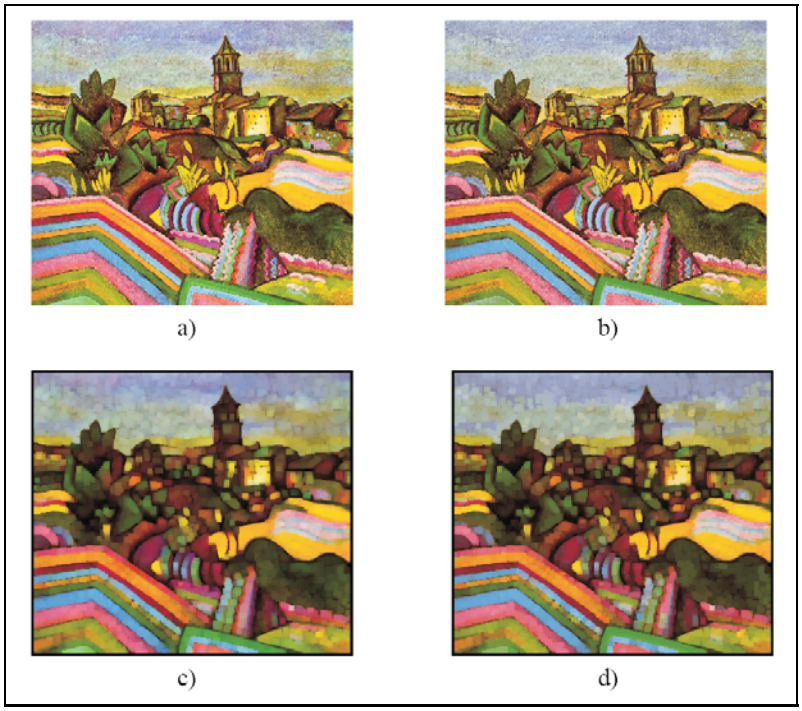


FIG. 5.29: Résultat de l'indexation des couleurs. a) image miro originale avec une représentation sur 24 bits, b) image miro indexée, c) image (a) obtenue par une ouverture morphologique, d) image (c) indexée

Cette méthode d'indexation des couleurs nous permet de passer de l'espace 5D à l'espace 3D. Cependant, contrairement aux cas d'images en niveaux de gris où les niveaux sont équirépartis entre 0 et 255 (deux niveaux successifs sont distants de 1), l'indexation des couleurs nous donne des couches successives d'épaisseurs différentes. Nous ne pouvons plus utiliser le principe de calcul des images de distances décrit précédemment puisque les niveaux ne sont pas équidistants. C'est pourquoi, nous avons développé un opérateur de distances non-stationnaire (voir 4.5.3).

5.7.1 Influence du déplacement spatial

L'image originale a été déplacée horizontalement de 5 pixels en 5 pixels, sur la gauche. La figure 5.30 permet de comparer le résultat de 4 mesures de dissimilarité :

- notre mesure de dissimilarité en fonction du déplacement notée D_{adapt} qui utilise un opérateur local de distances non-stationnaire (qui s'adapte à l'épaisseur des couches),
- notre dissimilarité utilisant un opérateur local 3D notée D_{nadapt} (sans adaptation entre les couches),
- la mesure du RMS calculée sur l'image RVB RMS_{RGB},
- la mesure du RMS calculée sur l'image indexée RMS_{index}

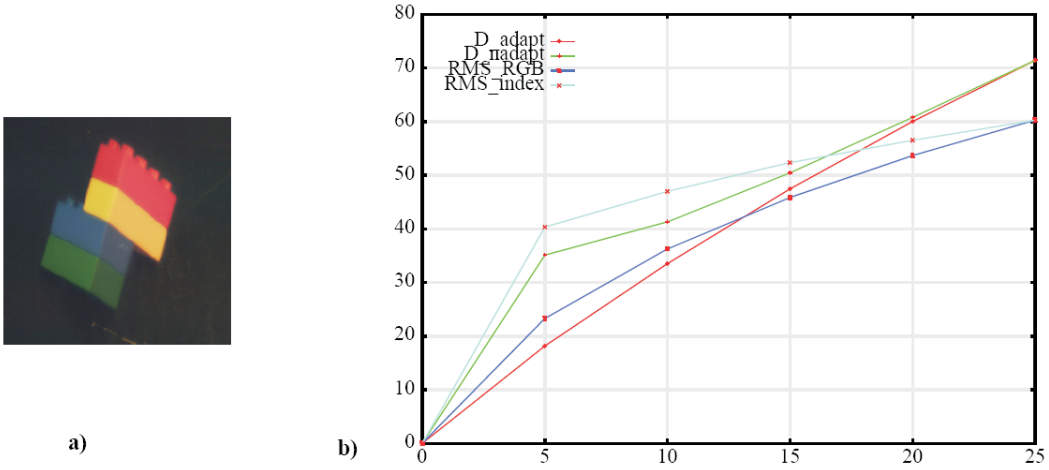


FIG. 5.30: a) Image originale et b) Dissimilarités $\mathcal{D}$ et RMS entre l'image déplacée et l'image originale

La figure 5.30 montre la linéarité de notre mesure de dissimilarité en fonction du déplacement.

Puisque l'échelle des couleurs n'est pas uniforme, différentes approches peuvent être envisagées pour calculer la dissimilarité (A, B) entre les deux images de couleurs indexées :

- la première approche consiste à utiliser un opérateur local de distances 3D non stationnaire, en adaptant la profondeur des couches à la taille du voxel P_r. Cette dissimilarité est notée D_1.
- la deuxième approche consiste à utiliser un opérateur local de distance 3D avec un pas d'échantillonnage constant P. Dans ce cas, la profondeur P est égale à la plus petite des profondeurs des différentes couches. Nous devons donc sur-échantillonner les images de couleurs indexées. Cette dissimilarité est notée D_2.
- la troisième approche consiste à utiliser un opérateur local de distances 3D avec un pas d'échantillonnage constant $\overline{P_r}$ (la moyenne de la profondeur de l'ensemble des couches). Cette dissimilarité est notée D_3.

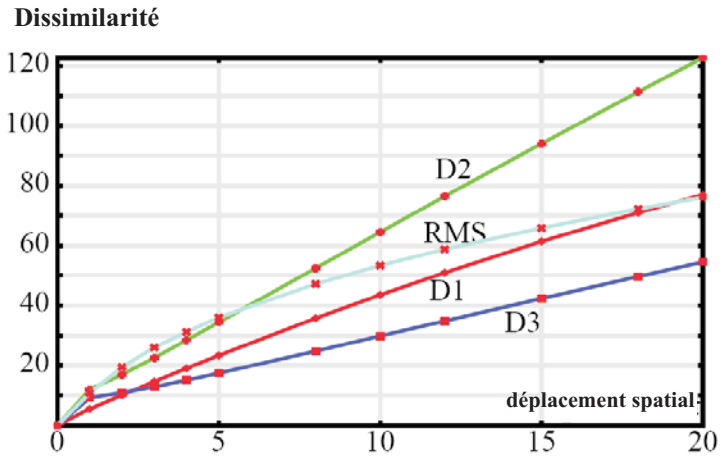


FIG. 5.31: Dissimilarités globales en fonction d'un déplacement spatial

Ces différentes dissimilarités sont représentées dans la figure 5.31. Nous avons également ajouté la mesure du RMS. Les dissimilarités D_1 et D_2 ont un comportement similaire par rapport au déplacement spatial d'un objet (courbes quasi-linéaires). Par contre, le nombre de voxels sur lequel il faut calculer la transformée de distances pour D_2 est 7 fois plus important que celui requis pour D_1 ou D_3. Cependant, puisque la profondeur des couches a été moyennée, nous observons une distorsion non-linéaire pour de faibles déplacements spatiaux.

5.7.2 Influence du nombre de couches

Afin de réduire la taille de l'espace mémoire et le temps de traitement, il est possible de réduire le nombre de couleurs lors de l'indexation des images. On peut voir l'effet de la réduction du nombre de couleurs dans la figure 5.32. La réduction du nombre de couleurs ne change pas le comportement de l'allure de la dissimilarité, qui reste quasi-linéaire lors d'un déplacement spatial de l'objet cube.

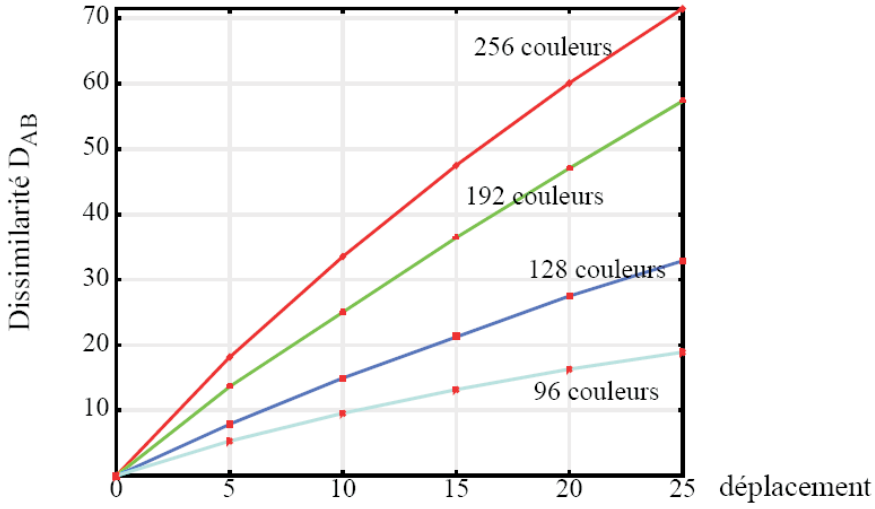


FIG. 5.32: Réduction du nombre de couches

Nous avons également étudié le comportement de ces dissimilarités dans les espaces de couleur (RVB, HSI, HSV, CIELab, CIELuv) vis-à-vis, d'une variation radiométrique, d'un déplacement spatial d'objets et d'une distorsion de forme. Cette étude a montré que la mesure de dissimilarité globale $\mathcal{D}$ dans l'espace couleur HSI a un comportement quasi-linéaire vis-à-vis des distorsions radiométriques ou spatiales. Davantage de détails sont donnés dans [Coquin 00b] et [Coquin 02b].

5.8 Conclusion

Dans ce chapitre, nous avons montré quelques applications que nous avons développées pour mettre en évidence des méthodes de **comparaison d'images** binaires, puis en niveaux de gris et enfin d'images couleurs.

Nous avons appliqué la distance de Baddeley pour reconnaître les gestes dynamiques de la main et de l'avant-bras, au moyen de la **comparaison** d'images binaires (considérées comme signatures dynamiques). Cette mesure, associée à celles issues d'un gant numérique, permet de lever des ambiguïtés et d'apporter davantage de robustesse. La démarche proposée dans [Coquin 06] permet de traiter les données complémentaires ou redondantes issues de plusieurs capteurs en associant un degré de confiance à chacune des sources et en fusionnant ces informations.

Nous avons proposé une mesure de dissimilarité permettant de **comparer** et d'analyser certains traitements associés aux images en niveaux de gris ou en couleur. La dissimilarité proposée est une extension de la distance de Baddeley. La distance est calculée à l'aide d'un opérateur local en maillage parallélépipédique. Elle permet de tenir compte simultanément des différences entre niveaux de gris ou couleurs et des éventuelles déformations géométriques des structures présentes dans l'image. Elle s'avère meilleure que les méthodes basées sur le seul critère de l'erreur quadratique moyenne. Elle peut être utilisée pour évaluer les performances de méthodes de filtrage et de compression ou les effets d'opérateurs de prétraitement d'images. La complexité de l'implantation de l'algorithme est réduite mais l'analyse d'images de grandes dimensions nécessite une taille de mémoire importante.

L'utilisation de la distance locale en remplacement de la distance euclidienne réduit la taille mémoire requise.

Nous avons vu qu'en considérant un maillage parallélépipédique, nous pouvions selon les valeurs L, H et P, privilégier, soit les déplacements spatiaux et donc mettre en évidence les erreurs et les distorsions géométriques, soit les déplacements en niveaux de gris et donc privilégier les différences d'intensité.

Nous avons également mis en évidence qu'il était possible de **comparer** et de caractériser, par leur signature, différentes déformations que peut subir une image en niveaux de gris. L'analyse de l'histogramme de la différence des images de distances fournit des informations sur l'amplitude de la déformation aussi bien en niveau de gris qu'en déplacement spatial.

Dans [Coquin 02a] nous avons proposé une nouvelle méthode pour calculer le champ des vecteurs déplacements entre deux images ou sur une séquence d'images à niveaux de gris. Cette méthode est basée sur la transformation de distance appliquée à la surface d'une image. L'image en niveaux de gris est vue comme une surface dans un espace à 3 dimensions. L'analyse de ces champs de vecteurs déplacements permet de caractériser une déformation de type déplacement spatial ou rotation. Des travaux supplémentaires pourraient être poursuivis dans ce sens pour affiner les résultats.

Chapitre 6

Analyse de séquences d'images

Résumé : *La vidéo est un média qui pose des problèmes complexes en raison du volume important de données à traiter et de la difficulté à représenter et à extraire des informations utiles pour comprendre son contenu. Nous avons travaillé sur l'analyse de séquences d'images plus particulièrement, dans le domaine des films d'animation. Nous exposons dans ce chapitre notre approche de la comparaison des séquences d'images. Nous avons décidé de comparer les séquences de manière indirecte, en comparant leurs résumés construits après la sélection d'images clés ou en comparant des attributs extraits de ces séquences. La comparaison nécessitera la définition et l'utilisation de distance.*

6.1 Introduction

Le développement du numérique - particulièrement l'apparition de standards de compression de la vidéo (les formats Mpeg) et des réseaux permettant l'échange de ces documents - a ouvert un nouveau champ d'étude aux communautés de chercheurs impliqués dans l'analyse des images et des sons.

Le passage de la forme d'enregistrement analogique à la forme numérique permet un accès plus rapide au contenu. Il devient en effet possible, en posant des index sur le flux audiovisuel numérique d'accéder directement au contenu sans avoir à parcourir séquentiellement celui-ci. Ces index permettent de pointer ou de situer le contenu et servent à la recherche d'information (l'index primordial étant le nom d'un fichier sur un disque dur d'ordinateur). Mais pour rendre vraiment exploitables les documents, le processus documentaire consiste à produire des métadonnées alignées sur le flux audiovisuel à l'aide d'index temporels (*exemple* : dans le fichier Mpeg, journal télévisé du 12/12/2000, monsieur $\mathbb{X}$ apparaît à l'image entre le temps t_1 et t_2).

Créer les métadonnées : Comment créer, représenter et exploiter ces métadonnées ? Les deux derniers points relèvent de la problématique des formats et du génie documentaire. La création des métadonnées était jusqu'à une période récente le fruit d'un travail de saisie manuelle effectué par des documentalistes. Récemment, l'application de techniques, issues de travaux de recherche effectués dans d'autres domaines que l'audiovisuel (- analyse d'images pour la reconnaissance des formes en imagerie satellitaire, médicale, vision par ordinateur et ses applications à la robotique, vidéo surveillance, etc. - ainsi que tous les travaux portant sur le flux sonore : reconnaissance de locuteur, dictée vocale) a permis la création automatique de métadonnées représentant le contenu du document. L'indexation des contenus vidéo peut être effectuée à différents niveaux d'abstraction élevés, comme le nom du programme et son sujet, ou à des niveaux plus basiques portant sur le contenu, comme la localisation des changements de plan ou la caractérisation en quantité de mouvement d'une séquence vidéo. L'automatisation du processus d'indexation devient inévitable lorsque la granularité de l'accès au contenu augmente. L'indexation automatique des flux audiovisuels est alors capable de nous renseigner sur la nature physique du document : ce que l'on voit (un visage) et ce que l'on entend (une explosion) mais ne permet pas d'obtenir des informations d'un niveau d'abstraction élevé.

Segmenter les programmes : L'indexation textuelle utilise les mots et les expressions comme

pointeurs vers les phrases, les paragraphes et les pages des documents. De manière similaire, indexer le matériel audiovisuel requiert la sélection d'images clés et de séquences comme pointeurs vers des unités de plus haut niveau comme les ***scènes*** grâce à la prise en compte des règles de la production cinématographique [Joly 05]. L'unité de base après l'image est le ***plan*** : une séquence audiovisuelle enregistrée de manière continue. Des plans successifs peuvent êtres assemblés de diverses manières, la plus simple étant la transition franche ou cut, caractérisée par un changement brutal d'une image à la suivante. Des transitions plus sophistiquées existent comme les fondus, les volets et les masquages. La détection des ***cuts*** est généralement opérée grâce à des méthodes de différence d'images basées sur la **comparaison** pixel à pixel ou sur **la comparaison** des histogrammes des couleurs sur les images entières (ou sur un ensemble de sous-régions de l'image) [Del Bimbo 99], et [Smeulders 00a].

Pour la détection automatique de plans [Hanjalic 97] et [Quénot 99], beaucoup d'algorithmes se fondent sur l'utilisation des **comparaisons** d'histogrammes (distribution de l'intensité, de la couleur, etc.) car la nature globale des histogrammes les rend moins sensibles aux changements typiques dans un plan, réduisant de ce fait le nombre de fausses détections par rapport aux comparaisons pixel à pixel. Les transitions progressives sont une classe importante de transitions, incluant le fondu enchaîné, le fondu à l'ouverture, le fondu au noir, les masquages et les volets. La présence des effets spécifiques peut alors être utilisée comme un indice pour détecter des changements de scène, par opposition à la détection simple de plans. L'évaluation et la comparaison des résultats peuvent être effectuées à l'aide d'une vérité terrain. En effet, il est possible à des experts du domaine de se mettre d'accord sur une segmentation idéale en plans et d'autoriser ainsi une comparaison des résultats des différents algorithmes. Les performances des algorithmes sont bonnes pour les cuts, ce qui n'est pas le cas pour les transitions progressives, particulièrement les fondus.

La détection des limites de plan est, comme on vient de le voir, une étape importante vers l'extraction automatisée des segments visuels significatifs. Une autre étape est de caractériser les plans en extrayant de l'information des images correspondantes. Ces informations peuvent être obtenues par l'identification d'objets dans les images (comme des vues frontales de visages), par la lecture des légendes ou par la détection des événements caractéristiques (comme des flashes lumineux). Les scènes contenant des flashes lumineux peuvent indiquer qu'il s'y produit des événements notables ou qu'y figurent des personnalités. De plus, fournir les temps d'apparition des flashes permet d'améliorer les algorithmes de segmentation en plans.

Localiser et identifier des visages dans les séquences visuelles peut fournir des informations utiles pour segmenter le flux vidéo (ainsi dans les journaux télévisés, quand les différents sujets sont présentés par un même journaliste) et pour l'indexation (en identifiant les personnes présentes dans une séquence donnée) [Smith 97] et [Viallet 02]. En outre, connaître la taille relative des visages détectés dans les images permet de classifier le plan en tant que gros plan, plan moyen, ou plan large. L'identification ou reconnaissance des visages (mettre un nom sur le visage localisé) nécessite la construction d'une base de connaissance contenant tous les visages et la mise en correspondance probabiliste entre le visage détecté et les visages présents dans la base. Là encore, des méthodes de **comparaison d'images** sont nécessaires.

Les légendes dans les programmes audiovisuels comme les journaux télévisés, les documentaires et les émissions sportives fournissent des informations importantes sur le sujet des émissions. Elles peuvent renforcer la bande sonore parlée, mettre en valeur des événements ou présenter de l'information supplémentaire. Dans les sports, elles informent sur les scores, les temps de jeux et mettent en avant des phases de jeu importantes. Il y a plusieurs approches pour localiser les légendes dans un flux visuel. Après qu'une légende a été localisée, elle peut être isolée du fond et lue par un programme de reconnaissance optique de caractères.

Choisir des images clés : Une fois que des plans ont été détectés, il est très intéressant de les récapituler en choisissant des images clés représentatives. Ceci permet d'obtenir la représentation du contenu d'un plan avec un nombre limité d'images. Différents algorithmes ont été proposés jusqu'ici, se basant sur la **comparaison d'images** pour extraire les images les plus différentes et sur des plans dans lesquels l'action est prédominante. Les producteurs emploient fréquemment les mouvements de caméra pour établir des messages complexes dans un plan simple. Les acteurs font une pause pour souligner des gestes. Ces deux observations conduisent à penser que la détection d'images clés basée

sur l'analyse du mouvement est une solution intéressante. Ces algorithmes procèdent par l'analyse du flux optique pour mesurer la quantité de mouvement dans un plan. Ensuite, ils choisissent des images clés qui correspondent aux minima locaux de mouvement, considérant ainsi qu'elles sont les plus significatives.

Afin d'obtenir une information sémantique plus précise et plus complète sur la vidéo, nous avons besoin de classifier des objets apparaissant dans une séquence audiovisuelle par des critères comme la forme, la couleur mais aussi à partir de leurs mouvements. En décrivant les mouvements, à l'aide d'analyses spécifiques, on peut alors représenter les objets dans l'espace et temporellement. Outre fournir des informations sur la trajectoire des objets, l'analyse du mouvement est également utile afin de détecter des objets en premier plan, de récupérer les mouvements de caméra (zoom, travelling, inclinaison, etc.), et pour créer des mosaïques, à partir de plusieurs images, pour la représentation résumée d'une scène.

La nature temporelle des signaux vidéo est alors employée pour faciliter la segmentation automatique des objets en termes de mouvement et pour identifier **par comparaison** des structures significatives, fournissant ainsi une représentation **résumée de la vidéo**. Des régions avec des mouvements particuliers sont identifiées itérativement en produisant des modèles de mouvement et en reliant chaque région de l'image à l'un des modèles. Ce modèle peut décrire des mouvements typiques comme la translation, la rotation, le zoom, et toute combinaison entre eux. Dans l'état actuel des techniques d'analyse d'image, il est généralement difficile d'extraire exactement les objets mobiles dans les documents audiovisuels, et encore plus difficile de les identifier. Cependant, il est possible d'estimer le mouvement des objets sans les détecter exactement. Il est alors possible de caractériser les séquences audiovisuelles en termes de quantité de mouvement pour les rechercher ultérieurement.

Préalablement à une indexation des informations contenues dans une vidéo, il est nécessaire de bâtir une représentation temporelle structurée de cette vidéo, correspondant au découpage en plans élémentaires. Il s'agit de détecter les "cuts" et les transitions progressives marquant les changements de plans. La méthode proposée dans [Peyrard 05] pour réaliser le découpage de la vidéo en plans élémentaires s'appuie sur la cohérence temporelle au sein d'un même plan d'une information liée au mouvement global dominant entre deux images successives.

Dans les documents film ou vidéo, il peut y avoir plusieurs milliers de plans par heure. Afin de parcourir rapidement ces documents, pour l'indexation et la recherche, il est nécessaire de trouver des séquences plus longues que les plans par exemple des scènes ou des unités narratives. Un ensemble de plans consécutifs est groupé dans une scène en se basant sur l'unité d'espace, de lieu, de temps et d'action. Comme la détection automatisée de scènes est actuellement à ses débuts, certains auteurs utilisent les effets de transition entre plans (tels les fondus) comme limites des changements de scène (et non seulement comme limite de plan). Une scène peut être aussi déduite quand une alternance de plans semblables est détectée (les champs/contrechamps). On peut aussi construire des hiérarchies de plans sur un critère visuel (la couleur) : en décidant d'un niveau de similarité dans cette hiérarchie, on obtient des classes de plans semblables ; la première image de chaque classe est alors souvent choisie pour être l'image représentative.

Indexer le son : Le problème de la recherche documentaire sonore est bien connu. Beaucoup de travaux tentent de trouver des voies pour localiser, classer, et parcourir automatiquement l'audio en utilisant les avancées récentes de la reconnaissance de la parole.

Une grande variété de méthodes a été utilisée pour indexer les documents sonores. Les méthodes proposées permettent la segmentation de la bande sonore en zones de parole, musique et bruit. Sur les zones de parole, on peut alors appliquer les techniques de reconnaissance automatique de la parole qui, en général, fournissent aussi une localisation temporelle des changements de locuteur. L'analyse de la parole nécessite un apprentissage sur un large corpus de texte. Elle permet d'obtenir une transcription avec de bons résultats et de retrouver les mots importants (issus d'un dictionnaire) dans la bande sonore. Enfin, l'identification des locuteurs permet de mettre un nom sur chaque nouveau locuteur. Mais ces techniques sont plus efficaces lorsqu'elles sont couplées à une segmentation en plans ou en scènes. La plupart des solutions industrielles proposées aujourd'hui (MediaSite, Mate, Virage, etc.) sont basées sur le couplage transcription automatique de la bande son et segmentation de la vidéo en scènes.

À l'INA (Institut National de l'Audiovisuel), l'approche adoptée pour l'indexation automatique des flux audiovisuels s'inspire des méthodes présentées précédemment et tend à produire des analyses conjointes de l'audio, de la vidéo et des textes d'accompagnement.

Citons également, les tâches d'indexation de la campagne TREC Vidéo (Rushes)[1] où le résumé est défini par le pourcentage de la longueur par rapport au document initial, et par la sémantique : "le résumé doit contenir tous les objets significatifs et les événements significatifs".

Pour notre part, nous avons débuté ces travaux en 2003 sur l'analyse de séquences d'images issues du CICA[2] en vue de l'indexation, pour la recherche et la navigation de ces séquences, dans une base de données. La majorité des résultats présentés ici sont issus du stage de Master de ***Laurent Ott*** effectué en 2005, et de la thèse soutenue par ***Bogdan Ionescu*** en mai 2007.

Nous présenterons dans ce chapitre les méthodes que nous avons développées pour construire automatiquement des résumés par **comparaison** des images clés extraites de la séquence. Enfin, nous terminerons en présentant une approche visuelle pour **comparer** les séquences d'images par le biais de gamuts sémantiques.

6.2 Construction d'un résumé vidéo

Comme nous venons de le mentionner, en ce qui concerne l'indexation de vidéo, on peut considérer qu'il existe trois niveaux de représentations attachés à la donnée vidéo :

- un niveau ***signal*** ou ***bas niveau*** qui s'attache à décrire les caractéristiques des segments d'une vidéo comme les couleurs, la texture, la taille, les formes reconnaissables ... Un segment peut être un plan, une scène, un épisode ou un passage de la séquence contenant un événement présentant un intérêt particulier (visuel, sonore ou textuel).
- un niveau ***structurel*** qui met en évidence une organisation hiérarchique de la vidéo en images, plans, scènes et séquences. Cette structuration est issue du monde de la production cinématographique.
- un niveau sémantique qui vise à fournir une description de ***haut niveau*** de ce que contient la vidéo, qu'il s'agissent de personnages, de lieux, ou d'actions et de leurs interactions. On cherche ici à modéliser l'"*histoire*" véhiculée par le contenu de la vidéo. La vidéo étant un média complexe, on veut également trouver à ce niveau une description des sous-titres ou/et de la bande sonore associés.

Selon le niveau de représentation ciblé, il est possible ou pas d'extraire automatiquement l'information recherchée. Ce travail de recherche et d'extraction des caractéristiques, de la structure, et de la sémantique à des fins de modélisation constitue le travail d'**indexation** de la vidéo. En règle générale, l'extraction des informations de bas niveau (couleur, texture, forme, ...) d'une image ou d'un segment d'images est un processus automatisable. La recherche de plans ou de scènes d'une vidéo a fait des progrès sensibles depuis une dizaine d'années et l'on peut parler d'un processus semi-automatisable, le recours à l'utilisateur pouvant être nécessaire pour valider le découpage ou lever les indécisions du système. Pour les informations de plus haut niveau, hormis quelques domaines particuliers dans lesquels les vidéo répondent à des schémas fixes (comme les journaux télévisés ou les reportages sportifs), il est évident que l'indexation nécessite encore aujourd'hui l'assistance d'un opérateur humain. Le résultat de l'indexation d'une vidéo est une description numérique (exploitable par une machine) de la vidéo dans un ou plusieurs formalismes qui permettent l'accès, la recherche, le filtrage, la classification et la réutilisation de tout ou partie de cette vidéo.

Une fois la description d'une vidéo établie, il est possible de créer dynamiquement un ou plusieurs **résumés** de cette vidéo. Un résumé est un extrait de la vidéo qui vise à écourter la présentation de la vidéo en en présentant les moments jugés ***essentiels***. En effet, s'il est possible de créer automatiquement un résumé par sélection "*aveugle*" ou "*aléatoire*", la génération automatique de résumés requiert l'expression de critères de préférence qui seront exploités pour la sélection des segments

[1] http ://www.-nlpir.nist.gov/projects/trecvid

[2] Cité de l'image en mouvement

d'images constituant le résumé. La problématique de la production de résumés de vidéo consiste à être capable de présenter de manière synthétique le contenu de la vidéo en préservant l'essentiel du message original [Pfeiffer 96].

La génération de résumés repose sur la connaissance de la segmentation temporelle de la séquence vidéo. Par vidéo, nous n'abordons ici que les **séquences d'images**, sans utiliser le son, ni le texte extraits de ces images. Une thèse est actuellement en cours au LISTIC sur la coopération image et texte, pour l'indexation de séquences d'images. Un résumé vidéo peut être conçu soit sous forme d'un document hypermédia composé d'un ensemble d'images représentatives du contenu de la vidéo, soit comme une séquence audio-visuelle, de durée réduite, construite en prenant des extraits jugés essentiels de la séquence originale. Nous nous sommes intéressés à ces deux approches.

Il existe deux types de **résumés de vidéo** : ceux réalisés à partir d'images fixes et ceux réalisés à partir de segments d'images qui ont chacun une intégrité sémantique [Li 01]. **Les résumés à base d'images fixes** ou **résumés en images** comme ceux proposés par Video Manga [Uchihashi 99], présentent les images clés de la vidéo, alors que les **résumés à segments d'images** ou **résumés dynamiques** rassemblent des séquences importantes de la vidéo originale. Dans la littérature, les premiers résumés sont appelés simplement résumés de vidéo (ou ***video summary***), les seconds sont appelés condensés de vidéo (ou ***video skimming***). Par la suite, pour éviter toute confusion, nous emploierons les termes résumés en images ou résumés dynamiques.

Ce domaine de recherche sur les résumés vidéo dynamiques est récent. Parmi les principaux travaux réalisés, on peut mentionner :

- le projet **Informedia**[3] qui propose une bibliothèque vidéo intelligente accessible à l'utilisateur à travers une interface de recherche dans une base de données. Cette équipe de recherche s'est concentrée sur le ***Video skimming ou condensé vidéo*** qui est une tâche proche de la création de résumés. La vidéo est segmentée en utilisant les histogrammes des couleurs. La sélection des scènes importantes se fait par la détection des visages, les mouvements de caméra, la capture de texte, et le signal audio. Ensuite, des règles de sélection basées sur ces différents critères sont appliquées.
- le projet **CueVideo**[4] d'IBM. L'accès à la vidéo se fait par visualisation rapide ou storyboard ou storyboard animé. Il fait partie d'un projet beaucoup plus large portant sur l'indexation, la recherche et la lecture rapide de documents vidéo et audio.
- le projet **MoCa**[5] à l'université de Mannheim porte sur les bandes annonces de film. La vidéo est segmentée en utilisant les vecteurs de cohérence ainsi que l'audio. Pour la sélection des scènes importantes, on fait appel à la détection et la reconnaissance d'acteurs principaux, de scènes de bagarre et d'explosion, de texte, etc. Puis le choix des scènes retenues se fait en fonction de considérations générales et d'événements spéciaux.
- le projet **VISU** (pour VIdeo SUmmarization) [Mulhem 03] consiste à annoter les segments d'images d'une vidéo à l'aide de Graphes Conceptuels [Sowa 84] et à stratifier la vidéo en distinguant diverses annotations communes à des segments d'images non contigus de la vidéo. Les Graphes Conceptuels sont un formalisme qui permet des descriptions complexes du contenu d'une vidéo. Ces descriptions peuvent ensuite être manipulées pour le traitement de requêtes par des algorithmes efficaces [Ounis 98].

Les deux types de résumés (**résumé en images et résumé dynamique**) sont nécessaires pour la caractérisation du contenu de la séquence, chacun apportant des informations distinctes : le résumé en images 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 [Li 01] et [Truong 06]. Dans la suite, nous allons brièvement rappeler ces deux techniques de génération de résumés.

[3] http ://www.informedia.cs.cmu.edu/dli2/index.html
[4] http ://www.almaden.ibm.com/projects/cuevideo.shtml
[5] http ://www.informatik.uni-mannheim.de/pi4/projects/MoCA/Project-videoAbstracting.html

6.2.1 Les résumés en images

Comme nous l'avons déjà mentionné dans les paragraphes précédents, un **résumé en images** 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 dans 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 classées en fonction de la manière dont les images clés sont extraites de la séquence. [Li 01] propose quatre catégories :

- ***l'extraction par échantillonnage*** qui consiste à sélectionner des images prélevées uniformément ou aléatoirement 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 les plus importants de la séquence.
- ***l'extraction au niveau des plans*** qui consiste à extraire des informations de bas-niveaux (couleur, mouvement, ...) les mieux adaptées pour contrôler le nombre d'images clés [Michal 95], [Doulamis 00b] et [Aner 01].
- ***l'extraction au niveau des segments***. L'intérêt de cette approche est de travailler sur des unités vidéo de plus haut niveau, et d'être bien adapté à la construction d'un résumé compact [Doulamis 00a].
- ***d'autres approches*** consistent à extraire des images clés à partir de passages comportant des visages, un nombre important d'objets, etc., [Dufaux 00].

Toutes ces méthodes nécessitent une technique pour **comparer les images** clés extraites de chaque plan et une ***distance*** pour ne retenir que les plus différentes.

6.2.2 Les résumés dynamiques

Le résumé dynamique ou "video skim", est une collection de segments audio-visuels extraits de la séquence. C'est également un document vidéo. Le résumé dynamique, $R_{seq}(S)$, de la séquence S peut se définir de la façon suivante :

$$R_{seq}(S) = seg_1 \cup seg_2 \cup ... \cup seg_M \tag{6.2}$$

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).

Le résumé en images peut également 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 consiste à utiliser directement 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 mineurs, ce qui aboutit souvent à un résumé de durée trop élevée.

Un "bon" résumé dynamique nécessite une compréhension sémantique du contenu de la séquence. Les limites des méthodes d'analyse sémantique des images ont réservé les méthodes d'extraction de résumés dynamiques à des domaines spécifiques comme *le sport* [Coldefy 04], *les documentaires* [Yu 03], *les vidéo personnelles* [Zhao 03], etc. La difficulté d'analyse est simplifiée par l'utilisation d'informations *a priori* sur le domaine visé.

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, et ne retenir que les moments jugés ***essentiels***. 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 et les résumés *personnalisés* par interaction avec l'utilisateur.

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 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 (événement rare, ...) [Radhakrishnan 04],
- les passages de la séquence préférés par l'utilisateur (ceux visualisés plusieurs fois, ...) [Yu 03].

Parmi ces résumés "video highlights", il en existe un, la bande-annonce ("movie trailer"), qui ne résume que certains passages particulièrement captivants ou riches en action [Wan 04]. 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.

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 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.

La méthode utilisée pour **la construction du résumé final** consiste à agréger tous les segments en respectant l'évolution temporelle de la séquence. Toutefois, pour le résumé "bande-annonce", qui est un résumé contenant seulement les parties les plus attrayantes de la séquence et qui a comme objectif de susciter un intérêt pour le film, on ne respectera pas toujours l'ordre temporel.

La génération automatique de résumés nécessite l'emploi de distances et d'indices de similarité que nous allons maintenant développer.

6.3 Distances et similarités

6.3.1 Distance entre histogrammes

Dans [Antani 02a], S. Antani et al. exposent la notion de similarité en indexation d'images et de vidéo. Un opérateur de similarité a pour but d'établir les ressemblances ou les relations qui existent entre les informations manipulées.

La similarité est déterminée par le calcul de distances entre certaines caractéristiques extraites ou d'un vecteur les combinant. La méthode la plus couramment adoptée est l'utilisation de distance entre histogrammes.

Si H_1 et H_2 sont des histogrammes calculés sur n "bins" (*cellule de l'histogramme ici une couleur issue d'une palette*) d'images de même taille et si i est l'index dans l'histogramme, alors la différence entre ces histogrammes est donnée par D dans l'équation (6.3). L'intersection entre deux histogrammes est définie par le minimum sur chaque "bin" entre les deux histogrammes. La définition est donnée par l'équation (6.4) qui, lorsqu'elle est normalisée, représente une forte similarité pour une valeur proche de 1.

$$D = \sum_{i=1}^{n} |H_1(i) - H_2(i)| \tag{6.3}$$

$$D_{int} = \frac{\sum_{i=1}^{n} min(H_1(i), H_2(i))}{\sum_{i=1}^{n} H_2(i)} \tag{6.4}$$

Le test de Yakimovsky a été proposé pour détecter la frontière entre deux régions. L'expression du taux de ressemblance de Yakimovsky est donnée par l'équation (6.5), où σ_1^2 et σ_2^2 sont les variances de chaque histogramme, alors que σ_0^2 est la variance des données mises en commun. Les nombres m et n sont les nombres d'éléments dans l'histogramme. Une valeur faible de y indique une forte similarité. Cette méthode de comparaison d'histogrammes peut être appliquée aux histogrammes couleur des images.

$$y = \frac{(\sigma_0^2)^{m+n}}{(\sigma_1^2)^m (\sigma_2^2)^n} \tag{6.5}$$

Le test du χ^2, pour **comparer** deux histogrammes, proposé par Nagaska et Tanaka est décrit par l'équation (6.6), où une valeur faible du χ^2 indique une bonne similarité.

$$\chi^2 = \sum_{i=1}^{n} \frac{(H_1(i) - H_2(i))^2}{(H_1(i) + H_2(i))^2} \tag{6.6}$$

Le test de Kolmogorov-Smirnov décrit par l'équation (6.7) est basé sur l'utilisation d'histogrammes cumulés notée CH_1 et CH_2 dans les équations.

$$D = \max_j [|CH_2(j) - CH_1(j)|] \tag{6.7}$$

Le test de Kuiper défini dans l'équation (6.8), est similaire à celui de Kolmogorov-Smirnov mais est plus sensible aux queues de distribution selon les auteurs.

$$D = \max_j[CH_2(j) - CH_1(j)] + \max_j[CH_1(j) - CH_2(j)] \tag{6.8}$$

La métrique de similarité perceptuelle définie par l'équation (6.9) est similaire à la distance quadratique, avec I et M, les histogrammes calculés sur n "bins" et la matrice $\mathcal{A} = [a_{ij}]$ contenant les coefficients de similitude entre les couleurs des "bins" i et j.

$$d = (I - M)^T \mathcal{A}(I - M) \tag{6.9}$$

6.3.2 Similarité entre plans

Chang et al. [Chang 99] décrivent une technique de génération d'images clés basée sur une nouvelle mesure de fidélité d'un jeu d'images clés. La mesure de fidélité est définie comme étant la distance semi-Hausdorff entre le jeu d'images clés S et le jeu d'images R de la séquence. Soit le jeu d'images clés constitué de m images S_i , $i = 1..m$, et le jeu d'images R contenant n images R_i, $i = 1..n$, soit la distance entre deux images S_i et R_i définie par $d(S_i, R_i)$, on peut alors définir pour chaque image R_i la distance d_i décrite par l'équation (6.10) :

$$d_i = \min[d(S_k, R_i)], k = 1..n \tag{6.10}$$

Alors, la distance Semi-Hausdorff entre S et R est donnée par l'équation (6.11) :

$$d_{SH} = \max(d_i), i = 1..n \tag{6.11}$$

En d'autres termes, pour chaque i, on mesure la distance d_i entre l'image R_i et son image la plus représentative dans le jeu d'images clés S. Puis, on recherche le maximum des distances d_i. De cette manière, on arrive à représenter à quel point le jeu d'images clés S est fidèle à R car, plus la distance Semi-Hausdorff est faible, meilleure est la représentation. Par exemple, si S et R sont identiques, la distance Semi-Hausdorff vaut 0. D'un autre côté, une grande valeur pour d_{SH} indique qu'au moins une des images de R ne correspond à aucune image du jeu d'images clés S.

Dans [Zhang 97], H. J. Zhang et al. proposent la méthode suivante pour définir la similarité entre des plans. Lorsque des images clés sont utilisées comme représentation pour chaque plan, on peut définir une similarité entre plans basée sur les similarités entre les deux jeux d'images clés. Si on note deux plans S_i et S_j, leurs jeux d'images clés, $K_i = \{f_{i,m},\ m = 1..M\}$ *et* $K_j = \{f_{j,n},\ n = 1..N\}$, alors la similarité entre les deux plans peut être définie par (6.12) :

$$S_k(S_i, S_j) = \begin{array}{l}\max[s_k(f_{i,1}, f_{j,1}), s_k(f_{i,1}, f_{j,2}), ..., s_k(f_{i,1}, f_{j,N}), ..., \\ s_k(f_{i,m}, f_{j,1}), s_k(f_{i,m}, f_{j,2}), ..., s_k(f_{i,M}, f_{j,N})]\end{array} \tag{6.12}$$

avec s_k une mesure de similarité entre deux images. Cette définition décrit le fait que la similarité entre deux plans peut être déterminée par la paire d'images clés la plus similaire.

Une autre définition de la similarité entre plans est définie par (6.13) :

$$S_k(S_i, S_j) = \frac{1}{M} \sum_{m=1}^{M} \max[s_k(f_{i,m}, f_{j,1}), s_k(f_{i,m}, f_{j,2}), ..., s_k(f_{i,m}, f_{j,N})] \tag{6.13}$$

Cette définition permet de dire que la similarité entre deux plans est la somme des paires les plus similaires.

Dans ce paragraphe, nous retrouvons la notion de distance entre un point et un ensemble et la notion de distance entre ensembles. Ainsi, d'autres distances, définies dans les chapitres précédents, pourraient être utilisées.

6.4 Utilisation du mouvement

6.4.1 Représentation du mouvement

Comme les caractéristiques de mouvement devraient représenter au mieux la perception humaine et qu'il n'a pas encore été clairement défini comment les humains perçoivent le mouvement, dans [Zhang 97] le choix de la représentation du mouvement est basée sur des caractéristiques statistiques plutôt que sur des trajectoires d'objets. Plus précisément ces caractéristiques comprennent des distributions directionnelles des vecteurs de mouvement (d_i) et des vitesses moyennes dans différentes directions ($\overline{s_i}$). Ces caractéristiques sont calculées à partir du flux optique entre deux images consécutives.

$$d_i = \frac{N_i}{N_{mt}},\ i = 1..M \tag{6.14}$$

avec i, une des M directions, N_i le nombre de points se déplaçant dans la direction i et N_{mt} le nombre de points en mouvement. N_{mt} peut être remplacé par le nombre total de points du flux optique de manière à ne prendre en compte que les surfaces en mouvement de grande taille. De la même manière, on peut estimer la vitesse moyenne et l'écart-type dans une direction donnée, décrits comme suit :

$$\overline{s_i} = \frac{\sum_{j=1}^{N} s_{ij}}{N_i}, \quad \sigma_i = \sqrt{\frac{\sum(s_{ij} - \overline{s_i})^2}{N_i - 1}},\ i = 1..M \tag{6.15}$$

avec s_{ij} la vitesse du j_{ieme} point dans la direction i.

Dans leur projet de positionnement du contenu (publicités, logos de chaînes) dans des vidéos, Wan et Xu [Wan 04] ont eu besoin de repérer les segments vidéo ayant une faible pertinence pour le spectateur. Cette recherche passe par la caractérisation du mouvement. L'information de mouvement pour l'image f est obtenue à partir du flux optique. L'amplitude des vecteurs du flux optique et l'uniformité des directions sont des indications du mouvement global. $mf_{i,j,f}$ est le déplacement normalisé du pixel (i, j) calculé à partir du vecteur du flux optique (dx, dy).

$$mf_{i,j,f} = \frac{\sqrt{dx_{i,j}^2 + dy_{i,j}^2}}{\max_{i,j,f}(\sqrt{dx_{i,j}^2 + dy_{i,j}^2})} \tag{6.16}$$

$\overline{mf_f}$ est la moyenne de tous les déplacements des pixels de l'image, avec N le nombre de pixels dans l'image.

$$\overline{mf_f} = \frac{1}{N} \sum_{i,j} mf_{i,j,f} \tag{6.17}$$

mp_f correspond à l'entropie des directions définie par :

$$mp_f = -\sum_{d}^{D} p_M(d) \log(p_M(d)) \tag{6.18}$$

D correspond à l'ensemble des directions, et $p_M(d)$ correspond à la proportion de pixels se déplaçant dans la direction d.

L'information de mouvement global mv_f pour l'image f est obtenue par :

$$mv_f = \overline{mf_f}(1 - \overline{mf_f} \times mp_f) \tag{6.19}$$

6.4.2 Utilisation du mouvement dans l'extraction d'images clés

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'informations. Dans la réalité, l'information du mouvement issue du flux MPEG n'est pas partout cohérente. Il existe, en effet, des régions qui nécessiteraient une étape de correction (voir Figure 6.1) car des incohérences sont détectées pour certains vecteurs de mouvement issus des 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 par conséquent nécessaire de décompresser la séquence et d'analyser précisément le mouvement [Kraemer 06].

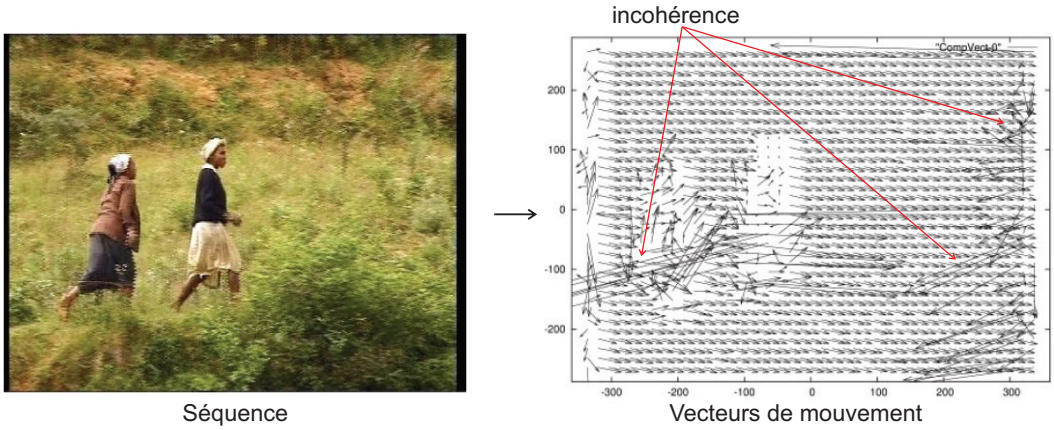


FIG. 6.1: Exemple des vecteurs de mouvement obtenus à partir du flux MPEG2 (source des données : Projet Analyse et Indexation Vidéo [Benois-Pineau 05]).

Une des méthodes consiste à rechercher des minima locaux sur la mesure de mouvement auxquels on attribue une image clé. Cette méthode est basée sur le fait que les passages clés d'une vidéo sont volontairement lents pour leur donner plus d'importance.

Dans [Brunelli 99], Brunelli et al. proposent une extraction d'images clés à partir de l'information sur le mouvement. Dans un premier temps, le flux optique est calculé puis, pour chaque image t, on calcule la somme des amplitudes des composantes du flux optique de chaque pixel.

$$M(t) = \sum_{i,j} |o_x(i,j,t)| + |o_y(i,j,t)| \tag{6.20}$$

où $o_x(i,j,t)$ et $o_y(i,j,t)$ sont les composantes horizontales et verticales du flux optique du pixel (i,j). La seconde étape identifie les maxima locaux de $M(t)$. L'algorithme parcourt progressivement la courbe $M(t)$ et la partitionne à partir des maxima locaux si ceux-ci ont des valeurs qui diffèrent d'au moins 30%. Les images clés sont alors prises sur le minimum de chaque intervalle.

Une autre méthode proposée par A. Divakaran et al. [Divakaran 01] consiste à dire que la mesure du mouvement est une mesure du changement du contenu de la vidéo image par image. Ainsi, en prenant l'intensité cumulée du mouvement, on doit avoir une bonne indication du changement du contenu de la vidéo. Ce qui suggère que la dernière image du plan doit être la plus différente par rapport à la première image du plan puisqu'elle présente la plus forte accumulation d'activité. Ainsi, si l'on choisit la première image du plan comme première image clé, on utilisera logiquement la dernière comme deuxième image clé. En continuant le raisonnement, on peut penser que la troisième image clé

sera celle où le mouvement cumulé sera la moitié du mouvement cumulé total. On peut poursuivre ce raisonnement récursivement pour augmenter le nombre d'images clés souhaité. Cette méthode a l'inconvénient de garder arbitrairement la première et la dernière image du plan.

L'analyse du mouvement dans une séquence d'images doit être posée comme un problème joint d'estimation et de segmentation, puisqu'il s'agit d'appréhender des informations partiellement observables et discontinues. Une partition de l'image en régions cohérentes au sens du mouvement nécessite, sous une forme ou une autre, une mesure de mouvement. Inversement, le calcul du champ des vitesses 2D dans le cas général impose une détection et une gestion simultanées des possibles dicontinuités (inconnues *a priori*) du mouvement. Ce problème peut se présenter en fait sous plusieurs variantes, suivant l'objectif prioritaire se trouvant être l'obtention d'une mesure dense ou paramétrique du mouvement, d'une partition de l'image en régions, ou l'extraction d'entités pertinentes. L'indexation vidéo par le contenu est un champ d'investigation motivant ces études sur la segmentation et la caractérisation du mouvement [Piriou 06].

6.5 Les méthodes proposées

Dans cette section, nous proposons et analysons différentes méthodes d'extraction de résumés en images ou dynamiques. Chacun des résumés proposés joue un rôle précis dans un système d'analyse de séquences d'images. 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.

6.5.1 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, proposée dans [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 ne retenant qu'une seule image clé par plan. Cette approche assure que les images extraites 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 couleur (comme par exemple un "*changement bref de couleur*") 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, toutes les images du plan peuvent 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 sont indépendants 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 [Chanussot 98].
D'un point de vue formel, cette médiane s'exprime par :

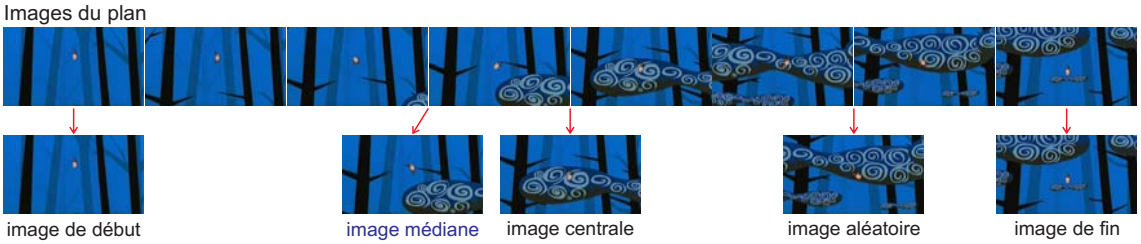
$$I_{méd} = argmin_{I_i \in P}\{D(I_i)\} \tag{6.21}$$

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 :

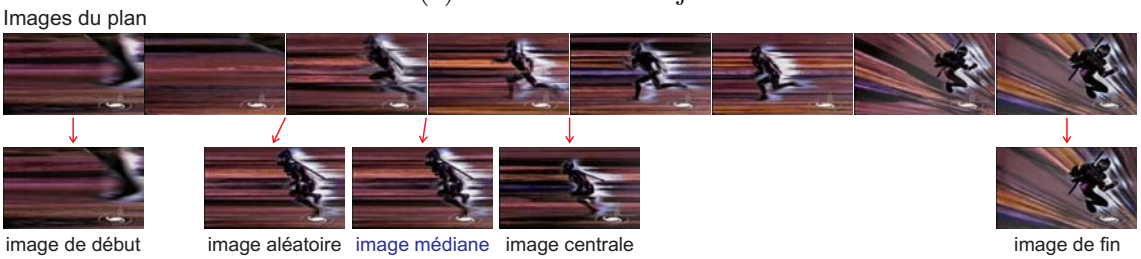
$$D(I_i) = \sum_{I_j \in P, j \neq i} d_{sim}(I_i, I_j) \tag{6.22}$$

où $d_{sim}(I_i, I_j)$ est une mesure de distance calculée entre les images I_i et I_j.

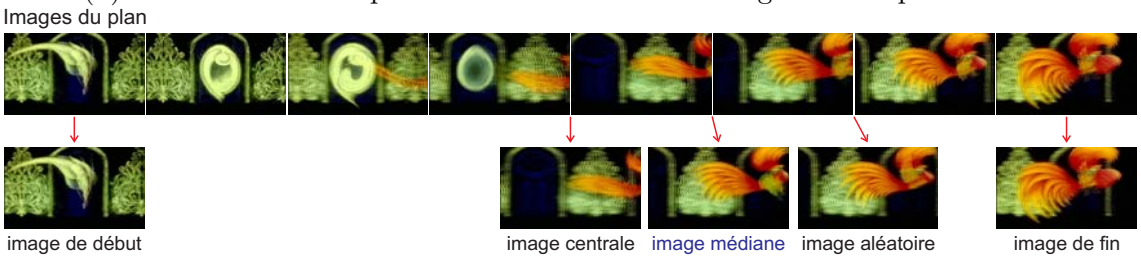
Dans la Figure 6.2, nous proposons quelques exemples utilisant ces stratégies à "une image par plan" (pour l'image médiane nous avons utilisé la distance de Manhattan [Jain 99]).



(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. Mais, 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 représente 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 :

$$R_{img}(S) = \{I_{plan_1} \cup I_{plan_2} \cup ... \cup I_{plan_N}\} \tag{6.23}$$

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 donc testé une technique d'extraction adaptative de résumé en images. Le contenu de chaque plan est résumé avec un *nombre d'images clés qui est adapté à l'action contenue* dans le plan.

Cette technique utilise la distance cumulée définie ci-dessus dans l'équation 6.22. 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.) [Ott 07].

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.

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. Par exemple, dans la Figure 6.3, 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).

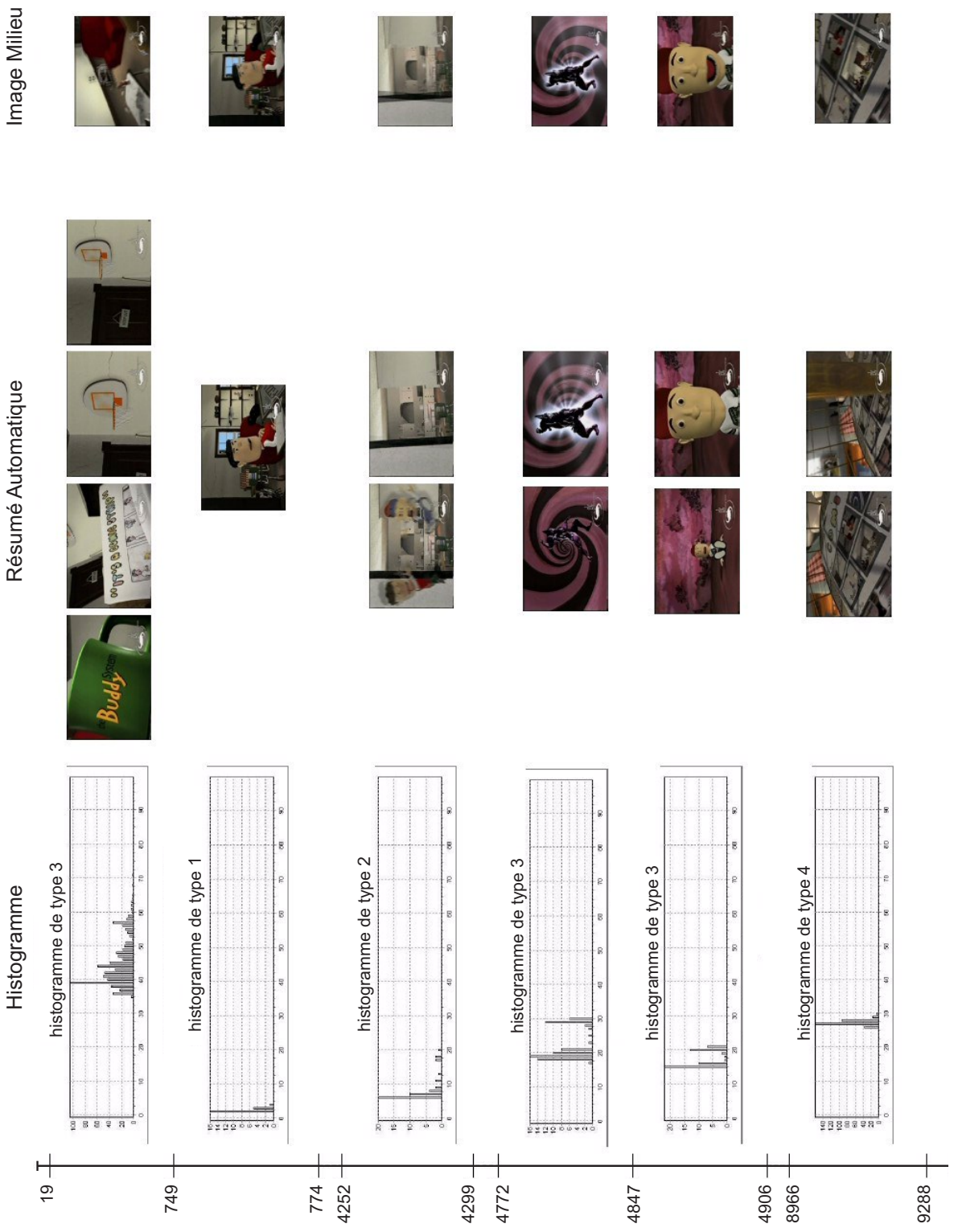


FIG. 6.3: Résultats de l'extraction adaptative des images sur quelques plans du film "The Buddy System" [CICA 06].

FIG. 6.4: Résultats de l'extraction adaptative des images sur quelques plans du film "Gazoon" [CICA 06].

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 d'informations succinctes sur le contenu visuel global de la séquence.

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}, ($N_{init} > N_{comp}$). L'algorithme d'extraction du résumé compact, permet de réduire de manière itérative le nombre d'images de départ, R_{init}, en utilisant un critère de similarité entre les images. La méthode assure que les images retenues dans le résumé sont toujours les images les plus différentes de l'ensemble initial R_{init} ce qui permet d'apporter à chaque image retenue davantage d'information.

Nous avons pu tester le résumé compact sur plusieurs films d'animation de [CICA 06]. Par la suite, nous présenterons et commenterons quelques exemples de résumés obtenus pour des extraits de 3 films d'animation : "François le Vaillant" (Figure 6.5), "Le Moine et le Poisson" (Figure 6.6) et "Le Roman de Mon Âme" (Figure 6.7). 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.

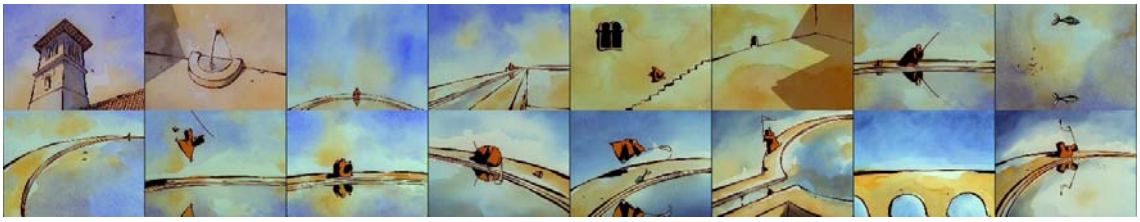
(a) une image par plan (image de milieu)

(b) résumé compact en 5 images

(c) résumé compact en 3 images

FIG. 6.5: Exemple de résumés compacts pour un extrait de 16 plans du film "François le Vaillant" [Folimage 06].

En analysant les résultats, nous avons observé que si l'ensemble d'images clés initial, R_{init}, contient un petit groupe composé d'images similaires 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 6.6 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.

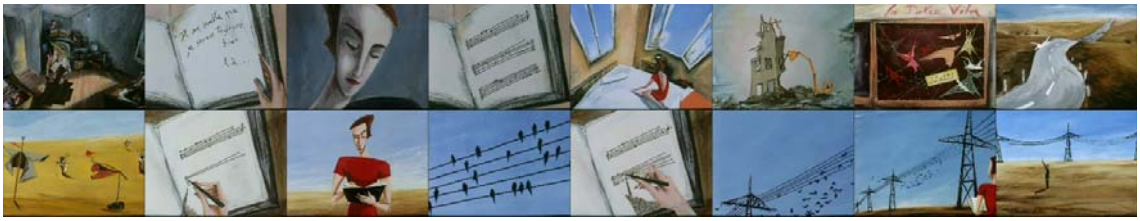


(a) une image par plan (image de milieu)

(b) résumé compact en 5 images

(c) résumé compact en 3 images

FIG. 6.6: Exemple de résumés compacts pour un extrait de 16 plans du film "Le Moine et le Poisson" [Folimage 06].



(a) une image par plan (image de milieu)

(b) résumé compact en 5 images

(c) résumé compact en 3 images

FIG. 6.7: Exemple de résumés compacts pour un extrait de 16 plans du film "Le Roman de Mon Âme" [Folimage 06].

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 entre les histogrammes des images en couleurs réduites que nous avons employée est la distance de Manhattan. La réduction couleur que nous avons retenue est la quantification uniforme de l'espace RVB en $5 \times 5 \times 5$ couleurs. Cette technique est sensible aux variations de l'intensité lumineuse dans l'image. Par exemple, dans la Figure 6.6.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 Âme", film qui comporte beaucoup de changements visuels (voir Figure 6.7). Les images du résumé compact sont toutes différentes entre elles.

Conclusion

Globalement, la durée d'un ***résumé en images*** 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. Par comparaison, pour un ***résumé adaptatif*** le nombre d'images est encore plus élevé, puisque plusieurs images peuvent 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.

Le ***résumé en images*** 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, sans prendre le temps de la regarder entièrement. Par exemple, les 100 images peuvent être visualisées, à une cadence d'une image toutes les 2 secondes, en 50 secondes, ou même seulement en quelques secondes si elles sont organisées sous la forme d'une planche (voir Figure 6.5.a). En conclusion, bien que volumineux, ce résumé en images 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 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 et nous avons constaté que la restriction à quelques images par plan n'altère pas beaucoup les résultats.

Pour 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 en images est très mal adapté à cette situation alors que le résumé compact (voir Figure 6.5.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.

6.5.2 Les résumés dynamiques

Les résumés dynamiques, quant à eux, 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 par la suite, le son ne sera pas exploité. Pour le moment, nous nous sommes limités à l'utilisation de l'image. Le son est néanmoins un élément important qui devra être pris en compte à l'avenir. Notons cependant que, dans la plupart des films d'animation utilisés pour nos expérimentations (42 sur 52) il n'a pas de dialogues ou de commentaires mais uniquement de la musique.

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 :

$$R_{mouv}(S) = \{seq^c_{1,p} \cup seq^c_{2,p} \cup ... \cup seq^c_{N,p}\} \tag{6.24}$$

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 donc plus d'information, seront 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). Ainsi, pour le film "Fini Zayo" [Folimage 06] 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 06], 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), nous proposons un résumé dynamique prennant en compte l'action contenue dans la séquence [Ionescu 06].

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 est présenté dans [Ionescu 07].

Le résumé dynamique proposé s'appuie sur le résumé du contenu de chaque segment d'action de la séquence. Construit de cette manière, *il ne contiendra que 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 de films conçues pour attirer le spectateur 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 :

$$R_{ba}(S) = \{pass_1 \cup pass_2 \cup ... \cup pass_M\} \tag{6.25}$$

où S est 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 :

$$pass_i = \{seq^c_{i,1,p} \cup seq^c_{i,2,p} \cup ... \cup seq^c_{i,N_i,p}\} \tag{6.26}$$

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 de 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), comme par exemple le film "Ferrailles" [Folimage 06], ou du fait de la fréquence élevée des changements de plan, la plupart des plans ayant été considérés importants pour le résumé.

Discussion

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 non 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 liées 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 des 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 6.8). 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 ne provient pas du rythme des changements de plan.

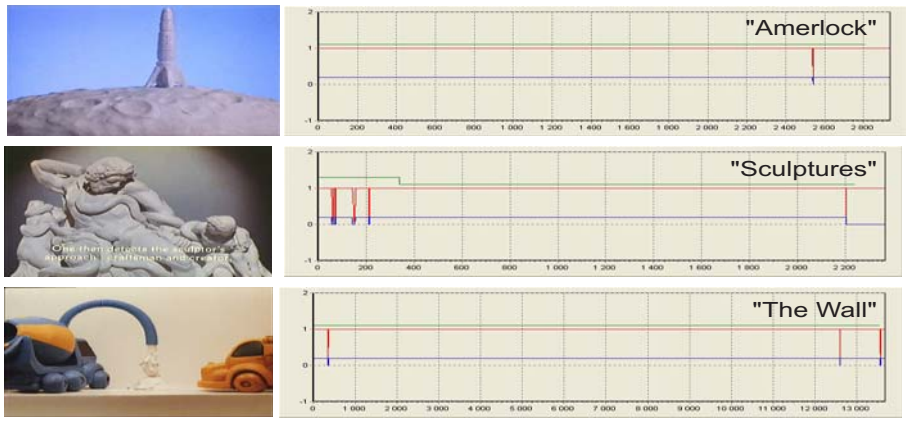


FIG. 6.8: Exemple de films d'animation [CICA 06] contenant un nombre réduit de plans. Chaque film est représenté par une image et l'annotation visuelle des transitions (l'axe oX est l'axe temporel et les lignes rouges verticale indiquent un changement de plan).

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.

Enfin des travaux intéressants sur la construction de résumés vidéo sont présentés dans [Guironnet 06]. L'auteur a développé trois méthodes d'extraction de résumés vidéo, à savoir :

- le résumé hiérarchique à partir de caractéristiques de bas-niveau,
- le résumé extrait à partir de l'information de mouvement de caméra,
- le résumé basé sur l'attention visuelle.

Il semble que, sur les tests effectués lors de cette thèse, ce soit le résumé vidéo basé sur le mouvement de caméra, qui donne les meilleurs résultats. Par exemple, un zoom avant informe sur l'importance d'un passage de vidéo. Pour notre part, nous avons également remarqué, que bien souvent les passages qui succédaient un plan dans lequel il y a une forte activité (beaucoup d'action), devaient être retenus dans le résumé.

6.6 Comparaison de séquences en utilisant les gamuts sémantiques

Comme nous l'avons déjà mentionné, nous souhaitons **comparer les séquences d'images**, mais celles-ci sont trop volumineuses pour être traitées de façon globales. Une solution consiste à extraire certains attributs de ces séquences et à les comparer.

Dans la thèse de Bogdan IONESCU nous avons mis en place un *système de traitement* capable d'analyser et de "comprendre" de manière automatique le contenu des films d'animation [Ionescu 07]. Le système est composé d'un ensemble d'outils qui traduisent des données de bas niveau (me-

sures mathématiques, statistiques caractérisant certaines propriétés de la séquence, ...), difficilement compréhensibles pour les non spécialistes, en des données de haut niveau (proche de la perception humaine) exprimées dans un langage accessible au plus grand nombre. Le système proposé a été appliqué au domaine particulier des films d'animation [CICA 06].

Dans cette partie, nous proposons une *représentation visuelle* des caractéristiques des films. Cette représentation peut être utilisée pour comparer le contenu de différents films. Elle permet également de trouver rapidement les caractéristiques communes, fonction indispensable de tout moteur de recherche d'une base de données vidéo. Les caractérisations de chaque film sont illustrées en utilisant une représentation graphique inspirée de la construction de gamuts de couleurs d'un dispositif de restitution d'images couleur (comme par exemple, l'écran ou l'imprimante). Nous appellerons cette représentation **le gamut sémantique** de la séquence.

6.6.1 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 différents axes dans l'espace 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 par l'ensemble des valeurs des paramètres représentés. Un exemple est présenté dans la Figure 6.9.

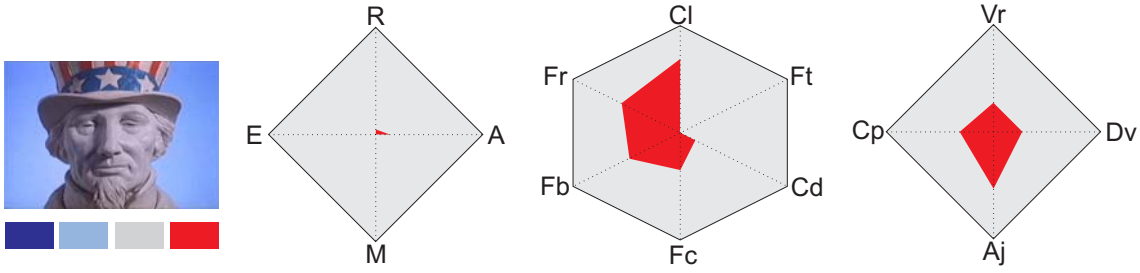


FIG. 6.9: Les gamuts sémantiques obtenus pour le film "Amerlock" : G^p, G^c et G^{rl} (les couleurs élémentaires prédominantes sont illustrées en bas de l'image).

Nous avons divisé les caractérisations des séquences en trois catégories. Pour chacune, nous avons associé un gamut sémantique, dont : *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 faiblement 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 couleur (paramètre $100 \cdot P_{var}$, valeur maximale 100%), **Dv**- diversité couleur (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%).

En annexe B, nous trouvons la définition de ces paramètres.

Ce type de *représentation visuelle compacte* permet de se faire une *idée globale* de l'ensemble des caractéristiques de la séquence. Ainsi, la tâche de comparaison des différents films s'en trouve 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 aux films analysés. Les films ayant différentes caractéristiques sémantiques auront différentes formes de gamuts sémantiques et inversement, les films similaires du point de vue du contenu comporteront des gamuts similaires.

6.6.2 Résultats expérimentaux

L'originalité de ces travaux réside dans l'efficacité de cette représentation, que nous avons testée sur plusieurs films d'animation. Les résultats sont donnés dans [Ionescu 07].

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).

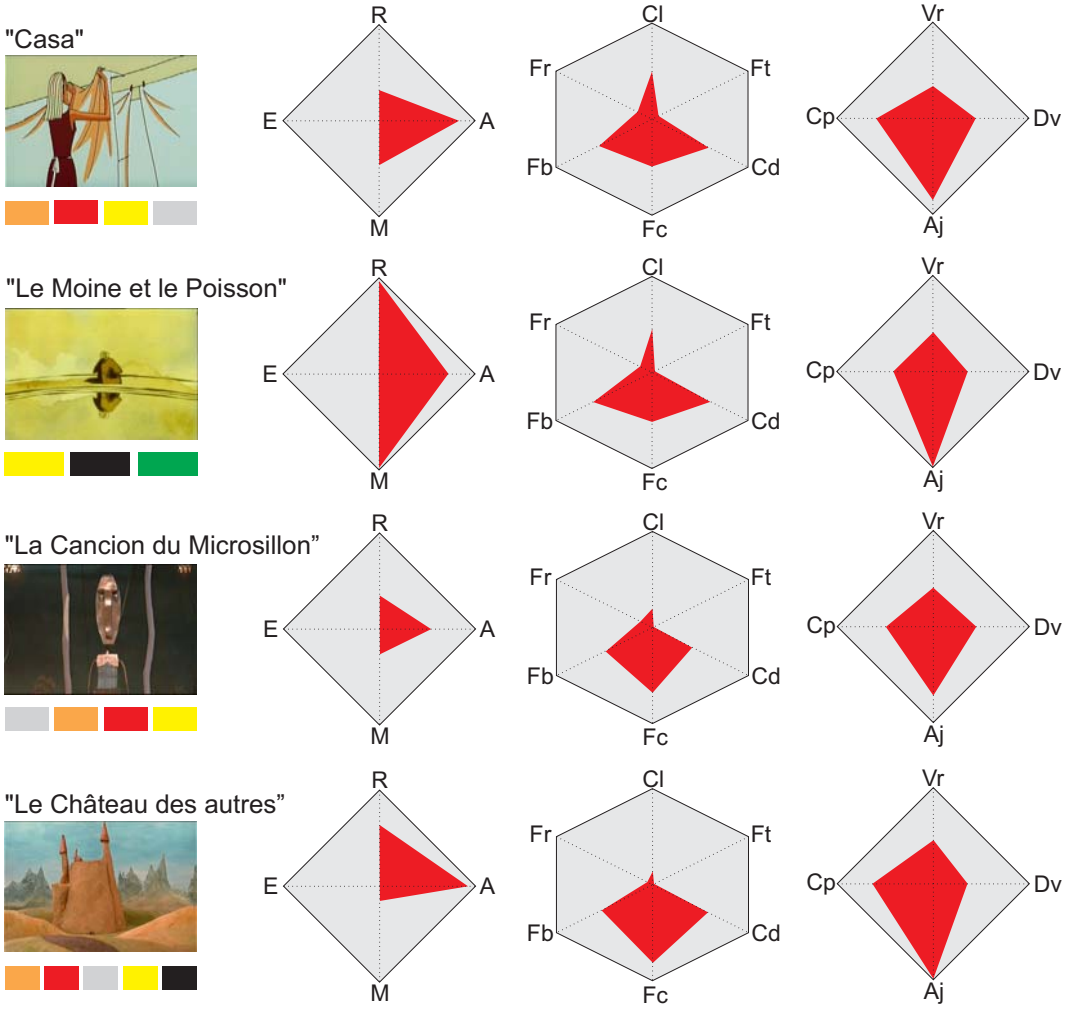


FIG. 6.10: Les gamuts sémantiques G^p, G^c et G^{rl} obtenus pour les films cités ci-dessus (les couleurs élémentaires prédominantes de chaque film sont illustrées en dessous de l'image).

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", "Rouge", "Jaune" et "Gris" et des techniques couleurs similaires. Les gamuts des propriétés couleurs sont ressemblants (voir la Figure 6.10).

6.6.3 Les applications

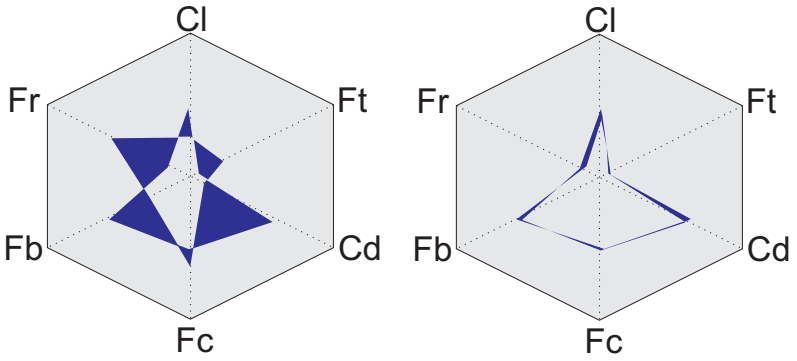
Une application immédiate de ce type de représentation graphique peut être faite dans un *outil de navigation* d'une base de données vidéo. En associant à chaque film les gamuts proposés (comme ils sont présentés dans la Figure 6.9), l'utilisateur pourra, d'un seul coup d'œil, avoir une impression rapide et efficace du *contenu* de la séquence, sans passer de 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 *signatures visuelles des caractéristiques* de la séquence.

Une autre application possible est l'utilisation de ces gamuts sémantiques 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 :

$$d_{gamut}(G_1, G_2) = Surf(G_1 \cup G_2 - G_1 \cap G_2) \tag{6.27}$$

où G_1 et G_2 sont deux gamuts sémantiques, représentant la même catégorie sémantique, des deux films à comparer et ou $Surf()$ est l'opérateur calculant 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. 6.11: Exemple de distance entre gamuts (les abréviations sont expliquées en annexe).

Pour montrer la capacité discriminatoire de la mesure de distance proposée nous présentons un exemple. L'exemple est illustré par la Figure 6.11 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". 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 6.11.b), les deux films étant 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 6.11.a) est importante car le film "Casa" et le film "François le Vaillant" sont très différents en ce qui concerne les caractéristiques des couleurs.

Discussion : Des améliorations peuvent être proposées sur ces gamuts sémantiques. Tout d'abord, les valeurs associées à ces gamuts n'ont pas la même échelle, il serait donc intéressant de trouver une normalisation adaptée. Nous pouvons également utiliser d'autres distances pour comparer ces gamuts. Enfin, pour comparer deux gamuts, il est impératif de respecter l'ordre dans lequel les caractéristiques sont représentées.

6.7 Conclusions

Dans ce chapitre, nous avons proposé deux approches permettant de **comparer les séquences d'images**. La première est basée sur la création automatique de résumés, tandis que la seconde repose sur la comparaison des caractéristiques extraites des séquences par le biais d'une représentation visuelle.

Les résumés : Les approches existantes sont orientées vers deux directions distinctes : les **résumés en images** 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 simplicité 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 long (plus long 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 comme 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 contenue dans la séquence.

  Aussi avons-nous proposé un résumé plus compact ne reproduisant que 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 plans, 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, l'évaluation d'un résumé, devant représenter le contenu global de la séquence, ne peut pas porter sur les mêmes critères que l'évaluation de résumé, devant contenir les événements jugés essentiels 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, l'information dynamique manque, mais sa durée est beaucoup plus courte que celle du résumé dynamique. 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.

Enfin nous avons proposé une méthode permettant de **comparer des séquences d'images** à partir de leurs gamuts sémantiques. C'est une piste qui nous semble intéressante car elle permet de comparer d'un coup d'œil les caractéristiques des films et de repérer, à l'aide d'une distance adaptée, les séquences qui sont similaires ou non.

Chapitre 7

Conclusions et Perspectives

Nous avons proposé dans ce mémoire une approche sur la **comparaison des images** pour l'évaluation des traitements, la reconnaissance des formes et l'indexation de séquences d'images.

Les travaux développés durant ces dernières années, au travers des différents DEA-Master et de la première thèse encadrée ont essentiellement porté sur l'optimisation des opérateurs de distance en 2D et 3D. Nous avons utilisé ces opérateurs de distance pour la **comparaison d'images** binaires, en niveaux de gris et en couleur. Cela nous a permis d'**évaluer** des algorithmes et des opérateurs de traitement d'images.

La deuxième thèse nous a permis de travailler sur les séquences d'images. Les travaux présentés dans cette deuxième 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 du contenu proches de la perception humaine qui peuvent servir d'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. Enfin nous avons proposé une méthode permettant de **comparer les séquences d'images** à partir de leurs gamuts sémantiques.

Suite à ces travaux, plusieurs pistes mériteraient d'être approfondies.

Tout d'abord, il existe des pistes intéressantes sur la façon dont il faut représenter une image. Les problèmes du bon choix de caractéristiques pour décrire une image et la pondération de ces caractéristiques ne sont pas encore résolus. Alors qu'un consensus minimal sur l'utilisation de la couleur et la texture semble se dégager des travaux présentés ces dernières années, certains systèmes ajoutent d'autres caractéristiques (points d'intérêts, ordre et contraste, ...), variables selon les systèmes et surtout selon les applications visées [Simand 05]. Quant à la pondération des caractéristiques dans la fonction de **mesure de similarité** entre les images, elle varie aussi fortement selon les systèmes, les applications et même selon les requêtes. La solution semble exister dans l'adaptation de la fonction et de ses poids, mais selon quels critères ? Ces critères ne sont pas forcément d'ordre symbolique, mais plutôt d'ordre sémantique.

Raisonner au niveau **sémantique** signifie que l'analyse de l'image se fait en termes d'objet et de contenu, et non pas seulement en termes de statistiques sur les couleurs, les textures ou d'autres caractéristiques bas niveau extraites de l'image. Puisque la sémantique n'est pas inscrite dans l'image, il faut rechercher des sources d'information extérieures nous donnant accès aux clés de décodage sémantique de l'image. Cette sémantique doit se retrouver selon deux approches complémentaires

et indissociables : l'une recherche des moyens pour lier la connaissance sémantique humaine et l'apparence de l'image (caractéristiques extraites de celle-ci, définition de concepts liés à l'image, ...), l'autre recherche des méthodes pour comprendre l'objectif de l'utilisateur, le sens de sa requête (une requête peut signifier différents buts selon le contexte) au moyen d'une interface interactive avec l'utilisateur [Santini 01]. Par ces interactions, le système apprend les intentions de l'utilisateur et rend des résultats capables de le satisfaire.

Dans [Mulhem 03], les auteurs utilisent un vocabulaire général pour identifier les visages, les foules, le ciel, le sol, l'eau, le feuillage, les montagnes, les bâtiments, etc. La représentation des concepts et le raisonnement sur ceux-ci se font à l'aide de graphes conceptuels, autre technique répandue. Sauf dans le cas d'applications spécifiques où l'on peut sur-spécialiser le système, il est difficile dans le cas général de faire une sélection pertinente de concepts qui seraient valides pour tous les usages imaginables. Pour faire coïncider les concepts sémantiques aux caractéristiques de l'image, l'apprentissage est une tendance populaire et efficace. C'est une manière d'ajouter de la connaissance au système sans lui imposer une vision humaine. Mais alors la difficulté est dans le choix de la base d'apprentissage. Une autre difficulté est liée à la diversité des requêtes des utilisateurs. Comment peut-on satisfaire toutes les requêtes possibles ?

Pour la recherche d'images par le contenu, une solution consiste à construire une requête multiple et repose sur une compétition de modèles pour le raffinement de la mesure de similarité. Quelques travaux ont cherché à transposer les techniques d'enrichissement de la requête, entre autre par le "bouclage de pertinence" [Cord 04]. L'objectif est de faire des interactions entre le système et l'utilisateur afin de faire refléter cette subjectivité dans les poids des descripteurs pour la composition de la réponse. Cette technique est surtout utilisée lorsqu'on a préalablement indexé les images de la base, et non lorsqu'on calcule sur demande la similarité entre images.

Quelle **mesure de dissimilarité** faut-il choisir ? Nous pensons qu'une mesure de dissimilarité qui combinerait des informations quantitatives et qualitatives, en associant l'avis de l'utilisateur, permettrait d'apporter plus de subjectivité. De nombreux travaux existent dans ce domaine, et pour le moment les mesures proposées affectent un coefficient de pondération aux informations quantitatives et qualitatives. Pour répondre à cette approche, un axe de recherche intéressant serait l'étude de nouveaux descripteurs qui permettent de caractériser l'***impression*** qui se dégage d'une image, ce que les japonais appellent le "***Kansei***". Ce type d'étude, peu encore développé en France, connaît un fort intérêt au Japon depuis plusieurs années [Yang 99], [Jiao 06]. Par exemple, le robot Kansei au Japon offre un nouveau degré d'interactivité, puisque le visage est composé d'un masque en silicone de 19 parties mobiles qui permet de montrer jusqu'à 36 expressions faciales en réagissant à certains mots. Il peut accéder à une base de données en ligne d'un demi-million de mots pour le moment, et le fait que la base de données se met à jour d'elle-même rend le robot Kansei encore plus dynamique. Par exemple, le robot fronce les sourcils quand il entend le mot "bombe", sourit quand "sushi" est mentionné.

Pour nos applications, il s'agit de donner la description d'une image en des termes proches de la perception humaine : *cette image dégage une impression de bonheur*. Cela peut se faire à partir de fonctions qui associent des images à des mots liés à une impression. Par exemple, on passe par une première étape qui permet de calculer une signature multi-dimensionnelle caractérisant l'image selon des attributs de bas-niveau, puis à l'aide d'une fonction de classification, on détermine la signature multi-dimensionnelle du terme qui désigne une impression. Cette approche avait été initiée lors de la thèse de Nadia Bouloudani.

Les applications potentielles de ces travaux sont multiples. On peut citer en particulier la recherche dans une base d'images. La plupart des travaux existants s'appuient sur des mesures de similitude entre des attributs de forme, de texture ou de couleur. Il serait extrêmement riche, pour répondre de manière plus pertinente aux requêtes, d'ajouter des descripteurs de ce type. Plus récemment avec la deuxième thèse soutenue, nous nous sommes penchés sur la caractérisation symbolique de séquences d'images et avons initié des travaux dans ce sens. Là aussi, des descripteurs issus du "***Kansei***" apporteraient une richesse supplémentaire.

Une deuxième piste de recherche qui me semble intéressante est le **rebouclage** dans les systèmes de traitement d'images. En effet, toute méthode de traitement d'images, ou plus généralement, tout système nécessite des paramètres. Or, il est bien souvent difficile d'ajuster ces paramètres et ceux-ci sont, la plupart du temps, liés à l'application. En utilisant une mesure de dissimilarité associée au jugement d'un utilisateur, nous pourrions comparer les résultats successifs, guider l'utilisateur et ajuster les paramètres pour tendre vers un "*bon*" résultat. Ma participation plus active à l'action 3 (SCATI) du thème B du GDR ISIS et les résultats encourageants que nous avons obtenus dans le projet BQR portant sur "*un système coopératif de fusion d'informations pour l'interprétation d'images 3D*", vont dans ce sens [Valet 07]. Nous pensons que l'**interaction** avec l'utilisateur par le biais d'une interface conviviale est une piste à ne pas écarter.

Une dernière piste qui me tient à cœur est l'**évaluation de performance des systèmes de fusion d'informations**, plus particulièrement, l'évaluation de ceux associés aux traitements des images. Il serait très intéressant de définir une méthodologie qui permette d'évaluer la performance des systèmes de fusion d'informations. Notamment, cette évaluation devrait permettre de :

- **comparer** différentes méthodes de fusion,
- de mieux comprendre l'apport de chaque paramètre de la méthode et leur interaction sur la sortie du système
- de mieux comprendre l'apport de chaque système de fusion, dans des systèmes coopératifs.

Les méthodes subjectives ont démontré leurs intérêts mais nécessitent des conditions d'expérimentation parfois lourdes (tests sur une base de données importante, bonnes conditions de visualisation, durée importante pour effectuer l'évaluation, choix du nombre d'utilisateurs expérimentés ou non, ...). Les méthodes objectives nécessitent quant à elles une vérité terrain, pour qu'elles soient pertinentes. Mais cette vérité terrain est parfois difficile à obtenir. Nous proposerons une mesure objective qui évaluera la quantité relative d'information qui est transférée de l'image d'entrée vers l'image de sortie du système de fusion. Nous pensons, par exemple, à une mesure de dissimilarité entre images, ou une mesure sur la préservation des contours ou des régions, ou une mesure basée sur des statistiques locales calculées sur une portion de l'image. Afin de valider de manière significative la méthode de fusion, un mécanisme de tests appropriés de comparaison subjective-objective sera défini, par exemple, à partir d'un classement subjectif ou d'un vote [Petrovic 07].

C'est vers ces trois dernières pistes, qui sont complémentaires, que je souhaite m'engager dans mes futurs travaux de recherche.

Bibliographie

[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, 2001.

[Aksoy 98] S. Aksoy & R. Haralick. *Textural features for image database retrieval.* IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR'98, 1998.

[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, Hawaï, USA, 2001.

[Antani 02a] R. Antani, R. Kasturi & R. Jain. *A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video.* Pattern Recognition, vol. 35, pages 945–965, 2002.

[Antani 02b] S. Antani, R. Kasturi & R. Jain. *A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video.* Pattern Recognition, vol. 35, no. 4, pages 945–965, 2002.

[Ardizzoni 99] S. Ardizzoni, I. Bartolini & M. Patella. *Windsurf : Region-based image retrieval using wavelets.* In DEXA Workshop, pages 167–173, 1999.

[ARGOS 06] ARGOS. *Campagne d'Evaluation d'Outils de Surveillance de Contenus Vidéo.* http ://www.irit.fr/argos, 2006.

[Bach 96] J.R. Bach & et al. *Virage image search engine : An open framework for image management.* Strorage and Retrieval for image and video Databases IV, IS&T/SPIE, 1996.

[Baddeley 92] A.J. Baddeley. *An error metric for binary images.* Robust Computer Vision, Wichmann, Karlsruhe, pages 59–78, 1992.

[Baudrier 05] E. Baudrier. *Comparaison d'images binaires reposant sur une mesure locale des dissimilarités : Application à la classification.* Thèse de l'Université de Reims Chapagne Ardenne, 2005.

[Benois-Pineau 05] J. Benois-Pineau. *Extraction des Objets Couleur en Mouvement des Séquences Vidéo.* LABRI UMR CNRS 5800, www.labri.fr/ImageetSon/AIV, 2005.

[Bolon 92] Ph. Bolon & J.L. Vila. *Opérateur local de distance en maillage rectangulaire.* Proc. 2ème colloque de géométrie discrète : Fondements et Applications, pages 45–56, 1992.

[Bolon 95] Ph. Bolon, J.M. Chassery, J.P. Cocquerez, D. Demigny, C. Graffigne, A. Montanvert, S. Philipp, R. Zeboudj & J. Zerubia. *Analyse d'images : Filtrage et Segmentation.* Eds J.P. Cocquerez et S. Philipp (coordinateurs), MASSON, page 457 pages, 1995.

[Borgefors 84] G. Borgefors. *Distance transformation in arbitrary dimensions.* Computer Vision, Graphics and Image Processing, vol. 27, pages 321–345, February 1984.

[Borgefors 86] G. Borgefors. *Distance transformations in digital images.* Computer Vision, Graphics and Image Processing, vol. 34, no. 3, pages 344–371, February 1986.

[Breu 95] H. Breu, J. Gil, D. Kirkpatrick & M. Werman. *Linear time Euclidean distance transform algorithms.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 5, pages 529–533, 1995.

[Brunelli 99] R. Brunelli, O. Mich & C.M. Modena. *A Survey on the Automatic Indexing of Video Data*. Journal of Visual Communication and Image Representation, vol. 10, pages 78–112, 1999.

[Brunelli 01] R. Brunelli & O. Mich. *Histograms analysis for image retrieval*. Pattern Recognition, vol. 34, no. 8, pages 1625–1637, 2001.

[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.

[Carson 99] C. Carson, M. Thomas, S. Belongie, J. Hellerstein & J. Malik. *Blobworld : A system for region-based image indexing and retrieval*. In Third International Conference on Visual Information Systems, Springer, 1999.

[Chang 99] H.S. Chang, S. Sull & S.U. Lee. *Efficient video indexing scheme for content-based retrieval*. IEEE Transaction on Circuits Systemes Video Technology, vol. 9, no. 8, pages 1269–1279, 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.

[Chehadeh 95] Y. Chehadeh, D. Coquin & Ph. Bolon. *A generalization to cubic and non cubic local distance operators on parallelepipedic grids*. Proc. 5th Discret Geometry for Computer Imagery, pages 27–36, 1995.

[Chehadeh 96] Y. Chehadeh, D. Coquin & Ph. Bolon. *A skeletonization algorithm using chamfer distance transformation adapted to rectangular grid*. IEEE International Conference on Pattern Recognition, vol. 2, pages 131–135, Vienne, Austria, 1996.

[Chehadeh 97] Y. Chehadeh. *Opérateurs locaux de distance en maillages rectangulaire et parallélépipédique : application à l'analyse d'images*. Thèse de doctorat de l'Université de Savoie, Octobre, 1997.

[Chen 99] Y. Chen & E.K. Wong. *Augmented Image Histogram for Image and Video Similarity Search*. SPIE Conf. Storage and Retrieval for Image and Video Database VII, pages 523–532, 1999.

[CICA 06] CICA. *Centre International du Cinema D'Animation*. http :// www.annecy.org, 2006.

[Coeurjolly 02] D. Coeurjolly. *Algorithmique et géométrie discrète pour la caractérisation des courbes et des surfaces*. Thèse de Doctorat de l'Université Lyon II, Décembre, 2002.

[Coeurjolly 07] D. Coeurjolly & A. Montanvert. *Optimal Separable Algorithms to Compute the Reverse Euclidean Distance Transformation and Discrete Medial Axis in Arbitrary Dimension*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 3, 2007.

[Coldefy 04] F. Coldefy & P. Bouthemy. *Unsupervised Soccer Video Abstraction Based on Pitch, Dominant Color and Camera Motion Analysis*. ACM Multimedia, pages 268–271, New York, USA, 2004.

[Coquin 93] D. Coquin & Ph. Bolon. *Comparaison d'opérateurs locaux de distance*. Proc. 3ième Colloque de Géométrie Discrète : Fondements et Applications, pages 182–191, 1993.

[Coquin 94] D. Coquin, Y. Chehadeh & Ph. Bolon. *3D local distance operator on parallelepipedic grid*. Proc. 4th Discret Geometry for Computer Imagery, pages 147–156, 1994.

[Coquin 95a] D. Coquin & Ph. Bolon. *Discrete distance operator on rectangular grids*. Pattern Recognition Letters, vol. 16, pages 911–923, 1995.

[Coquin 95b] D. Coquin, Ph. Bolon & Y. Chehadeh. *Opérateur de distance 3D - Application à la comparaison d'images*. 15ème colloque GRETSI, pages 761–764, Juan-les-Pins, France, 1995.

[Coquin 97] D. Coquin, Ph. Bolon & Y. Chehadeh. *Evaluation quantitative d'images d'images filtrées*. 16ème colloque GRETSI, vol. 2, pages 1351–1354, Grenoble, France, 1997.

[Coquin 00a] D. Coquin, Ph. Bolon & A. Onea. *3D Nonstationary Local Distance Operator*. IEEE International Conference on Pattern Recognition, vol. 3, pages 963–966, Barcelona, Spain, 2000.

[Coquin 00b] D. Coquin, Ph. Bolon & A. Onea. *Objective metric for colour image comparison*. in European Signal Processing Conference, pages 39–56, Tempere, Finlande, 2000.

[Coquin 01a] D. Coquin & Ph. Bolon. *Application of Baddeley's distance to dissimilarity measurement between gray scale images*. Pattern Recognition Letters, vol. 22, pages 1483–1502, 2001.

[Coquin 01b] D. Coquin & Ph. Bolon. *Quantitative assessment of image filtering : comparison of objective metrics*. Imaging and Vision Systems : Theory, Assessment and Applications, Nova Science Publishers, vol. 9, no. 7, pages 129–140, Huntington, New-York, 2001.

[Coquin 02a] D. Coquin & Ph. Bolon. *A new method to compute the distortion vector field from two images*. IEEE International Conference on Pattern Recognition, pages 279–282, Quebec City, Canada, 2002.

[Coquin 02b] D. Coquin, Ph. Bolon & B. Ionescu. *Dissimilarity measures in color spaces*. IEEE International Conference on Pattern Recognition, pages 612–615, Quebec City, Canada, 2002.

[Coquin 06] D. Coquin, E. Benoit, S. Hideyuki & B. Ionescu. *Gestures recognition based on the fusion of Hand positioning and Arm gestures*. Journal of Robotics and Mechatronics, vol. 18, no. 6, pages 751–759, December, 2006.

[Cord 04] M. Cord, J. Fournier & S. Philipp-Foliguet. *Approche interactive de la recherche d'images par le contenu*. RTSI, Technique et Science Informatique, vol. 23, no. 1, pages 93–123, 2004.

[Cuisenaire 99a] O. Cuisenaire. *Distance transformations : Fast algorithms and applications to medical image processing*. Thèse, UCL : Université Catholique de Louvain-la-Neuve, Belgique, 1999.

[Cuisenaire 99b] O. Cuisenaire & B. Macq. *Fast Euclidean distance transformation by propagation using multiple neighborhood*. Computer Vision and Image Understanding, vol. 76, pages 163–172, 1999.

[Danielsson 80] P.E. Danielsson. *Euclidean Distance Mapping*. Computer Graphics and Image Processing, vol. 14, pages 227–248, 1980.

[Del Bimbo 99] A. Del Bimbo. *Visual Information Retrieval*. Morgan Kaufmann Publishers, San Francisco, 1999.

[Derrode 99] S. Derrode, M. Daoudi & F. Ghorbel. *Invariant content-based retrieval using a complete set of Fourier-Mellin descriptors*. IEEE Int. Conf. on Multimedia Computing and Systems (ICMCS'99), 1999.

[Detyniecki 03] M. Detyniecki & E. Marsala. *Discovering Knowledge for Better Video Indexing Based on Colors*. IEEE International Conference on Fuzzy Systems, vol. 2, pages 1177–1181, Paris, France, 2003.

[Devijver 82] P.A. Devijver & J. Kittlet. *Pattern Recognition : a Statistical Approach*. Prentice Hall, Englewood Cliffs, London, 1982.

[Di Gesù 99] V. Di Gesù & V. Starovoitov. *Distance-based function for image comparison.* Pattern Recognition Letters, vol. 20, pages 207–214, 1999.

[Dijkstra 59] E.W. Dijkstra. *A note on two problems in connexion with graphs.* Numerische Mathematik, vol. 1, pages 269–271, 1959.

[Divakaran 01] A. Divakaran, R. Radhakrishnan & K.A. Peker. *Video Summarization using Descriptors of Motion Activity : A Motion Activity based Approach to Key-Frame Extraction from Video Shots.* Journal of Electronic Imaging, vol. 10, no. 4, pages 909–916, 2001.

[Doulamis 00a] A.D. Doulamis, Doulamis N. & S. Kollias. *A Fuzzy Video Content Representation for Video Summarization and Content-Based Retrieval.* Signal Processing, vol. 80, no. 6, pages 1049–1067, 2000.

[Doulamis 00b] A.D. Doulamis, Doulamis N. & 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.

[Dubuisson 93] H. Dubuisson. *Caractérisation de la trajectoire du grimpeur.* DEA AII, Université de Savoie, 1993.

[Dubuisson 94] M.P. Dubuisson & A.K. Jain. *A modified Hausdorff Distance for object matching.* Proceeding 12th International Conference on Pattern Recognition, pages 566–568, Jerusalem, Israel, 1994.

[Duda 01] R.O. Duda, P.E. Hart & D.G. Stork. *Pattern Classification.* Second Edition, A Wiley-Interscience Publication, John Wiley and Sons, INC., 2001.

[Dufaux 00] F. Dufaux. *Key Frame Selection to Represent a Video.* IEEE International Conference on Multimedia and Expo, vol. 2, pages 275–278, 2000.

[Eggers 97] H. Eggers. *Fast parallel Euclidean distance transformation in $\mathbb{Z}^n$.* In SPIE Proceedings, Vision Geometry IV, vol. 3168, pages 33–40, San Diego, 1997.

[Eggers 98] H. Eggers. *Two fast Euclidean distance transformations in $\mathbb{Z}^2$ based on sufficient propagation.* Computer Vision and Image Understanding, vol. 69, no. 1, pages 106–116, 1998.

[Flickner 95] M. Flickner & et al. *Query by image and video content : the QBIC system.* IEEE Computer, vol. 28, no. 9, pages 23–32, 1995.

[Folimage 06] Studio Folimage. *Présentation du studio et de ses productions.* http ://www.folimage.com, 2006.

[Forchhammer 89] S. Forchhammer. *Euclidean distances from chamfer distances for limited distances.* In 6th Scandinavian Conference on Image Analysis, pages 393–400, Oulu, Finland, 1989.

[Forsyth 03] D.A. Forsyth & J. Ponce. *Computer Vision - a Modern Approach.* Prentice-Hall, 2003.

[Fouard 05] C. Fouard & G. Malandain. *3D Chamfer distances and norms in anisotropic grids.* Image and Vision Computing, vol. 23, pages 143–158, 2005.

[Fournier 01] Cord M. Philipp-Foliguet S. Fournier J. *Retin : A content-based image indexing and retrieval system.* IEEE Pattern Analysis and Applications, vol. 4, no. 2, 3, pages 153–173, 2001.

[Fournier 02] J. Fournier. *Indexation d'images par le contenu et recherche interactive dans les bases généralistes.* Thèse de l'Université de Cergy-Pontoise, 2002.

[Gagalowicz 83] A. Gagalowicz. *Vers un modèle de texture.* Thèse d'état, Université Pierre et Marie Curie, Paris VI, 1983.

[Galmar 05] E Galmar & B. Huet. *Méthode de segmentation par graphe pour le suivi de régions spatio-temporelles.* CORESA 2005, 10èmes journées Compression et représentation des signaux audiovisuels, 2005.

[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.

[Gong 98] Proietti G. Faloutsos-C. Gong Y. *Image indexing and retrieval based on human perceptual color clustering.* Proc. of International Conference on Computer Vision and Pattern Recognition, 1998.

[Gong 03] Y. Gong & X. Liu. *Video Summarization and Retrieval Using Singular Value Decomposition.* ACM Multimedia Systems Journal, vol. 9, pages 157–168, 2003.

[Guironnet 06] M. Guironnet. *Méthodes de résumés de vidéo à partir d'informations bas-niveau, du mouvement de caméra ou de l'attention visuel.* Thèse de l'Université Joseph Fourier, Grenoble I, 2006.

[Hagedoorn 99] M. Hagedoorn & R. Veltkamp. *Measuring Resemblance of Complex Patterns.* Lecture Notes in Computer Science, vol. 1568, pages 286–298, 1999.

[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, 1997.

[Hirata 96] T. Hirata. *A unified linear-time algorithms for computing distance maps.* Information Processing Letters, vol. 58, no. 3, pages 129–133, 1996.

[Hu 62] M. Hu. *Visual Pattern Recognition by Moment Invariants.* IRE Transactions on Information Theory, pages 179–187, 1962.

[Huang 94] C.T. Huang & O.R. Mitchell. *A euclidean distance transform using greyscale morphology decomposition.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 4, pages 443–448, 1994.

[Huang 97] J. Huang, S. Kumar, M. Mitra, W.J. Zhu & R. Zabih. *Image indexing using color correlograms.* Proc. of Conference on Computer Vision and Pattern Recognition, 1997.

[Huttenlocher 93] D.P. Huttenlocher, G.A. Klanderman & W.J. Rucklidge. *Comparing Images using the Hausdorff Distance.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 9, pages 850–863, 1993.

[Hyvärinen 01] A. Hyvärinen, J. Karhunen & E. Oja. *Independent Component Analysis.* John Wiley and Sons, 2001.

[Ikonen 05] L. Ikonen & P. Toivanen. *Shortest routes on varying height surfaces using gray-level distance transforms.* Image and Vision Computing, vol. 23, 2005.

[Ikonen 07] L. Ikonen & P. Toivanen. *Distance and nearest neighbor transforms on gray-level surfaces.* Pattern Recognition Letters, vol. 28, 2007.

[Ionescu 03] B. Ionescu, D. Coquin & P. Lambert. *Reconnaissance de gestes dynamiques de la main.* 19ème colloque sur le traitement du signal et des images (GRETSI'03), Paris, France, vol. III, pages 22–25, 2003.

[Ionescu 05] B. Ionescu, D. Coquin, P. Lambert & V. Buzuloiu. *Dynamic Hand Gesture Recognition Using the Skeleton of the Hand.* EURASIP : Journal on Applied Signal Processing, vol. 2005, no. 13, pages 2101–2109, 2005.

[Ionescu 06] 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 07] B. Ionescu. *Caractérisation Symbolique de Séquences d'Images : Application aux film d'Animation.* Thèse de doctorat de l'Université de Savoie, Mai, 2007.

[Issa 96] I. Issa & Ph. Bolon. *Adaptive weighted d_α filter.* in European Signal Processing Conference, vol. III, pages 1921–1924, Trieste, Italy, 1996.

[Jacobs 00] Weinshall D. Jacobs D.W. *Classification with nonmetric distances : image retrieval and class representation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 6, pages 583–600, 2000.

[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.

[Jiao 06] J. Jiao, Y. Zhang & M. Helander. *A Kansei mining system for affective design.* Expert Systems with Applications, vol. 30, 2006.

[Joly 05] P. Joly. *Enjeux d'une normalisation pour l'indexation des contenus multimédia.* Paradigmes et enjeux de l'informatique, Hermès Science Publications, pages 157–169, 2005.

[Jonker 96] P.P. Jonker & O. Vermeij. *On skeletonization in 4D images.* in Proceedings of Advances in Structural and Syntactical Pattern Recognition, P.Perner, P. Wang and A. Rosenfeld Editions, Springer-Verlag, Berlin/New-York, pages 79–89, 1996.

[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, 2005.

[Kharbouche 05] S. Kharbouche, P. Vannoorenberghe, C. Lecomte & P. Miche. *Histogramme spatiaux couleur optimisés pour l'indexation d'images par le contenu.* Actes du 20ème colloque GRETSI : Traitement du signal et des images, Louvain-la-Neuve, Belgique,, 2005.

[Kiselman 96] C. Kiselman. *Regularity properties of distance transformations in image analysis.* Computer Vision and Image Understanding, vol. 64, pages 390–398, 1996.

[Kraemer 06] P. Kraemer, J. Benois-Pineau & J.P. Domenger. *Scene Similarity Measure for Video Content Segmentation in the Framework of a Rough Indexing Paradigm.* International Journal of Intelligent Systems, Wiley, vol. 21, no. 7, pages 765–783, 2006.

[Lee 97] Y.H. Lee, S.J. Horng, T.W. Kao & Y.J. Chen. *Parallel computation of the Euclidean distance transform on the mesh of trees and hypercube compute.* Computer Vision and Image Understanding, vol. 68, no. 1, pages 109–119, 1997.

[Leonard 91] D.M. Leonard. *Analysis of object in bianry images.* Technical report, NASA, 1991.

[Leymarie 92] F. Leymarie & M.D. Levine. *Fast raster scan distance propagation on the discrete rectangular lattice.* Computer Vision Graphics and Image Processing, vol. 55, no. 1, pages 84–94, 1992.

[Li 97] C.S. Li & V. Castelli. *Deriving texture feature set for content-based retrieval of satellite image database.* International Conference on Image Processing, 1997.

[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.

[Li 06] J. Li & J.Z. Wang. *Real-time Computerized Annotation of Pictures.* Proceedings of the ACM Multimedia Conference, Santa Barbara, Californie, 2006.

[Liu 96] F. Liu & R.W. Picard. *Periodicity, directionality and randomness : Wold features for image modeling and retrieval.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 7, 1996.

[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.

[Maurer 03] C.R. Jr. Maurer, Rensheng Qi & V. Raghavan. *A Linear Time Algorithm for Computing Exact Euclidean Distance Transforms for Binary Images in Arbitrary Dimensions*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 2, pages 265–270, 2003.

[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, pages 1191–1194, 2000.

[Medioni 05] G. Medioni & Sing Bing Kang. *Emergins Topics in Computer Vision*. IMSC Press Multimedia Series, Prentice-Hall PTR, vol. 8, 2005.

[Mehtre 97] B.M. Mehtre, M.H. Kankanhalli & W.F. Lee. *Shape Measure for Content Based Image Retrieval : A Comparison*. Information Processing and Management, vol. 33, no. 3, pages 319–337, 1997.

[Meijster 00] A. Meijster, J. Roerdink & W.H. Hesselink. *A general algorithm for computing distance transforms in linear time*. Math. Morphology and its Applications to Image and Signal Processing, pages 331–340, 2000.

[Mezaris 04] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris & M.G. Strintzis. *Real-Time Compressed-Domain Spatiotemporal Segmentation and Ontologies for Video Indexing and Retrieval*. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 5, 2004.

[Michal 95] I. Michal, P.Anandan & S. Hsu. *Mosaic Based Representations of Video Sequences and Their Applications*. Computer Vision, pages 605–611, 1995.

[Montanari 68] U. Montanari. *A method for obtaining skeleton using a quasi-euclidean distance*. Journal of ACM, vol. 15, pages 600–624, 1968.

[Mulhem 03] P. Mulhem, J. Gensel & H. Martin. *Adaptive Video Summarization*. in Handbook on Video Databases, CRC Press, Chapter 11, pages 279–298, 2003.

[Mullikin 92] J.C. Mullikin. *The vector distance transform in two and three dimensions*. Computer Vision Graphics and Image Processing, vol. 54, no. 6, pages 526–535, 1992.

[Niblack 93] W. Niblack & M. Flikner. *Find me the pictures that look like this : IBM's Image Query Project*. Advanced Imaging, vol. 8, 1993.

[Ott 05] L. Ott. *Résumé Automatique de Films d'Animation*. Rapport de fin d'étude, LISTIC, ESIA, juin, Annecy, France 2005.

[Ott 07] L. Ott, P. Lambert, B. Ionescu & D. Coquin. *Animation Movie Abstraction : KeyFrame Adaptative Selection based on Color Histogram Filtering*. Computational Color Imaging Workshop, Modena, Italy, 2007.

[Ounis 98] I. Ounis & M. Pasça. *RELIEF : Combining expressivenss and rapidity into a single system*. ACM SIGIR, pages 266–274, Melbourne, Australia, 1998.

[Paglieroni 92] D.W. Paglieroni. *Distance transforms : properties and machine vision applications*. Computer Vision Graphics and Image Processing, vol. 54, no. 1, pages 56–74, 1992.

[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, USA, 2001.

[Pass 96] G. Pass, R. Zabih & J. Miller. *Comparing images using color coherence vectors*. ACM International Multimedia Conference, pages 65–73, 1996.

[Pentland 96] A. Pentland, R. Picard & S. Sclaroff. *Photobook : Tools for content-based manipulation of image databases.* International Journal on Computer Vision, vol. 18, no. 3, 1996.

[Petrovic 07] V. Petrovic. *Subjective tests for image fusion evaluation and objective metric validation.* Information Fusion, vol. 8, pages 208–216, 2007.

[Peyrard 05] N. Peyrard & P. Bouthemy. *Motion-Based Selection of Relevent Video Segments for Video Summarization.* Multimedia Tools and Applications, Springer Science, vol. 26, 2005.

[Pfeiffer 96] S. Pfeiffer, S. Fisher & W. Effelsberg. *Automatic Audio Content Analysis.* ACM Int. Conference Multimedia, novembre, Boston, USA, 1996.

[Philipp-Foliguet 05a] S. Philipp-Foliguet & L. Guigues. *Critère multi-échelles d'évaluation de la segmentation.* GRETSI, 2005.

[Philipp-Foliguet 05b] S. Philipp-Foliguet, M.B. Vieira & M. Lekkat. *Recherche d'images par appariement d'ensembles de régions floues.* Information, Interaction, Intelligence, vol. 5, no. 2, pages 9–40, 2005.

[Piccard 96] R.W. Piccard, A. Pentland & S. Sclaroff. *Photobook : Content-based manipulation of image databases.* International Journal of Computer Vision, vol. 18, no. 3, pages 233–254, 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, 1997.

[Piriou 06] F. Piriou, P Bouthemy & J.F. Yao. *Recognition of Dynamic Video Contents with Global Probabilistic Models of Visual Motion.* IEEE Transactions on Image Processing, vol. 15, no. 11, 2006.

[Quack 04] T. Quack, U. Monich, L. Thiele & B. Manjunath. *Cortina : A system for large-scale, content-based web image retrieval.* In ACM Multimedia, 2004.

[Quénot 99] G. Quénot & P. Mulhem. *Two Systems for Temporal Video Segmentation.* CBMI'99, pages 187–193, Toulouse, France, 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.

[Ragnemalm 92] I. Ragnemalm. *Neigborhoods for distance transformations using ordered propagation.* Computer Vision Graphics and Image Processing, vol. 56, no. 3, pages 399–409, 1992.

[Ragnemalm 93] I. Ragnemalm. *The Euclidean distance transformation in arbitrary dimensions.* Pattern Recognition Letters, vol. 14, pages 883–888, 1993.

[Remy 00] E. Remy & E. Thiel. *Optimizing 3D chamfer mask with norm constraints.* in Proceedings of International workshop on Combinatorial Image Analysis, pages 39–56, Caen, France 2000.

[Rhodes 92] F. Rhodes. *Discrete euclidean metrics.* Pattern Recognition Letters, vol. 13, pages 623–628, 1992.

[Rosenberger 06] C. Rosenberger. *Contribution à l'évaluation d'algorithmes de traitement d'images.* Habilitation à Diriger des Recherches, Université d'Orléans, 2006.

[Rosenfeld 66] A. Rosenfeld & J.L. Pfaltz. *Sequential operations in digital picture processing.* Journal ACM, vol. 13, pages 471–494, 1966.

[Rosenfeld 68] A. Rosenfeld & J.L. Pfaltz. *Distance function on digital pictures.* Pattern Recognition, vol. 1, pages 33–61, 1968.

[Saito 94] T. Saito & J.I. Toriwaki. *New algorithms for Euclidean distance transformation of an n-dimensional digitized picture with applications.* Pattern Recognition, vol. 27, pages 1551–1565, 1994.

[Santini 01] S. Santini, A. Gupta & R. Jain. *Emergent Semantics though interaction in Image Databases.* IEEE transactions on Knowledge and Data Engineering, vol. 13, no. 3, pages 332–351, 2001.

[Shih 92] F.Y. Shih & O.R. Mitchel. *A mathematical morphology approach to Euclidean distance transformation.* IEEE Transaction on Image Processing, vol. 1, no. 2, pages 197–204, 1992.

[Simand 05] I. Simand & J.M. Jolion. *Représentation d'images par chaînes de symboles : application à la recherche par le contenu.* Actes du 20ème colloque GRETSI : Traitement du signal et des images, Louvain-la-Neuve, Belgique,, vol. 2, pages 925–928, 2005.

[Simon 84] J.C. Simon. *Reconnaissance des formes par algorithmes.* Editions Masson, 1984.

[Sintorn 01] I.M. Sintorn & G. Borgefors. *Weighted distance transforms in rectangular grids.* 11th International Conference on Image Analysis and Processing, pages 322–326, Palermo, Italy, 2001.

[Sintorn 02] I.M. Sintorn & G. Borgefors. *Weighted distance transforms for images using elongated voxel grids.* Proc. 10th Discret Geometry for Computer Imagery, pages 244–254, Bordeaux, France, 2002.

[Sintorn 04] I.M. Sintorn & G. Borgefors. *Weighted distance transforms for volume images digitized in elongated voxel grids.* Pattern Recognition Letters, vol. 25, no. 5, pages 571–580, 2004.

[Smeulders 00a] 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, 2000.

[Smeulders 00b] Worring M. Santini-S. Gupta A. Jain R. Smeulders A.W.M. *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, 2000.

[Smith 97] M. Smith & T. Kanade. *Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques.* IEEE Computer Vision and Pattern Recognition, Puerto Rico, 1997.

[Soille 99] P. Soille. *Morphological Image Analysis : Principles and Applications.* Springer Verlag, 1999.

[Sowa 84] J.F. Sowa. *Conceptual Structures : Information Processing in Mind and Machines.* Addison Wesley, Reading(MA), USA, 1984.

[Stoica 98] R. Stoica, J. Zerubia & J.M. Francos. *Indexation et recherche dans une base de données multimédia grâce à une modélisation paramétrique de texture utilisant la décomposition de Wold 2D.* Rapport de Recherche 3594, INRIA, 1998.

[Strand 05] R. Strand & G. Borgefors. *Distance transforms for three-Dimensional Grids with Non-Cubic voxels.* Computer Vision and Image Understanding, vol. 100, no. 3, pages 294–311, 2005.

[Stricker 96] Dimai A. Stricker M. *Color indexing with weak spatial constraints.* Storage and Retrieval for Image and Video Databases, SPIE Proc., vol. 2670, pages 29–40, 1996.

[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.

[Suykens 03] Horvath G. Basu-S. Micchelli C. Vandewalle J. (Eds.) Suykens J.A.K. *Advances in Learning Theory : Methods, Models and Applications.* NATO Science Series III : Computer and Systems Sciences, IOS Press Amsterdam, vol. 190, page 436, 2003.

[Svensson 02] S. Svensson & G. Borgefors. *Digital distance transforms in 3D images using information from neighbourhoods up to 5x5x5.* Computer Vision and Image Understanding, vol. 88, pages 24–53, 2002.

[Swain 91] Ballard D.H. Swain M.J. *Color indexing.* International Journal of Computer Vision, vol. 7, no. 1, pages 11–22, 1991.

[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.

[Thiel 92a] E. Thiel & A. Montanvert. *Etude et amélioration des distances du chanfrein pour l'analyse d'images.* Technique et Science Informatiques, vol. 11, no. 4, pages 9–41, 1992.

[Thiel 92b] E. Thiel & A. Montanvert. *Chamfer masks : discrete distance functions, geometrical properties and optimization.* 11th International Conference on Pattern Recognition, vol. 3, no. 4, pages 244–247, The Hague, The Netherlands, 1992.

[Thiel 01] E. Thiel. *Géométrie des distances de Chanfrein.* Habilitation à diriger des Recherches, Université de la Méditerranée, Aix-Marseille II, 2001.

[Truong 06] B.T. Truong & S. Venkatesh. *Video Abstraction : A Systematic Review and Classification.* ACM Transactions on Multimedia Computing, Communications and Applications, vol. 3, no. 1, 2006.

[Uchihashi 99] S. Uchihashi & J. Foote. *Video Manga : Generating Semantically Meaningful Video Summaries.* ACM Multimedia'99, Orlando, USA, 1999.

[Unser 95] M. Unser. *Texture classification and segmentation using wavelet frame.* IEEE transaction on Image Processing, vol. 4, no. 11, 1995.

[Valet 07] L. Valet, D. Coquin, S. Jullien & S. Teyssier. *A 3D image-segmented evaluation procedure in a cooperative fusion system context.* 10textsuperscriptth International Conference on Information Fusion, Quebec, Canada, July, 2007.

[Verwer 91] B. Verwer. *Local distances for distance transformations in two and three dimensions.* Pattern Recognition Letters, vol. 12, pages 671–682, 1991.

[Viallet 02] J.E. Viallet & O. Bernier. *Face detection for Video summaries.* CIVR'02 : the Challenge of Image and Video Retrieval, Londres,, 2002.

[Vincent 91] L. Vincent. *Exact Euclidean distance function by chain propagations.* In Computer Vision and Pattern Recognition, pages 520–525, 1991.

[Visibone 06] Visibone. *Webmaster Palette.* http ://www.visibone.com/colorlab, 2006.

[Wan 04] K. Wan, J. Wang, C. Xu & Q. Tian. *Automatic Sports Highlights Extraction with Content Augmentation.* Advances in Multimedia Information Processing, PCM, pages 19–26, LNCS, 2004.

[Wilson 97] D.L. Wilson, A.J. Baddeley & R.A. Owen. *A new metrics for grey scale image comparison.* International Journal on Computer Vision, vol. 24, pages 5–18, 1997.

[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.

[Yamada 84] H. Yamada. *Complete euclidean distance transformation by parallel operation.* 7th International Conference on Pattern Recognition, pages 69–71, 1984.

[Yang 99] S. Yang, M. Nagamachi & S. Lee. *Rule-based inference model for the Kansei Engineering System.* International Journal of Industrial Ergonomics, vol. 24, 1999.

[Ye 88] Qin-Zhong Ye. *The signed Euclidean distance transform and its applications.* in 9th ICPR, pages 495–499, 1988.

[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.

[Zamperoni 96] P. Zamperoni & V. Starovoitov. *On measures of dissimilarity between arbitrary gray-scale images.* International Journal of shape Modeling, vol. 2, no. 2,3, pages 189–213, 1996.

[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 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.

[Zouagui 04] T. Zouagui, H. Benoit-Cattin & C. Odet. *Image segmentation functional model.* Pattern Recognition, vol. 37, pages 1785–1795, 2004.

Troisième partie

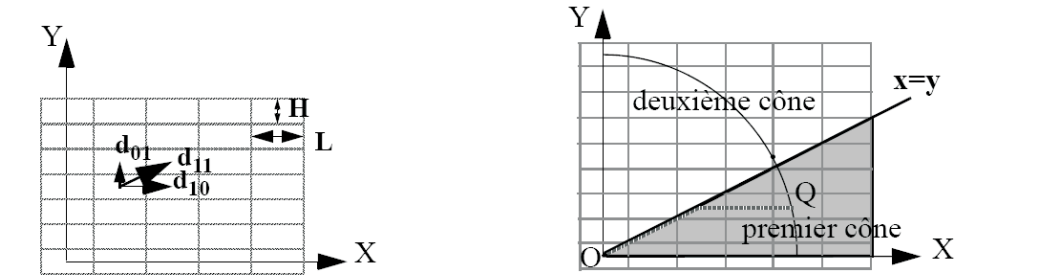
Annexes

Annexe A

Optimisation d'un opérateur local de distance 3×3

Détail du calcul de l'optimisation d'un opérateur local de distance 3×3.

En maillage rectangulaire, l'opérateur de taille 3×3 est caractérisé par trois coefficients déterminant les directions principales notées d_{10}, d_{11}, et d_{01}, comme le montre la figure A.1a. Nous pouvons remarquer que les déplacements horizontaux et verticaux sont différents. D'après Montanari [Montanari 68], il existe toujours un chemin minimal formé, au plus, de deux segments de droite discrets entre les pixels $O(0,0)$ et $Q(x,y)$, quelle que soit la position de Q. Nous considérons que le pixel Q appartient au premier quadrant, donc $x \geq 0$ et $y \geq 0$, ce qui n'enlève rien à la généralité du problème, puisque les autres quadrants sont équivalents. Les résultats se retrouveront par symétrie soit, par rapport à l'axe OX soit, par rapport à l'axe OY. Les trois directions principales divisent le quart du cercle en deux cônes. Le premier est délimité par les directions principales d_{10} et d_{11}. Le deuxième est celui situé entre les directions principales d_{11}, d_{01} (figure A.1b).



(a) Déplacements élémentaires (b) Les cônes du premier quadrant du cercle

Fig. A.1: Opérateur local de distance.

Nous suivons la démarche adoptée par Borgefors [Borgefors 86] pour déterminer les coefficients réels optimaux, mais, contrairement à Borgefors qui a fait son calcul pour le pixel Q décrivant une droite verticale (ce qui accentue l'erreur par rapport à la distance euclidienne sur la diagonale), nous considérons tout comme Verwer [Verwer 91], que le pixel Q décrit une trajectoire circulaire de rayon $R = \sqrt{(Lx)^2 + (Hy)^2}$ de manière à répartir l'erreur par rapport à la distance euclidienne sur l'ensemble des directions possibles.

Coefficients réels

La distance locale $d_L(O,Q)$ entre l'origine $O(0,0)$ et le pixel $Q(x,y)$ est calculée dans chaque cône du quart de cercle en fonction des coefficients correspondant aux déplacements élémentaires délimitant ce cône.

Dans le premier cône, pour $0 \leq y \leq R/\sqrt{L^2+H^2}$ la distance locale $d_L(O,Q)$ est donnée par l'expression suivante :

$$d_L(O,Q) = d_{10}.(x-y) + y.d_{11} = x.d_{10} + y.(d_{11} - d_{10}) \tag{A.1}$$

Dans le deuxième cône, pour $0 \leq x \leq R/\sqrt{L^2+H^2}$ la distance locale $d_L(O,Q)$ est donnée par l'expression suivante :

$$d_L(O,Q) = x.d_{11} + d_{01}.(y-x) = x.(d_{11} - d_{01}) + y.d_{01} \tag{A.2}$$

Le but est de trouver les coefficients d_{10}, d_{11}, et d_{01}, qui minimisent l'erreur absolue maximale réalisée pour Q. Dans le premier cône, l'erreur $E = d_L - d_E$ est une fonction continue de y et dérivable par morceaux. Pour $0 \leq y \leq R/\sqrt{L^2+H^2}$, l'erreur a pour expression :

$$E_1(y) = y.(d_{11} - d_{10}) + \frac{d_{10}}{L}.\sqrt{R^2 - (Hy)^2} - R \tag{A.3}$$

et dans le deuxième cône, l'erreur est une fonction continue de x et dérivable par morceaux. Pour $0 \leq x \leq R/\sqrt{L^2+H^2}$, l'erreur a pour expression :

$$E_2(x) = x.(d_{11} - d_{01}) + \frac{d_{01}}{H}.\sqrt{R^2 - (Lx)^2} - R \tag{A.4}$$

L'erreur est donc extrémale aux extrémités des intervalles ou lorsque la dérivée s'annule. Dans le premier cône, l'erreur aux bornes de l'intervalle est calculée à partir de l'équation A.3, deux expressions sont donc obtenues, E_{11} et E_{12} :

pour $y = 0$

$$E_{11} = R\left(\frac{d_{10}}{L} - 1\right) \tag{A.5}$$

pour $y = R/\sqrt{L^2+H^2}$

$$E_{12} = R\left(\frac{d_{11}}{\sqrt{L^2+H^2}} - 1\right) \tag{A.6}$$

L'expression de la dérivée de la fonction de l'erreur donnée par l'équation A.3 est la suivante :

$$\frac{\partial E_1}{\partial y} = (d_{11} - d_{10}) - 2d_{10}.\frac{y.H^2}{2L\sqrt{R^2 - y^2H^2}} \tag{A.7}$$

L'erreur est extrémale lorsque cette dérivée partielle s'annule, c'est à dire pour $y = y_0$ donnée par :

$$y_0 = \frac{R.(d_{11} - d_{10})}{H^2\sqrt{\frac{d_{10}^2}{L^2} + \frac{(d_{11}-d_{10})^2}{H^2}}} \tag{A.8}$$

d'où la valeur extrémale de l'erreur $E_{13} = E_1(y_0)$ donnée par l'expression suivante :

$$E_{13} = R.\left(\sqrt{\frac{d_{10}^2}{L^2} + \frac{(d_{11}-d_{10})^2}{H^2}} - 1\right) \tag{A.9}$$

De la même façon, on calcule l'erreur extrémale aux bornes du deuxième cône. Elle est donc calculée à partir de l'équation A.4. On obtient ainsi les deux expressions E_{21} et E_{22} :

pour $x = R/\sqrt{L^2+H^2}$

$$E_{21} = R\left(\frac{d_{11}}{\sqrt{L^2+H^2}} - 1\right) = E_{12} \tag{A.10}$$

pour $x = 0$

$$E_{22} = R\left(\frac{d_{10}}{H} - 1\right) \tag{A.11}$$

et la dérivée s'annule pour $x = x_0$

$$\frac{\partial E_2(x)}{\partial x} = 0 \iff x_0 = \frac{R.(d_{11} - d_{01})}{L^2\sqrt{\frac{(d_{11}-d_{01})^2}{L^2} + \frac{d_{01}^2}{H^2}}} \tag{A.12}$$

En reportant cette valeur dans l'équation A.4, on obtient la troisième expression de l'erreur dans ce deuxième cône :

$$E_{23} = R.\left(\sqrt{\frac{(d_{11} - d_{01})^2}{L^2} + \frac{d_{01}^2}{H^2}} - 1\right) \tag{A.13}$$

On cherche maintenant la valeur optimale des coefficients d_{10}, d_{11}, et d_{01} en minimisant l'erreur. En étudiant les variations de l'erreur $E_1(y)$ et de l'erreur $E_2(x)$, on s'aperçoit que c'est dans le deuxième cône que l'erreur est la plus importante. Elle est d'autant plus importante que le rapport L/H augmente. Par contre $E_1(y) = E_2(x)$ pour $L = H$. On va donc s'intéresser à la minimisation de l'erreur dans le second cône. Une façon de minimiser l'erreur est d'écrire que $E_{21} = -E_{23} = E_{22}$ [Rosenfeld 66] et [Borgefors 86]. En résolvant ces équations on obtient les expressions des coefficients :

$$d_{01} = \frac{-2H + 2H\sqrt{1+\lambda}}{\lambda} \tag{A.14}$$

avec

$$\lambda = \frac{1}{L^2}\left(\sqrt{L^2 + H^2} - H\right)^2 \tag{A.15}$$

et

$$d_{11} = \sqrt{L^2 + H^2}.\frac{d_{01}}{H} \tag{A.16}$$

la valeur optimale de d_{10} s'obtient en résolvant l'équation $E_{11} = -E_{22}$, d'où

$$d_{10} = L.\frac{d_{01}}{H} \tag{A.17}$$

L'erreur maximale notée E_{max} entre la distance locale et la distance euclidienne est proportionnelle au rayon R de la trajectoire et égale à E_{22} donnée par l'expression de l'équation A.11, on peut donc écrire :

$$E_{max} = \mid 1 - \frac{d_{01}}{H} \mid .R \tag{A.18}$$

ou encore

$$E_{max} = |1 - \frac{-2 + 2\sqrt{1+\lambda}}{\lambda}|.R \tag{A.19}$$

Cas particuliers : **maillage carré** ($L = H = 1$) on obtient alors les coefficients réels suivants :

$$d_{10} = d_{01} \simeq 0.9604 \qquad d_{11} \simeq 1.3583 \tag{A.20}$$

ces valeurs sont sensiblement égales à celles obtenues par l'optimisation sur une trajectoire rectiligne [Borgefors 86] dont on rappelle les valeurs :

$$d_{10} = d_{01} \simeq 0.9551 \qquad d_{11} \simeq 1.3507 \tag{A.21}$$

Annexe B

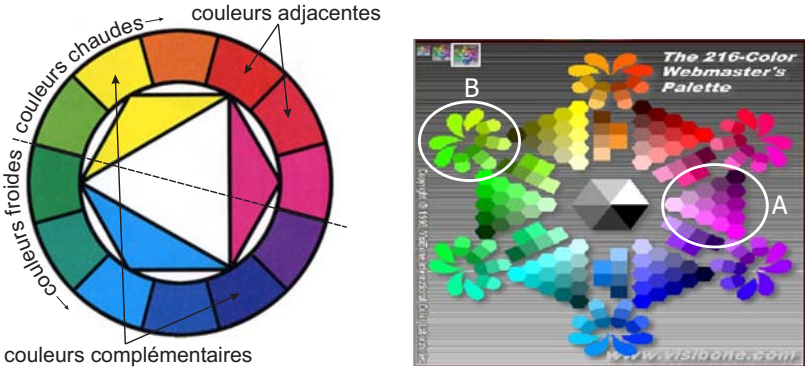
Les gamuts sémantiques

Nous donnons la définition des termes employés dans les gamuts sémantiques.

G^p : **Gamut des plans** :
- **R** : rythme. Ce terme est associé au rythme de la séquence défini à partir de la vitesse moyenne $\overline{v}_T$ des changements de plans, calculé sur la séquence entière, par tranche de T secondes (ici $T = 5$ secondes).
- **A** : action. Ce terme est calculé par le pourcentage de segments d'action par rapport à la séquence entière.
- **M** : mystère : Ce terme est proportionnel au nombre de transitions de type "dissolves" ou "fades" présentes dans la séquence entière.
- **E** : explosivité. Ce terme est proportionnel au nombre de transitions de type "changement bref de couleur" présentes dans la séquence.

G^c : **Gamut des propriétés couleurs** :
- **CL** : couleurs claires. Ce paramètre est lié à la proportion de couleurs claires présentes dans la séquence (sur une palette fixe de 216 couleurs).
- **Ft** : couleurs saturées. Ce paramètre est lié à la proportion de couleurs saturées présentes dans la séquence.
- **Cd** : couleurs chaudes. Ce paramètre est lié à la proportion de couleurs chaudes présentes dans la séquence.
- **Fc** : couleurs foncées. Ce paramètre est lié à la proportion de couleurs foncées présentes dans la séquence.
- **Fb** : couleurs faiblement saturées. Ce paramètre est lié à la proportion de couleurs faiblement saturées présentes dans la séquence.
- **Fr** : couleurs froides. Ce paramètre est lié à la proportion de couleurs froides présentes dans la séquence.

G^{rl} : **Gamut de la richesse couleur et des relations entre couleurs** :
- **Vr** : variété des couleurs qui est lié au pourcentage de couleurs utilisées sur l'ensemble de la palette contenant 216 couleurs.
- **Dv** : diversité des couleurs.
- **Aj** : couleurs adjacentes.
- **Cp** : couleurs complémentaires.



(a) (b)

FIG. B.1: (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 : mélanges entre les variations de couleurs élémentaires)

Résumé

L'essor du traitement de l'image et son utilisation dans de nombreux secteurs d'activité aussi différents que la biométrie, la médecine ou la vidéosurveillance ont donné naissance à de nombreuses bases de données d'images fixes ou de séquences d'images. Le support informatique des images permet notamment de les comparer. Cela peut être utile à plusieurs titres : pour visualiser les différences ou les similitudes entre les images, pour visualiser l'évolution dans le temps d'un phénomène en comparant des images prises à des instants différents, pour exploiter la redondance des informations recherchées et permettre ainsi un traitement plus robuste, pour rechercher des images ou un genre de films dans une base de données, … Mais selon le type d'image et l'application visée, les techniques de comparaison sont bien différentes.

Pour comparer, il faut une distance. Notre contribution porte sur la définition et l'implémentation de cette distance. Nous avons proposé une mesure de dissimilarité permettant de comparer l'effet de certains traitements associés aux images comme le filtrage, et un cycle de compression/décompression. Le calcul de la dissimilarité est réalisé en utilisant une extension de la distance de Baddeley, calculée à l'aide d'opérateurs locaux de distance en maillage parallélépipédique. Cette mesure permet de tenir compte simultanément, des différences entre niveaux de gris ou couleurs et des éventuelles déformations géométriques des structures présentes dans l'image. Nous avons également mis en évidence qu'il était possible de comparer et de caractériser, par leur signature, différentes déformations que peut subir une image. La comparaison est présentée ici comme une application de cette distance avec les différentes versions selon le type d'images (images binaires, à niveaux de gris, puis couleurs).

L'extension à l'analyse de séquences d'images pour la construction de résumés est également présentée. Nous avons proposé plusieurs méthodes de génération automatique de résumés construits à partir de la sélection d'images clés issues de la séquence d'images. Mais apprécier la qualité d'un résumé est une tâche difficile et subjective. Elle reste en effet liée à la manière de percevoir de chacun d'entre nous et à l'objectif visé. Enfin, nous proposons une représentation visuelle compacte par la définition de « gamuts sémantiques » permettant de se faire une idée globale de l'ensemble des caractéristiques d'une séquence d'images. Ainsi, la tâche de comparaison des différentes séquences s'en trouve simplifiée par la comparaison des formes de ces gamuts.

Mots-clé : Comparaison d'images, Mesures de dissimilarité, Comparaison de séquences d'images.

Abstract

The increasing popularity of image and its use in many activity sectors; as different as biometrics, medicine or video surveillance has given birth to many databases of still images or animated movies. The use of digital images allows them to be compared. This can be useful on many accounts: to visualise differences or similarities between images; or to visualise the evolution in time of a phenomenon by comparing images taken at different instances; also to exploit the redundancy of information that is being looked for and to allow a more robust treatment, it also allows image retrieval or a category of film on a database,… but depending on the type of image and its intended use the methods of comparison are totally different.

In order to compare one must take a distance. Our contribution covers the definition and the implementation of this distance. We have proposed to measure the dissimilarities, thus allowing the comparison of the effect of certain treatments applied to images like filters and a cycle of compression/decompression or the calculation of Baddeley's distance, using 3D local distance operators adapted to a parallelepipedic grid. This measure allows us to take into account simultaneously the difference in grey scale or colour and eventually the displacement of the structure (if present) in an image. The comparison is presented here as an application of this distance, with the different versions, according to the type of images, grey scale or colour.

The extension of the analysis of animated movies for the construction of summaries is presented. We have proposed many methods of automatic generation of summaries from the selection of key images taken from the movie. Appreciating the quality of a summary is a difficult and subjective task. It is effectively bound to every individual's manner of perceiving the images and the goal. Lastly, we propose a compact visual representation through the definition of "semantic gamuts" allowing one to make a global idea of the characteristics of animated movies. Therefore, the job of comparing different movies is simplified by the comparison, of the shapes of these gamuts.

Key words : Image comparison, Dissimilarity measure, Video comparison.