N$^o$ d'ordre: 2657

**THÈSE**

Présentée devant

L'UNIVERSITÉ DE RENNES I

pour obtenir

le grade de: **Docteur de l'Université de Rennes I**

Mention: Traitement du Signal

par

Emmanuel VENEAU

Équipe d'accueil: VISTA (IRISA, RENNES)

École doctorale: Mathématiques, Informatique, Signal, Électronique et Télécommunications

Composante Universitaire: SPM

Titre de la thèse:

# Macro-segmentation multi-critère et classification de séquences par le contenu dynamique pour l'indexation vidéo

Soutenue le 26 février 2002, devant la commission d'examen composée de:

COMPOSITION DU JURY

| | | | |
|---|---|---|---|
| M. | Claude | LABIT | Président |
| M. | Jean-Michel | JOLION | Rapporteur |
| M. | Bernard | MÉRIALDO | Rapporteur |
| M. | Patrick | BOUTHEMY | Directeur de thèse |
| M. | Laurent | VINET | Encadrant INA |
| M. | Ferran | MARQUÉS | Examinateur |
| Mme | Michèle | WAUTELET | Membre invité |

"Une société capitaliste exige une culture assise sur des images. Elle doit fournir de la distraction en grosse quantité afin de stimuler la consommation et d' « analgésier » les blessures de classe, de race et de sexe. Et elle doit rassembler des informations innombrables, afin de mieux exploiter les ressources naturelles, d'accroître la productivité, de maintenir l'ordre, de faire la guerre, de donner du travail aux bureaucrates."
Susan Sontag, *Sur la photographie*, Éditions du Seuil, collection 10/18, p. 208, 1983.

"Tellement tentant de vouloir distribuer le monde entier selon un code unique ; une loi universelle régirait l'ensemble des phénomènes : deux hémisphères, cinq continents, masculin et féminin, animal et végétal, singulier pluriel, droite gauche, quatre saisons, cinq sens, six voyelles, sept jours, douze mois, vingt-six lettres. Malheureusement ça ne marche pas, ça n'a même jamais commencé à marcher, ça ne marchera jamais. N'empêche que l'on continuera encore longtemps à catégoriser tel ou tel animal selon qu'il a un nombre impair de doigts ou des cornes creuses."
Georges Perec : "Penser/Classer", *Le Genre humain* n°2, pp. 111-127, 1982.

## Remerciements

Je tiens à remercier Claude Labit, pour avoir présidé le jury de thèse, et Jean-Michel Jolion ainsi que Bernard Mérialdo qui ont accepté d'être les rapporteurs de ce travail. Mes remerciements vont également à Ferran Marqués et Michèle Wautelet qui ont participé au jury comme examinateurs.

Je suis reconnaissant à Patrick Bouthemy qui a su nourrir cette étude de ses idées et qui a assuré le suivi de celle-ci en dépit de l'éloignement géographique. Je suis également redevable à Bruno Bachimont et Philippe Poncin pour m'avoir donné, à l'INA, les moyens nécessaires à la réalisation des travaux, ainsi qu'à Rémi Ronfard qui a proposé ce sujet de thèse.

Je souhaite exprimer toute ma gratitude à Laurent Vinet, grâce à qui cette étude a pu être menée à terme. Qu'il soit sincèrement remercié pour ses compétences, sa confiance et son soutien, ses conseils et sa dextérité dans la difficile tâche qu'est la chasse aux *bugs*. Ce fut un réel plaisir que de travailler sous sa responsabilité technique au cours de ces années.

Je remercie également celles et ceux qui m'ont aidé sur différents aspects de mes travaux : Vincent Brunie et Jean Carrive pour leur aide inestimable ; Dominique Brault et Sandrine Lambert pour m'avoir sympathiquement accueilli respectivement au sein de l'Inathèque et du Département Droits et Archives de l'INA ; Alain Perrier pour son soutien logistique et son efficacité ; Anne Jaigu, qui m'a donné accès avec gentillesse et compétence aux lointaines ressources du Centre de Documentation de l'IRISA ; Anne Bony pour sa participation à la constitution des annotations manuelles de référence ; Odile Phidias et Edith Blin pour m'avoir secondé en de nombreux problèmes administratifs et logistiques.

Je suis aussi reconnaissant à celles et ceux qui m'ont permis d'avancer au travers d'échanges enrichissants : Ronan Fablet, pour les nombreuses et intéressantes discussions que nous avons eues, et pour les échanges scientifiques qui s'en sont suivis ; Denis Pellerin, qui m'a guidé dans le monde parfois inamical des filtres de Gabor spatio-temporels ; Jean-Michel Jolion, dont les commentaires et les encouragements ont su cristalliser certaines de mes réflexions ; Michèle Wautelet, qui m'a éclairé sur le monde de la documentation audiovisuelle, et dont les réflexions m'ont grandement fait avancer lors des travaux menés en collaboration. Qu'elle soit aussi remerciée pour son enthousiasme communicatif, pour ses critiques pertinentes, et pour sa présence amicale.

J'aimerais saluer aussi celles et ceux que j'ai cotoyés, à l'INA comme à l'IRISA : Agnès, Alexandre, Antoine, Bastien, Carine, Estelle, Étienne, Guillaume, Gwendal, Karine, Marie-Luce, Thomas, Raphaël, Véronique, Younès, et les autres. Toute mon amitié va notamment à David, Pierre et Solène qui m'ont chaleureusement accueilli, logé et nourri, lors de mes passages à Rennes. Je remercie aussi Bastien dont les multiples talents et l'humeur facétieuse furent une forme de réconfort moral. Enfin, je tiens à remercier celles qui chantent dans les couloirs de l'INA, à l'heure où il est beaucoup trop tard pour y laisser traîner ses oreilles.

Mes pensées vont aussi à ma grand-mère, mes parents, ma famille et mes amis qui m'ont entouré de leur précieuse affection au cours de toutes ces années. Merci à Marion et Xavier pour leur soutien tout à la fois dominical et vespéral lors de la rédaction du manuscrit. Ma tendresse accompagne Léonie, que je remercie aussi pour sa présence et ses relectures efficaces. Enfin, enrichi du bout de chemin parcourru à leurs côtés, qu'il me soit permis de ne pas oublier les petits habitants du boulevard Soult. *Al andar se hace el camino...*

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction générale

## Contexte de la recherche

### Un contexte institutionnel

Les travaux de cette thèse se sont déroulés dans un contexte de recherche industrielle au sein du *Groupe de Recherches Audiovisuelles et Multimédias* (GRAMM) de la *Direction de la Recherche et Expérimentation* (DRE) de l'*Institut National de l'Audiovisuel* (INA). Ceux-ci ont été menés en collaboration avec le projet *Vision Spatio-Temporelle et Active* (VISTA) de l'*Institut de Recherche en Informatique et Systèmes Aléatoires* (IRISA/INRIA Rennes), dans le cadre d'une *Convention Industrielle de Formation pour la Recherche* (CIFRE) signée avec l'*Association Nationale de la Recherche Technique* (ANRT).

L'INA est un *Établissement Public à caractère Industriel et Commercial* (EPIC) fondé en janvier 1975, suite à la réforme de l'ORTF (loi du 7 août 1974). Les travaux de cette thèse sont liés à deux des principales missions de l'INA, réaffirmées au cours des années : "[L'INA] assure la conservation des archives audiovisuelles des sociétés nationales de programme et contribue à leur exploitation (...). Il est responsable du dépôt légal des documents sonores et audiovisuels radiodiffusés et télédiffusés." [1]

Au-delà de ce cadre institutionel, la gestion de telles missions peut se traduire par quelques chiffres[2] qu'il est toujours plaisant de citer :

- un fonds de 1,5 million d'heures de radio et de télévision, soit 2,5 millions de documents, occupant plus de 80 kilomètres de rayonnage;
- plus de 80 000 heures d'archives collectées chaque année, dont la moitié au titre du dépôt légal;
- plus de 5 millions de notices rendues disponibles par les documentalistes.

Retenons de ces chiffres la masse des documents déjà collectés et présents dans le fonds, l'importance du flux entrant des nouveaux documents, et l'étendue du travail nécessaire à la rédaction des notices générées manuellement.

### Un contexte applicatif

Les recherches menées selon les problématiques de l'INA, et associées à la collaboration avec l'IRISA, ont trouvé des applications concrètes à travers la participation à deux projets.

1. article 10 de la loi n°2000-79 du 1er août 2000.
2. sources : *Au cœur du patrimoine audiovisuel à l'heure du numérique*, brochure de l'Institut National de l'Audiovisuel, 2000 et *Rapport d'activité 2000*, Institut National de l'Audiovisuel.

Il s'agit des projets[3] *Distributed Audio Visual Archives Network* (DiVAN) [Bouthemy 99a], projet européen dont l'objectif était de fournir notamment des solutions pour la numérisation, l'annotation, la gestion et le stockage des documents professionnels d'archives audiovisuelles, et *Architecture Globale pour l'Indexation et la Recherche* (AGIR), projet national RNRT, dont un des objectifs était de fournir une station de production de descriptions de contenus multimédia conformes à la norme MPEG-7, utilisable pour la mise à disposition de contenus indexés.

## Problématique générale

Nos travaux s'inscrivent dans le cadre général de l'indexation automatique des documents audiovisuels. Nous avons défini deux axes d'étude privilégiés : la segmentation temporelle d'un document vidéo et la caractérisation du contenu dynamique de celui-ci.

Dans le cadre de la segmentation temporelle, nous nous sommes intéressés à la segmentation en séquences ou macro-segmentation. Pour cela, nous fondant sur des travaux antérieurs concernant la représentation hiérarchique de l'information visuelle, nous avons proposé une méthode de regroupement de plans utilisant la distance cophénétique et des descripteurs de nature composite.

Afin de caractériser le contenu des documents, nous nous sommes placés dans un cadre d'indexation général que nous avons appliqué à un domaine particulier - le mouvement dans les documents de sport. Plus précisément, nous avons utilisé une méthode de classification par apprentissage fondée sur les machines à vecteurs de support (*Support Vector Machines* - SVM) et étudié des descripteurs extraits de séquences courtes. Pour l'extraction de ceux-ci, nous avons mis en œuvre deux familles de méthodes, le première liée aux travaux d'Aaron Bobick et de son équipe sur les images de l'historique du mouvement (*Motion History Image* - MHI), la seconde issue de l'utilisation de filtres de Gabor tridimensionnels (spatio-temporels).

## Plan du mémoire

Ce mémoire débutera par des réflexions menées sur le cadre défini par le domaine d'application de notre travail, et notamment sur les problématiques de l'indexation de documents audiovisuels. Nous verrons comment ces problématiques se retrouvent au cœur des recherches en indexation automatique. Une fois posés ces quelques jalons, nous cernerons notre sujet d'étude autour de deux grands axes : la segmentation temporelle et la caractérisation du contenu dynamique d'un document audiovisuel.

La segmentation temporelle est abordée dans la deuxième partie en termes de macro-segmentation. Après avoir défini la problématique et proposé un état de l'art, nous décrirons la méthode de macro-segmentation que nous avons développée. Nous aborderons notamment les points suivants : l'extraction et la gestion coopérative des primitives au sein d'une structuration hiérarchique, le calcul d'une mesure de cohérence entre plans, la gestion de la paramètrisation des algorithmes. Des expérimentations quantitatives et qualitatives seront détaillées.

La caractérisation du contenu dynamique de segments audiovisuels sera traitée dans la dernière partie autour de deux points principaux : l'extraction de descripteurs pour des séquences courtes et l'apprentissage d'une classification de mouvements et d'activités. Les expérimentations menées afin de valider les différents niveaux de reconnaissance envisagés seront décrites.

3. Pour plus de détails, ces projets sont présentés en ligne sur le site de l'INA aux adresses `http://www.ina.fr/Recherche/divan.fr.html` et `http://www.ina.fr/Recherche/agir.fr.html`.

Enfin, nous établirons un bilan de nos travaux en vue de conclure à la fois sur nos propositions algorithmiques, et sur l'adéquation des recherches menées avec nos objectifs et le cadre d'étude défini précédemment.

# Première partie

# Positionnement des travaux

# Un contexte applicatif

La motivation initiale de nos travaux est de développer des outils d'analyse et d'indexation automatique en mesure d'assister les documentalistes dans leur tâche au sein de la chaîne de documentation de l'INA. Nous avons naturellement procédé à un état de l'art de la recherche dans ce domaine. Si les développements issus de cette communauté scientifique nous ont paru nombreux et prometteurs, certains auteurs semblent suggérer que les systèmes et prototypes dédiés à l'indexation des documents audiovisuels peinent à remplir pleinement les tâches qui leur avaient été prescrites.

Envisager des limites à l'automatisation de l'indexation nous a amenés à nous interroger sur les pratiques de documentation, notamment lors de l'annotation de documents audiovisuels, et plus particulièrement bien sûr à propos de celles des métiers de l'INA.

En outre, les visées documentaires et le contexte applicatif de nos travaux ont justifié l'attention particulière que nous avons portée au cours de notre étude aux questions de l'évaluation et de la paramétrisation des outils.

Ces réflexions, préliminaires à notre étude et présentées dans cette première partie, nous ont permis de positionner nos travaux par rapport aux deux communautés auxquelles nous nous adressons : celle de la recherche en analyse automatique et celle de la documentation audiovisuelle.

# Chapitre 1

# L'indexation, les documents audiovisuels et l'INA

## 1.1 Définition de l'indexation du point de vue de la documentation

L'indexation est une pratique élaborée pour faire face à l'inflation d'informations sous toutes ses formes (textes, sons, images). Cette pratique consiste, en toute généralité, à substituer au document originel une représentation de ce document sous la forme d'une description abrégée, le plus souvent textuelle. Sa finalité est de permettre de repérer rapidement au sein d'un ensemble (ou d'un document), les documents (ou les extraits) pertinents en fonction d'une requête donnée. Cette définition, très générale, est conforme à celle donnée par la norme en vigueur [1] [AFNOR 96].

Pour des documents analogiques, les données descriptives (ou descripteurs) ont longtemps été gérées indépendamment du support physique du document auquel elles sont associées. Avec l'utilisation extensive de documents numérisés, les descripteurs et le document coexistent sur un même support. Il devient alors possible d'établir des liens et d'assurer des correspondances entre un des éléments du documents (objet, segment, événement) et sa description. Ces données, appelées métadonnées [Hunter 99], attachées au document, en facilitent la lecture, la description ou l'accès : "l'indexation est une tâche centrale du traitement documentaire puisqu'elle préside à la constitution d'une description qui sert de substitut au contenu dans les phases de recherche et de consultation du document" [Auffret 00].

## 1.2 Spécificité de l'indexation des documents audiovisuels

### 1.2.1 Principales spécificités du document audiovisuel

Les documents audiovisuels présentent un certain nombre de spécificités [Auffret 00] par rapport aux autres formes de documents, et notamment aux textes :

- la temporalité de sa forme de restitution. D'une part, contrairement au texte ou à l'image fixe, la lecture d'un document audiovisuel nécessite encore un système mécanique de lecture,

1. La norme NF Z 47-102 de l'AFNOR (octobre 1993) qui "a pour objet d'établir des méthodologies valables et cohérentes pour la description et la caractérisation des documents à l'aide de représentations de concepts" indique notamment que "l'indexation est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse."

d'autre part ces systèmes proposent un accès au contenu informationnel par une lecture exhaustive du document.

- l'icônicité. On entend par ce terme que l'image est une icône de ce qu'elle représente. À la différence du langage, le rapport entre l'icône et la réalité à laquelle elle se réfère est fondé sur un rapport de ressemblance. Par contre, les liens d'interprétation entre l'icône, ou plus généralement la forme du document (les images et les sons), et le contenu informationel de celui-ci sont arbitraires. Une des conséquences est que des concepts peuvent être présents dans le document d'un point de vue de l'information tout en restant absents des flux audio et vidéo.
- la polysémie de l'image, liée à l'ambiguïté de toute forme d'illustration, et qui concerne donc aussi l'image fixe[2]. La polysémie de l'image est liée à la fois à la temporalité (nécessité de recontextualiser l'image pour la comprendre) et à l'icônicité[3] (nature des concepts suggérés par la forme du document).
- la nature composite. Un document audiovisuel est un document composite, résultant de la production conjointe de l'image et du son, voire parfois d'un texte incrusté. L'interaction de ces différents médias n'est donc pas sans conséquence sur la compréhension de l'ensemble.

**Conséquences pour l'indexation.** Ces spécificités influent sur les pratiques de l'indexation audiovisuelle. Ainsi, l'indexation des objets se fait en fonction du contexte, de ce que l'ensemble du document suggère. De plus, la reformulation de l'information perçue par les documentalistes est liée aux termes du langage, et se fait donc dans une forme différente de celle du document audiovisuel.

La polysémie peut être envisagée principalement à trois moments de la vie d'un document audiovisuel. Lors de la production, les images sont réalisées avec une finalité donnée, liée à un contexte de création. Ensuite, le traitement documentaire induit à la fois la définition d'un usage et un point de vue permettant une certaine lecture du document. Notons que, pour un usage donné, le contexte interprétatif, et donc la lecture, peuvent être variables. Enfin, une séquence visuelle, extraite de son contexte d'origine, peut être réutilisée pour illustrer d'autres propos. Ainsi, un reportage consacré aux conséquences de la guerre du Golfe a montré un cormoran mazouté ; cette image était en fait extraite d'un reportage sur la marée noire, intervenue sur le littoral breton.

Le commentaire des images, l'influence de la production du document (du tournage au montage), bref la composition du document audiovisuel, permet de lever les ambiguïtés ou de biaiser l'interprétation des images, et indique le point de vue de l'auteur. En résumé, la nature composite d'un document audiovisuel influence notre compréhension d'une image polysémique au travers de contraintes interprétatives croisées.

### 1.2.2 Une indexation nécessairement subjective

Les différents points soulevés, et notamment la nature composite et polysémique de l'image, peuvent rendre le lecteur pessimiste sur les possibilités d'indexation du document audiovisuel : "Est-ce à dire que l'audiovisuel n'est qu'un enregistrement « objectif » de la réalité et que comprendre la

2. Nous reprenons ici une idée couramment répandue. Sans chercher à rentrer dans un débat d'experts qui dépasse le cadre de notre travail, il nous a semblé pour notre part que les sons et les textes peuvent aussi présenter des aspects polysémiques dans leur interprétation.

3. Dans une certaine mesure, il semble contradictoire d'affirmer à la fois que l'image ne véhicule intrinsèquement aucun sens, puis qu'elle est polysémique. En fait, il s'agit d'un abus de langage : l'image entretient un rapport arbitraire aux concepts mais son interprétation est polysémique. Il serait donc plus juste de dire que l'image est poly-interprétable.

construction du sens documentaire par le lecteur revient à comprendre comment le monde fait sens? Est-ce à dire que le document audiovisuel n'a que des sens subjectifs construits par des consciences individuelles et totalement indépendantes les uns des autres?" [Auffret 00].

Comme nous venons de le montrer, lire une image c'est l'interpréter et, par conséquent, l'indexation des documents audiovisuels est intrinsèquement subjective[4]. Ainsi, dans [Bachimont 98], indexer c'est réaliser une "paraphrase d'un contenu en une forme sémiotique interprétable permettant de rendre exploitable le contenu indexé dans le cadre d'une pratique donnée (...). L'indexation est donc une interprétation d'un contenu, une réécriture ou reformulation dans une forme propre à l'exploitation dans un contexte."

Par conséquent, du point de vue du traitement documentaire, l'indexation est un processus à la fois subjectif et éphémère qui reflète un usage donné dans un temps donné. Cette conclusion pose la relativité de toute indexation mais n'affirme pas son impossibilité. D'une part, il ne s'agit pas tant de nier toute possibilité d'indexation que de prendre conscience de sa subjectivité. D'autre part, l'ensemble des subjectivités individuelles est contraint par des pratiques culturelles, des usages plus ou moins figés qui guident l'interprétation. C'est ce que nous nous proposons d'aborder brièvement ci-après.

### 1.2.3 Une indexation guidée par un usage

La recherche ou la consultation d'un document, raisons d'être de la pratique de l'indexation, peut s'établir à des niveaux d'approches différents selon que l'on s'intéresse à une thématique (l'écologie) ou à un objet spécifique (un oiseau mazouté). Les questions auxquelles nous sommes confrontés sont alors : Quel est le niveau de granularité requis pour décrire un document? Quel est le point de vue à adopter? Quelles sont les informations qui fonderont la pertinence de la recherche ou de la consultation du document? Les réponses sont apportées par l'identification d'un usage prescrit par l'organisme gestionnaire des documents, le domaine d'application visé, etc.

Puisqu'il n'est pas possible d'épuiser tous les points de vue sur le document, la perspective d'un usage est le seul guide en matière d'indexation.

**En résumé,** l'indexation est une pratique se référant aux usages qui utilise des techniques de représentation (construction d'objets de substitution) destinées à rechercher et à manipuler des documents au sein d'un ensemble. Dans le cadre de notre étude, nous avons donc essayé de nous donner pour guide l'usage tel qu'il semble prescrit par l'INA.

## 1.3 Pratiques de l'indexation des documents audiovisuels à l'INA

Les pratiques de l'indexation et plus généralement la chaîne documentaire audiovisuelle à l'INA ont déjà fait l'objet d'une présentation détaillée dans des travaux précédents [Auffret 00, Chap. II]. Outre nos propres observations, nous nous contenterons de reprendre quelques-uns des éléments utiles à notre réflexion et au positionnement de notre sujet, et nous renvoyons le lecteur soucieux de plus de détails au document précédemment cité ainsi qu'à [Troncy 01].

Comme nous avons déjà eu l'occasion de le préciser, l'INA gère, au travers de ses deux principales missions que sont la conservation des archives et le dépôt légal, un fonds de 2,5 millions de documents, et 80 000 heures d'archives sont collectées chaque année. Ajoutons à cela que la numérisation

4. par *subjectif* nous n'entendons pas arbitraire ou gratuit, mais variable ou lié au point de vue.

au format MPEG-1 et -2 des documents est en cours[5]. Ainsi, ces deux missions mettent l'INA, en tant qu'institution, et, plus directement ses documentalistes, au cœur des problématiques évoquées plus haut.

Les deux missions de l'INA, confiées à deux départements distincts (le Département Droits et Archives et l'Inathèque de France), correspondent à deux types d'exploitation des documents. "Les fonds conservés au titre de l'archivage professionnel sont communiqués à des fins professionnelles aux producteurs, diffuseurs, éditeurs, établissements éducatifs et culturels. (...) Les fonds conservés au titre du dépôt légal sont communiqués à des fins scientifiques aux étudiants, chercheurs et enseignants"[6]. Ajoutons qu'à moyen terme, l'INA a pour objectifs d'améliorer "le service offert aux clients de l'archivage professionnel en développant la thématisation [et] l'accessibilité aux images et aux sons dans l'environnement Internet, (i) par l'articulation des données documentaires, techniques, juridiques et commerciales sur les programmes afin de disposer rapidement d'une information complète, et de favoriser la communication des documents sélectionnés (...), (ii) par l'amélioration des outils de recherche et d'accès aux images et aux sons dans l'environnement Internet"[7]. Ainsi, outre les usages professionnels et scientifiques pouvant être faits des documents, l'INA tend à s'ouvrir sur le grand public dont les usages seront plus disparates et moins facilement modélisables.

Dans le cadre de la captation du flux entrant, l'INA procède à la documentation, c'est-à-dire au catalogage et à l'indexation des documents audiovisuels. Il s'agit donc pour les documentalistes d'interpréter le contenu des documents pour en dégager les caractéristiques les plus représentatives, en vue des usages ultérieurs précédemment décrits. L'indexation des documents débouche concrètement sur la rédaction de notices documentaires (voir annexe A.3), fondées sur un ensemble de mots-clefs issus d'un thesaurus normalisé sémantiquement. Observons qu'une notice est le résultat subjectif et éphémère d'une pratique, d'usages et d'une lecture à un moment donné. Cette relativité de l'indexation est maîtrisée par un ensemble de règles et d'usages propres aux métiers des documentalistes qui contraint la description des documents et est compilé dans un manuel d'indexation[8].

Les missions d'archivage et de dépôt légal ont été confiées à l'INA à différentes étapes de son histoire. L'INA gère les archives professionnelles depuis 1975, mais n'est chargé du dépôt légal que depuis 1992. Le Département Droits et Archives (DDA) et l'Inathèque de France produisent des descriptions différentes même si les informations contenues dans les notices sont sensiblement similaires. Notons par exemple la présence des *timecodes* de segmentation temporelle dans les notices de l'Inathèque, et la spécification d'un certain nombre d'annotations sur la forme audiovisuelle dans les notices DDA. Ces pratiques liées à des usages différents sont amenées à évoluer d'une part à cause de la numérisation du flux entrant, d'autre part parce que la sérialisation et la normalisation des traitements Inathèque et DDA sont envisagées ou en cours[9]. Ces évolutions sont récentes. Depuis deux ans, les notices ne sont plus rédigées que par l'Inathèque et les correspondants de chaînes, et les logiciels *Mediascope* et *Thesaurus Rex*, dont il sera question plus loin, ont été déployés à l'INA au deuxième semestre 2001.

Dans le cadre de l'Inathèque de France et du dépôt légal, les documentalistes disposent de la

5. 30 091 heures de télévision et 13 720 heures de radio ont été numérisées fin 2000, selon le rapport d'activité de l'INA.

6. *Rapport d'activité 2000*, Institut National de l'Audiovisuel, p. 13

7. Article 2 du *Contrat d'objectifs et de moyens 2000-2003*, signé par l'État et l'INA le jeudi 27 avril 2000.

8. plus précisément, le manuel d'indexation de l'Inathèque, intitulé *Le traitement documentaire des programmes de radio et de télévision à l'Inathèque* datant de 1996 et mis à jour en mars 2001.

9. projets *Captation*, *Intranormes* et *Extranormes*, voir *Rapport d'activité 2000*, Institut National de l'Audiovisuel, p. 13.

possibilité de visionner le document et, le cas échéant, d'une documentation écrite fournie par les diffuseurs[10]. Le flux entrant est segmenté, selon un processus descendant, de la récupération des grilles de programmes sur une semaine à l'obtention d'unités de traitement documentaire[11]. Une fois celles-ci repérées, le flux est segmenté en séquences avec les logiciels *VideoScribe* (flux analogique) ou *MédiaScope* (flux numérique). Notons que *MédiaScope* propose un résumé vidéo par simple extraction d'une image à intervalles de temps réguliers, laissés au choix de l'utilisateur qui peut donc ainsi décider de la granularité de sa visualisation. La segmentation temporelle est ensuite transférée dans l'outil *MédiaIndex* qui permet une annotation fine et synthétique du contenu au niveau des séquences audiovisuelles segmentées. Six principaux types de traitement, correspondant à six niveaux de description des documents, ont été rapportés dans [Troncy 01]. Ils vont d'informations purement signalétiques concernant la production, la diffusion ou les droits, à des descriptions plus analytiques, du type résumé chronologique et annotations sur "qui parle ? à qui ? de quoi ? où ? quand ? dans quel ordre ?". Mentionnons le cas particulier, et pourtant fréquent, des émissions périodiques, qui, respectant un même canevas, sont aussi indexées dans leur ensemble au moyen de *fiches collection* [Carrive 00].

Le DDA n'est plus en charge depuis peu de l'écriture des notices documentaires. La nouvelle organisation des services doit déboucher sur la mise en œuvre de la *thématisation*[12] du fonds d'archives et la constitution de corpus centrés sur des personnalités, des événements ou des thématiques classiquement demandées ou d'actualité. La thématisation consiste à regrouper un ensemble de documents ou d'extraits sur un même thème. Pour cela, les documentalistes disposent de l'outil *Thesaurus Rex* de la société *Question d'Images*. Cet outil offre entre autres fonctionnalités une navigation par extraction d'imagettes, soit à intervalles de temps réguliers au choix de l'utilisateur, soit en fonction d'un découpage automatique en plans[13]. Chaque extrait ou document versé à la base thématique sera représenté par une image représentative choisie manuellement. En général, il s'agit simplement de la première image ou d'une image avec texte incrusté tirée d'un générique.

Enfin, notons que l'INA recueille des documents de types très variés, journaux télévisés, magazines (d'actualités, littéraires, de société, ou sportifs), émissions de variétés, documentaires historiques ou scientifiques, émissions artistiques et musicales. Les documents sont caractérisés selon deux listes d'autorité[14], la typologie médiamétrie utilisée par les publicitaires, et la typologie INA fondée sur les genres, les thèmes et les publics.

**En conclusion,** l'approche de l'indexation à l'INA illustre et confirme les principes évoqués plus haut, et notamment la possibilité de contraindre la subjectivité de la perception des documents audiovisuels par une pratique et des usages. Toutefois, il ne s'agit pas d'un problème clos dans la mesure où l'usage et la pratique ne sont pas figés. Ils sont au contraire amenés à évoluer, notamment en fonction des techniques et des outils de production, de stockage, d'indexation et de mise à disposition, liées aux supports numériques.

10. Dans le cadre de la numérisation, il est prévu de collecter, outre les documents sur support numérisé, un ensemble de données fournies en amont dans la chaîne de production/diffusion (projet Captation).

11. pour plus de détails sur ces sujets se référer à [Auffret 00].

12. projet *InaCom*, voir *Rapport d'activité 2000*, Institut National de l'Audiovisuel, p. 16.

13. les documentalistes peuvent visualiser la courbe de variations d'un histogramme de couleurs correspondant au découpage en plans et à l'extraction des images-clefs.

14. listes de termes normalisées.

# Chapitre 2

# Peut-on automatiser l'indexation des documents audiovisuels?

La question peut paraître quelque peu provocatrice dans le cadre de nos travaux, pourtant elle semble hanter même les chercheurs les plus impliqués dans ce domaine si l'on en juge par ce constat de A. Del Bimbo : "Bien qu'un certain nombre de systèmes prototypaux ont été récemment rendus disponibles, l'utilisation concrète de cette discipline n'a pas encore été reconnue comme ayant eu des applications pratiques" [1].

**L'indexation automatique n'est pas l'automatisation de l'indexation manuelle.** L'indexation manuelle est, pour G. Van Slype [Slype 87], une activité fondamentalement distincte à la fois de la classification et de l'indexation automatique, en raison des unités significatives reconnues (descripteurs vs. concepts) et de la sélectivité mise en œuvre lors de l'analyse (classification synthétique locale vs. analyse critique globale).

Force est de constater que les communautés, issues des domaine de l'analyse d'image, de la reconnaissance des formes, de la vision par ordinateur, du traitement de la parole et du son, des bases de données, de la représentation des connaissances, de l'intelligence artificielle (IA), ou de l'interaction homme-machine, qui ont investi le champ des recherches sur l'indexation automatique de documents audiovisuels se sont heurtées à l'ambition de leurs objectifs. Ainsi que le note R.M. Bolle, "idéalement, la vidéo sera automatiquement annotée d'après le résultat de l'interprétation par la machine du contenu sémantique de la vidéo ; toutefois, étant donné l'état de l'art de la vision par ordinateur, une abstraction des données si sophistiquée ne serait pas réaliste dans la pratique. L'ordinateur devrait offrir plutôt une assistance intelligente dans l'annotation manuelle de la vidéo, ou l'ordinateur pourrait réaliser une annotation automatique avec une interprétation sémantique réduite" [2].

Automatique ou semi-automatique, la recherche sur l'indexation assistée par ordinateur s'inscrit dans le cadre de l'indexation manuelle décrite précédemment et doit donc se confronter aux problématiques évoquées, avec la difficulté supplémentaire d'automatiser un traitement *sémantique*, c'est-à-dire d'être capable d'extraire et de gérer, dans la mesure du possible, des informations qui

1. "Although a number of prototype systems have been made available, recently, nevertheless this discipline has not yet been credited as of concrete use in practical applications." [Bimbo 00]

2. "Ideally, the video will be automatically annotated as a result of machine interpretation of the semantic content of the video; however, given the state of the art in computer vision, such sophisticated data abstractions may not be feasible in pratice. Rather, the computer may offer intelligent assistance in the manual annotation of video, or the computer may perform automatic annotation with limited semantic interpretation." [Bolle 98]

font sens. Le principal écueil, et l'explication généralement admise aux limites des résultats obtenus, est le saut qualitatif à franchir[3] pour proposer des annotations sémantiques à partir des descripteurs numériques extraits du flux. Cette difficulté et la nécessaire définition d'un lien aux usages se retrouvent d'ailleurs dans les problématiques liées à l'évaluation et la validation des différents outils et méthodes proposés.

## 2.1 De l'analyse automatique à l'interprétation sémantique : un saut qualitatif à franchir

En suivant l'idée de G. Van Slype [Slype 87], nous pouvons presque parvenir à opposer les processus impliqués dans les indexations automatique et manuelle. En effet, du côté de l'indexation manuelle, les documentalistes sélectionnent parmi les concepts présents - implicitement ou explicitement - dans le document ceux à propos desquels le document apporte une information susceptible d'intéresser les utilisateurs du système documentaire. Il y a donc nécessité d'une analyse critique, contextuelle et globale du document. Inversement, l'analyse automatique souvent fondée sur des méthodes de traitement du signal, renvoie à une information numérique intrinsèquement dépourvue de sens et bien souvent locale. Ainsi, l'indexation manuelle propose une compréhension globale du document permettant une interprétation contextuelle des éléments locaux, tandis que l'analyse numérique des documents voudrait à partir d'indices locaux reconstruire une interprétation globale.

L'hypothèse sur laquelle se fonde le processus ascendant que se propose de mettre en œuvre l'indexation automatique n'est donc pas triviale. Elle suppose l'existence de liens entre les trois types d'éléments qui constituent un document audiovisuel numérisé : les données sémantiques, les données liées à la forme audiovisuelle, les données physiques (ou numériques). Les données sémantiques sont manipulées sous forme de concepts, objets, segments temporels, etc., par les documentalistes. Les données physiques sont liées au support numérique (les descripteurs physiques seront détaillés à la section 2.3.1). Les données liées à la forme audiovisuelle[4] sont liées à la production et au montage. Un inventaire forcément non exhaustif de ces dernières a été proposé en s'appuyant sur des pratiques professionnelles, par exemple dans [Garrett 84,Davenport 91,Chandler 94].

Des travaux ont étudié parfois avec succès les liens et les "constantes" permettant de remonter de l'information numérique vers des données liées à la forme du document, puis vers des concepts de nature sémantique [Yoshitaka 97]. Il semble même que, dans [Bimbo 00], l'auteur, se référant à la sémiotique, en fasse une des pistes principales pour franchir le saut qualitatif entre les données perceptuelles et le niveau sémantique. Toutefois, il est fort probable que ces travaux ne débouchent que sur des résultats limités dans la mesure où, en raison des spécificités des documents audiovisuels évoquées à la sous-section 1.2.1, il n'y a pas, au sens strict, de grammaire universelle de l'audiovisuel [Metz 68].

Différentes stratégies ont cependant été mises en œuvre dans le but de réduire le fossé entre les données numériques et les annotations sémantiques recherchées ; nous nous proposons d'en évoquer quelqu'unes.

3. "Virtuellement tous les systèmes proposés jusqu'à présent utilisent seulement les données visuelles dont la représentation a une signification perceptive de bas-niveau." ("Virtually all the systems proposed so far use only low-level perceptively meaningful representations of pictorial data, which have limited semantics.") [Bimbo 00].

4. Certains articles les décrivent aussi sous le terme de données syntaxiques. Nous n'utiliserons pas ce terme qui, comme la notion d'*information sémantique*, a fait l'usage d'utilisations ambiguës et contradictoires [Marsicoi 97].

## 2.2 À la recherche du sémantique

Dans [Bimbo 00], sont présentées différentes méthodes ayant trait à cette problématique, ainsi que de nombreuses références dépassant parfois le cadre de notre étude. Nous nous contenterons d'évoquer celles qui nous ont semblé les plus centrales par rapport à la problématique abordée.

La première tentation est sans doute d'appeler "sémantique" toute information qui ne serait plus purement numérique. Ainsi, dans [Bimbo 00], si les descripteurs de couleur ou de texture sont considérés comme "perceptuels", les primitives de contours ou de relations spatiales sont dites "sémantiques". Par ailleurs, une des pistes suggérées[5] est l'utilisation conjointe des informations et notamment celles issues des flux audio et vidéo. D'une manière générale, la combinaison de descripteurs numériques est généralement considérée comme étant une bonne méthode pour produire du "sémantique".

Si, en effet, l'utilisation conjointe de descripteurs de nature différente [Vertan 00,Vasconcelos 98a], et notamment l'utilisation du son [Saraceno 97], permet d'enrichir pertinement l'information extraite automatiquement et d'améliorer les résultats obtenus pour certaines tâches, ces algorithmes n'ont cependant pas fait émerger des outils d'indexation capables de créer et de manipuler des concepts sémantiques.

Toutefois, certains auteurs affirment pouvoir extraire des informations sémantiques. Dans [Vasconcelos 98a], la classification de séquences proposée selon les classes *action, gros-plan, foule, décors* semble en effet soit de nature sémantique, soit liée à la forme du document. Néanmoins, le contexte d'utilisation reste à inventer dans la mesure où il n'apparaît pas clairement dans quelle stratégie d'indexation l'étiquetage selon les quatre classes proposées serait pertinent. La démarche de N. Vasconcelos s'apparente en fait à une démarche d'enrichissement incrémentale de l'information (utilisation conjointe du mouvement et de la couleur, ou des textures et de la couleur) ; démarche tout à fait justifiée et nécessaire afin d'enrichir l'information rendue disponible mais qui ne peut prétendre répondre à la question du sémantique. Dans le même ordre d'idée, N. Dimitrova suggère de recontextualiser temporellement l'information afin de s'affranchir du "syndrome image-séquence"[6] [Dimitrova 95], ce qui a pour conséquence d'enrichir l'information sans toutefois nous faire franchir de saut qualitatif sensible. Il semble donc qu'à ce jour, le niveau d'information atteint par ce type de méthodes ne soit pas suffisant pour permettre une formulation sémantique d'un processus automatique de recherche, de navigation ou de structuration documentaire.

Une autre piste envisagée, où l'on retrouve des influences des méthodes de l'intelligence artificielle, a été de se placer à un niveau symbolique, de définir des objets d'un certain niveau sémantique en s'intéressant à leur manipulation, à leurs rapports spatiaux ou temporels. Si les solutions proposées sont souvent séduisantes [Chang 87,Shibata 92,Naphade 00a], l'extraction des primitives manipulées n'étant pas traitée ou étant gérée manuellement, la question du saut qualitatif à franchir ne se trouve donc pas résolue.

D'autres auteurs semblent placer de nombreux espoirs sur l'interaction homme-machine, où l'utilisateur apporterait lui-même l'information sémantique au sein des algorithmes, soit par la gestion directe de paramètres[7], soit par la définition de profils d'utilisateurs [Merialdo 99]. Une telle solution, nécessitant le développement d'interfaces dédiées, aurait le double avantage de laisser à

5. Notons à ce propos que la fusion de données est en soi un domaine de recherche à part entière [Dimitrova 95]. Sur le sujet de la combinaison des similarités, on pourra se référer aux méthodes citées dans [Jolion 00].

6. Sous ce terme, N. Dimitrova évoque la difficulté de décrire les documents vidéo à partir d'un ensemble limité d'images ("Processing of video as a sequence of selected images is very limiting (I call this the image-sequence video syndrome)").

7. Dans de nombreux travaux, la gestion des paramètres est plus ou moins explicitement laissée aux bons soins de l'utilisateur, c'est par exemple le cas dans [Rui 99b, Sec. 4, p. 365].

l'utilisateur le choix des paramètres dont la gestion est toujours délicate dans les algorithmes de traitement de l'image, et d'apporter aux algorithmes la "compréhension" contextuelle de la tâche par une intervention humaine. La difficulté est alors qu'il est nécessaire d'avoir affaire à des utilisateurs experts [Jolion 98] ayant assimilé le fonctionnement des algorithmes[8]. Le problème a donc été repoussé vers la définition de paramètres idéaux censés être à la fois intuitifs et constituer le pont entre les concepts manipulés par l'utilisateur et les données numériques générées par les algorithmes. Des méthodes ont certes été proposées où l'interaction se fait simplement par l'indication des résultats pertinents ou non-pertinents par l'utilisateur [Nastar 98a,Aksoy 00], toutefois ces méthodes semblent dédiées à des applications de recherche dans une base de données. De plus, quelle que soit la méthode d'interaction retenue, l'interaction serait limitée car fondée sur des implicites perceptifs et non conceptuels [Auffret 00].

Les seules méthodes ayant réellement réussi à franchir ce fameux saut qualitatif semblent finalement être celles qui utilisent l'information *a priori*, dans un cadre particulier, par la mise en œuvre de règles, de modèles, ou d'algorithmes d'apprentissage [Castel 96,Vasconcelos 97c,Carrive 00]. Ces méthodes semblent apporter satisfaction dans les limites du cadre dans lequel elles ont été développées et lorsqu'il s'agit de schémas répétables[9]. Les critiques adressées à ces méthodes sont alors, outre le manque intrinsèque de généralité, le travail de modélisation (sous ses différentes formes) souvent très pointu qui est demandé à l'opérateur humain. Ce travail de modélisation peut, en effet, s'avérer trop coûteux dans la mesure où il doit être fait pour chaque typologie de documents et chaque usage envisagé, et doit être maintenu dans le temps pour tenir compte des évolutions des formes de production et des modes de lecture.

## 2.3 Les acquis de l'indexation automatique

Si les algorithmes d'indexation automatique se heurtent au problème du passage du niveau numérique au niveau sémantique, de nombreux résultats ont cependant vu le jour, dans les limites que nous avons tenté d'expliciter plus haut. En effet, outre le cas de la modélisation de problèmes particuliers que nous venons d'évoquer et qui a su faire ses preuves, il existe de nombreuses tâches où la problématique numérique/sémantique soit ne se pose pas, soit se pose dans des termes envisageables notamment dans le cadre d'une stratégie incrémentale d'utilisation de l'information. Ainsi, le domaine de l'extraction et de la représentation de primitives perceptuelles se situe en amont de la problématique évoquée. Par ailleurs, un certain nombre de relations numérique-forme, ou numérique-forme-sémantique, sont assez générales pour être traitées par les algorithmes d'indexation automatique. À titre d'exemple, nous pouvons citer le cas de la segmentation en plans, où les données liées à la forme du document (le découpage de la vidéo en plans) sont fortement corrélées aux données physiques (la discontinuité du signal audiovisuel). Le cas de la détection de visage est aussi révélateur de ce qu'il est possible de faire. Les caractéristiques physiques (couleur, texture) de l'objet sémantique "visage" sont suffisamment sélectives pour permettre la mise en œuvre d'algorithmes de détection. La localisation et la connaissance du rapport de surface entre le visage et l'image peuvent ensuite permettre d'identifier les gros plans (information liée à la forme du document).

8. Dans des domaines comme l'imagerie médicale ou satellitaire, l'existence de tels utilisateurs a pu rendre adéquat ce type d'approche. Dans le domaine audiovisuel, et en particulier à l'INA, cela semble moins évident. Il paraît peu probable que les principaux utilisateurs finaux (professionnels, scientifiques et grand public évoqués à la section 1.3) puissent devenir des utilisateurs experts à moyen terme. L'appropriation de prototypes algorithmiques par les documentalistes nécessiterait aussi un effort non négligeable d'adaptation et de formation.

9. ce qui est le cas à l'INA, au moins partiellement, dans le cadre de la constitution des *fiches collection*.

Nous nous proposons donc de dégager succinctement les principaux axes d'intérêt qui constituent la recherche en indexation automatique [Aigrain 96,Idris 97,Brunelli 99,Rui 99a].

### 2.3.1 Extraction de descripteurs

L'extraction de descripteurs, souvent de nature numérique ou perceptuelle, a été l'objet d'une attention certaine dans la mesure où de nombreux espoirs se fondent sur les possibilités d'enrichir et d'affiner l'information extraite du flux vidéo. Contrairement à une annotation textuelle du document, ces descripteurs physiques ne peuvent rendre compte d'une information sémantique, et en particulier pas d'un concept absent physiquement du flux ; toutefois ils présentent l'intérêt d'être de même nature que le flux qu'ils décrivent. C'est pourquoi ils ont été largement utilisés dans le cadre de tâches pouvant se fonder sur une similarité entre objets (notamment dans le cas de requête par l'exemple), ou en vue de la mise en œuvre d'outils de visualisation (outils dont il sera question à la sous-section 2.3.5). Nous allons évoquer quelques-uns de ces descripteurs, qu'ils soient globaux ou locaux.

#### *2.3.1.1 Primitives de couleurs*

Ces descripteurs assez intuitifs ont été abondamment étudiés [Swain 91,Stricker 95]. L'idée principale est de considérer les couleurs des points d'une image comme une distribution statistique. Ainsi, une signature possible est l'extraction des moment statistiques (i.e. moyenne, variance, etc.) [Stricker 95]. Une autre signature courante est l'histogramme de couleurs [Swain 91] qui est une approximation de la probabilité jointe des intensités des trois canaux de couleur. Comme les histogrammes sont en général non-denses, et par conséquent sensibles au bruit, les histogrammes cumulés ont été suggérés [Stricker 95]. Une signature alternative proposée dans [Smith 96] est l'extraction d'une liste ou d'un ensemble de couleurs dominantes.

Si ces primitives globales sont décrites comme étant relativement robustes aux petits changements du contenu ou de l'orientation des images, un défaut souvent énoncé est l'absence d'utilisation des relations spatiales, ce qui serait à l'origine de nombreuses fausses détections dans les systèmes d'indexation automatique. Par conséquent, des auteurs ont proposé différentes améliorations afin de distinguer les images non identiques mais similaires en terme de couleurs globales. Des descripteurs localisés ou régionalisés ont été proposés pour les distributions de couleurs ainsi que pour les histogrammes [Stricker 96]. D'autres améliorations ont été proposées pour prendre en compte davantage d'informations comme la densité des contours, la texture, l'amplitude du gradient d'intensité, ou la cohérence locale (CCV) [Pass 99]. Une autre possibilité est de pondérer la contribution des points à l'histogramme à travers différentes grandeurs comme le Laplacien, des mesures probabilistes ou floues [Vertan 00]. Une autre approche pour introduire l'information spatiale est d'utiliser des corrélogrammes de couleurs ou des auto-corrélogrammes [Huang 97].

#### *2.3.1.2 Primitives de textures*

Ces descripteurs sont extraits principalement par des méthodes fréquentielles ou statistiques. Une des méthodes classiques est la mise en œuvre de bancs de filtres de Gabor spatiaux ; les coefficients obtenus sont ensuite résumés par des grandeurs statistiques (moyenne, variance) [Manjunath 96]. Une autre méthode usuelle est de calculer des matrices de cooccurrences pour différentes distances et orientations. L'information est alors décrite par des descripteurs globaux, notamment ceux d'Haralick [Haralick 73,Gotlieb 90]. Par ailleurs, [Vehel 00] signale l'utilisation possible de paramètres fractaux (les exposants d'Hölder), et [Tamura 78] s'est intéressé à des grandeurs génériques

censées être plus intuitives[10], liées à des notions de contraste, de directionnalité, de régularité, etc.

#### *2.3.1.3 Autres primitives pour image fixe*

D'autres descripteurs sont utilisés comme les contours ou les formes des objets. L'algorithme de Canny-Deriche pour l'extraction des contours semble être toujours la méthode de référence [Canny 86,Deriche 87]. Par ailleurs, on pourra se référer à l'article de [Scassellati 94] qui étudie certaines primitives de formes en relation avec la perception humaine. Les principaux descripteurs de formes cités par [Idris 97,Rui 99a] sont les descripteurs de Fourier et les moments invariants, notamment ceux proposés par [Hu 62]. D'autres méthodes proposent de rechercher des régions ou des points caractéristiques (ou points d'intérêt) dans l'image [Schmid 97].

#### *2.3.1.4 Primitives de mouvement*

La prise en compte d'indices dynamiques a aussi été un domaine de recherche largement abordé [Cedras 95,Mitiche 96]. Comme nous le verrons plus en détail dans le chapitre 7, il est difficile de donner une vue d'ensemble synthétique des différents travaux. D'une part, les niveaux d'étude du mouvement ou de l'activité sont très variés, de l'extraction d'information de mouvement de bas-niveau à des classifications d'activités [Bobick 97]. D'autre part, l'étude du mouvement comprend un grand nombre de travaux qui dépassent le cadre de nos recherches en indexation automatique et qui ont trait à la vision par ordinateur, la robotique ou la vidéo-surveillance. Nous nous contenterons d'évoquer ici quelques descripteurs de bas-niveau. Ceux-ci sont déjà fort nombreux dans la mesure où la plupart des primitives visuelles déjà citées peuvent être proposées sous une forme temporelle. Citons par exemple les modèles de textures temporelles suggérés par [Nelson 92], ou les *Trajectory Primal Sketch* (TPS) de [Gould 89] qui sont liées aux formes et aux contours.

Par ailleurs, de nombreuses études ont été menées sur l'extraction du flot optique [Nagel 87, Konrad 92,Barron 94,McCane 98], différentes méthodes (méthodes différentielles, méthodes fréquentielles, méthodes fondées sur la phase ou la correspondance entre régions) sont présentées dans [Cedras 95]. Autre sujet de recherche classique, la mesure du mouvement dominant, par exemple en utilisant une modélisation paramétrique et un estimateur robuste [Odobez 95], des modélisations paramétriques du mouvement sont aussi présentées dans [Black 97].

#### *2.3.1.5 Primitives audio*

L'analyse de la bande audio utilise bien souvent des méthodes de filtrage fréquentiel. Ainsi, les coefficients cepstraux semblent être les descripteurs de base du flux audio, même s'ils ont été au départ définis pour l'analyse de la parole [Rabiner 93]. Certains auteurs ont cherché à définir des grandeurs globales plus intuitives [Scheirer 97,Pfeiffer 99].

### 2.3.2 Segmentation de la vidéo

La segmentation temporelle d'un document vidéo en plans peut s'appuyer sur les discontinuités du signal. Par conséquent, de nombreuses méthodes ont été proposées [Dailianas 95, Boreczky 98], fondées sur la l'évolution temporelle des descripteurs présentés ci-dessus: mouvement [Bouthemy 99b], couleur [Yeo 95], ou contour [Zabih 96]. Des formulations alternatives ont

10. Plus précisément [Tamura 78] propose des descripteurs censés correspondre à la perception visuelle humaine, ceux-ci sont au nombre de six, soit, en anglais: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity*, *roughness*.

été suggérées comme celle statistique de [Vasconcelos 97a], ou celle liée à des modèles de production [Hampapur 95]. Dans [Ngo 01], l'auteur considère des sections spatio-temporelles (*temporal slices*) ; l'étude de descripteurs qui en sont issus permet de retrouver les ruptures de plans et certaines transitions graduelles. Les transitions entre plans peuvent être brutales (cas des *cuts*) ou graduelles (cas des transitions progressives). Le premier cas est évidemment le plus simple, et les résultats de la détection des ruptures franches atteignent les 90% [Dailianas 95]. Toutefois, l'utilisation de plus en plus fréquente de techniques d'insertion, de partage de l'image, ou d'effets de transition complexes, amène à brouiller la perception de la continuité et des discontinuités du signal et prend à défaut une grande partie des algorithmes classiques. Ainsi, l'étude des transitions graduelles et des effets d'édition qui ne cessent d'évoluer demeurent une direction de recherche ouverte [Aigrain 94,Hampapur 95,Zabih 99], et ceux-ci ne sont pas toujours bien pris en compte par les algorithmes classiques.

D'autres niveaux et d'autres approches de la segmentation ont été envisagés, nous verrons cela en détail dans le chapitre 4.

La segmentation temporelle consiste aussi en la segmentation du flux audio. Celle-ci est souvent réalisée à partir de la caractérisation des segments selon les classes *Silence*, *Parole*, *Musique*, *Autres* [Spina 96].

### 2.3.3 Caractérisation d'objets

La détection et l'extraction d'objets [Ohba 97], l'étude de leur apparence [Nagasaka 91] et leur suivi [Marques 97,Courtney 97,Gelgon 99] est aussi un axe de recherche qui a été particulièrement étudié, si bien que la durée de vie des objets a été proposée comme segment fondamental en remplacement du plan [Gunsel 98b]. Dans le cadre de l'approche par stratification, l'objet est d'ailleurs le fondement du processus d'indexation du flux audiovisuel [Prie 98]. Enfin, dans un contexte symbolique, des travaux ont été menés afin d'étudier les relations spatiales entre objets à des fins d'indexation ou de recherche [Chang 87,Bimbo 95].

### 2.3.4 Extraction d'entités d'intérêt

Des objets plus spécifiques ont aussi été recherchés dans le flux et notamment les logos [Soffer 98], les visages [Yang 01] et les textes incrustés [Kim 96]. Dans le cadre d'applications dédiées [Haering 99], et donc dans un cadre se prêtant à la modélisation de liens entre les éléments physiques, de forme et sémantiques, des travaux ont abouti à la détection d'événements de nature sémantique (flash, explosion) [Naphade 98] ou liés à la forme du document (gros plan, ralentis) [Vasconcelos 98a,Kobla 99].

### 2.3.5 Représentation, accès, navigation et recherche

Des travaux se sont naturellement intéressés aux sujets connexes à l'indexation des documents, tels que la représentation des données extraites du flux, l'accès aux documents et aux données qui lui sont attachés (la navigation), et bien évidemment la recherche de document ou d'extraits.

La représentation, la navigation et l'accès aux données commencent avec l'extraction d'images-clefs ou l'utilisation de mosaïques [Anandan 95,Taniguchi 97,Gelgon 98] afin de visualiser les informations liées aux plans ou aux mouvements de caméra, et consistent aussi en la construction de résumés vidéo [Lienhart 97,Smith 97,Saarela 99,Lacoste 02], de tables des matières ou d'index [Llach 99]. Ces tâches sont bien souvent associées à une recherche de la structuration du document

audiovisuel [Aoki 96,Kobla 97] ou fondées sur des notions de similarité [Zhang 95,Zhong 96] ou de caractérisation des segments [Rigoll 96,Fischer 95].

De très nombreux travaux ont en particulier abordé le problème de la recherche de documents (ou d'extraits) dans une base (ou dans un document). Ces problématiques rejoignent des questions liées à la gestion et à la structuration des bases de données et à la construction d'interfaces de requête. Ces points ne seront pas abordés dans le cadre de cette étude, notons cependant que la majorité des travaux de recherche d'information par le contenu[11] traite en fait de recherches par l'exemple[12] fondées sur des notions de similarité physique [Vinod 98,Jain 99]. La recherche sur ce sujet a connu un succès important dans la mesure où son champ d'application dépasse largement le cadre de l'indexation des documents audiovisuels (imagerie satellitaire, imagerie médicale, applications orientées vers l'internet[13]) [Gudivada 95]. Si la recherche par similarité semble unanimement utilisée pour évaluer la pertinence des différents descripteurs, la requête par l'exemple reste toutefois limitée aux informations physiquement présentes dans le flux (ainsi qu'il a été dit à la section 2.2) et par la capacité de ces descripteurs à rendre compte de concepts sémantiques.

### 2.3.6 Développement de prototypes et travaux de normalisation

L'ensemble de ces recherches a débouché sur la mise en œuvre d'un certain nombre de systèmes ou de prototypes, notamment *QBIC* [Niblack 93], *Photobook* [Pentland 94], *Virage* [Hampapur 97], *VideoQ* [Chang 97], *VisualGREP* (du projet MoCA) [Lienhart 98] ou *Surfimage* [Nastar 98b] qui ont été décrits par exemple dans [Rui 99a,Fablet 01]. Ces prototypes intègrent nombre des différents algorithmes décrits plus haut, et proposent en général des outils de recherche par l'exemple, assistés d'outils de recherche par mots-clefs. L'exception notable est le prototype VideoQ où l'interface permet à l'utilisateur de dessiner sa requête, toutefois "on peut mettre en doute les capacités de dessinateur de l'utilisateur moyen (...) [et] l'exemple reste un vécu perceptif unique et individuel qui n'accède jamais à l'universalité du concept" [Auffret 00].

Enfin, même si l'INA n'a pas inscrit notre travail de recherche dans le cadre de cette normalisation, signalons l'ensemble des travaux du groupe MPEG (*Moving Pictures Expert Group*), comité de standardisation ISO lié à l'audiovisuel dont les premiers travaux ont consisté en la mise en place d'une norme de compression du signal audiovisuel (MPEG-1 & 2), pour ensuite évoluer vers une tentative de normalisation de la composition multimédia (MPEG-4) [Auffret 00]. Les dernières évolutions liées à la norme MPEG-7 sont fortement liées aux différents sujets abordés dans ce bref état de l'art puisqu'une normalisation de la représentation des contenus a été proposée fin 2001[14] [Salembier 01].

## 2.4 La question de l'évaluation

Comme dans les domaines de la transcription audio [Gauvain 90] et de la catégorisation de texte [Dumais 98b], la validation des performances des différents outils automatiques d'indexation audiovisuelle sur une base d'expérimentations de référence s'est révélée nécessaire. La mise en œuvre

11. Soit, en anglais, *Content-Based Information Retrieval* (CBIR).

12. Soit, en anglais, *Query By Example* (QBE).

13. On pourra juger de l'engouement médiatique pour ce type d'applications liées au développement de la nouvelle économie en se reportant aux articles exagérément enthousiastes publiés par exemple dans *Libération* : "Les Nouveaux filets de la Toile" par David Groison et "Lookthatup, les images supplantent les mots clés" par Nidam Abdi (*Libération*, jeudi 10 mai 2001).

14. Voir l'adresse `http://mpeg.telecomitalialab.com/standards.htm`, où des informations sont disponibles à ce sujet.

d'évaluations de ce type dépend à la fois de la construction d'une indexation de référence commune à tous, et de la définition d'indicateurs de fiabilité (taux d'erreur, oubli, précision, fausses détections, etc.).

Pour un certain nombre d'algorithmes, des ensembles de validation ont été proposés, notamment pour des bases d'images plus ou moins dédiées à des domaines précis, mais aussi pour la segmentation en plans [Ruiloba 99]. De même, la définition des indicateurs de fiabilité a été largement abordée. Toutefois, la constitution de corpus vidéo étiquetés en fonction d'un usage et la définition des protocoles d'évaluation associés restent à effectuer pour de nombreuses tâches d'analyse automatique de documents audiovisuels. En particulier, l'évaluation d'outils d'indexation cherchant à proposer des résultats sémantiques reste très problématique.

En effet, le fait que les résultats d'une indexation *sémantique* soient parfois difficilement mesurables avec des indicateurs purement numériques et quantitatifs n'est qu'une des difficultés rencontrées lors de l'évaluation. Lorsque qu'aucun usage précis n'a été défini préalablement, l'évaluation se fait souvent soit sur des corpus minimalistes dédiés aux algorithmes mis en œuvre, soit sur des corpus construits autour de notions sémantiques imprécises. Il arrive fréquemment que les expérimentations menées soient peu détaillées quant au corpus utilisé, et à la méthodologie appliquée.

Lorsqu'une réflexion a été menée, elle semble souvent mettre l'accent sur le fossé entre le niveau sémantique de la formulation des requêtes et celui des résultats d'algorithmes, et paraît susciter plus de question qu'elle n'en résout. Ainsi, dans le contexte des *vidéos à la demande*, où une étude des requêtes a été menée [Rowe 94], l'auteur donne comme exemple : "Les utilisateurs veulent retrouver des vidéos à partir de leur contenu. Les exemples d'index de contenu sont : 1) les ensembles d'images-clefs significatives de la vidéo (comme les images représentant chaque acteur ou les segments et scènes principaux) (...), 3) des index d'objets indiquant les images d'entrée et de sortie pour chaque objet ou individu significatif" [15]. À la lecture de telles spécifications, les questions restent nombreuses : qu'entend-on par *contenu* ? qu'est-ce qu'une *image représentative pertinente* ? comment définit-on un *segment principal* ? qu'est-ce qu'un *objet significatif* ?

D'éventuelles réponses à ces questions ne pourront être suggérées que dans le cadre défini précédemment au chapitre 1, c'est-à-dire en proposant une interprétation subjective des documents, guidée par des usages pré-établis.

15. "Users want to retrieve videos based on their content. Examples of content indexes are: 1) sets of keyframes that represent key images in a video (e.g., frames that show each actor in the video or a sequence of images that depict the major segments and scenes in the video) (...), 3) object indexes that indicate entry and exit frames for each appearance of a significant object or individual".

# Chapitre 3

# Définition d'une stratégie d'indexation automatique dans le cadre de l'INA

Après avoir essayé de situer les problématiques liées à l'indexation de documents, et plus précisément de documents audiovisuels dans le cadre des pratiques de l'INA, nous nous sommes intéressés aux acquis de l'indexation automatique en nous focalisant sur les principaux axes de recherche mis en œuvre, ainsi qu'aux limites de l'automatisation de l'indexation de ces documents. Nous nous proposons, à présent et pour conclure cette première partie, d'évoquer les quelques travaux préparatoires à notre étude qui nous ont permis de mieux appréhender les problématiques développées plus haut, puis de définir plus précisément les objectifs qu'il nous a paru pertinent de nous donner dans le cadre de cette étude, et ce en liaison avec les besoins de l'INA.

## 3.1 Travaux préparatoires

Les travaux préparatoires que nous avons menés avaient pour objectif principal de nous permettre d'appréhender pratiquement les différentes difficultés abordées plus haut. D'une part, nous avons pu nous entretenir avec différents documentalistes de l'Inathèque et de la DDA sur leur pratique de l'indexation des documents audiovisuels au sein de l'INA, et nous avons suivi quelques séances du stage *Journalisme audiovisuel : formation de rédacteur reporteur d'images*[1] (JRI). D'autre part, nous avons mené quelques expérimentations préliminaires sur les liens entre données physiques, données liées à la forme du document et données sémantiques en nous limitant à diverses signatures de couleur et leurs distances associées dans le cadre d'une recherche par l'exemple.

### 3.1.1 Entretiens avec des documentalistes l'INA

Dans le cadre des entretiens avec les documentalistes, nous avons pu enrichir notre information sur les points développées dans le chapitre 1. En particulier, il nous est apparu plus clairement que si l'usage donnait un sens à une interprétation ou une lecture des documents et ainsi à une indexation donnée, l'usage lui-même n'est pas fixé une fois pour toutes. Comme nous l'avons vu, il existe deux types d'usage à l'INA, celui de l'Inathèque axé vers la recherche scientifique, et celui des archives axé vers un public de professionnnels, auxquels on peut ajouter un usage émergent qui

1. Pour en savoir plus : `http://www.ina.fr/Formation/index.fr.html`.

serait celui d'une ouverture vers le grand public via l'internet. À ces évolutions institutionnelles (ouverture du dépot légal en 1992, élargissement de celui-ci aux chaînes du câble et du satellite[2] d'ici 2003 et évolution des missions de service public vers le grand public), s'ajoutent des évolutions pratiques liées aux outils techniques, et à de nouveaux types de requêtes des utilisateurs qui seront ensuite pris en compte en amont lors de l'indexation. Ainsi, l'usage même lorsqu'il est défini n'est qu'un guide relatif[3].

Nous avons aussi constaté une forte interaction entre une pratique d'indexation, les outils techniques mis à la disposition des documentalistes et les requêtes envisagées par l'utilisateur. Ainsi, si nous avons évoqué jusqu'à présent l'idée selon laquelle les pratiques d'indexation et les requêtes forment un usage qui définit des besoins de développement d'outils d'indexation automatiques, il est non moins vrai que le développement de nouveaux outils fera naître de nouvelles requêtes qui induiront des évolutions des pratiques d'indexation.

La prise en compte d'un usage doit donc tenir compte d'une certaine variabilité intrinsèque de celui-ci ainsi que d'une dimension prospective. Cette vision prospective n'est pas simple à définir dans la mesure où les professionnels rencontrés ne semblent pas avoir une intuition immédiate et évidente des évolutions techniques possibles à apporter à leurs outils et à leurs pratiques. La définition d'un tel schéma prospectif est un travail de grande ampleur dépassant largement le cadre de cette étude.

Par ailleurs, nous avons constaté que tous les résultats des outils d'extraction automatique ne sont pas systématiquement souhaitables ou pertinents du point de vue des documentalistes, ou de l'organisme prescripteur. Ainsi, dans le cas où les algorithmes repèrent un certain nombre d'entités sans caractérisation suffisamment fine et sans structuration suffisamment formalisée, le résultat est perçu négativement. À titre d'exemple, les outils de segmentation en plans fonctionnent relativement bien d'un point de vue algorithmique, cependant ils aboutissent à la mise à disposition d'un très grand nombre d'entités qui exigeraient de l'organisme prescipteur un surcoût de travail et des documentalistes une indexation fastidieuse et répétitive, là où ils n'indexaient que quelques séquences par document. Par conséquent, la segmentation en plans, *en tant que résultat documentaire* ne semble pas présenter d'intérêt pour l'INA. Le cas de figure serait similaire avec les visages si leur détection n'était pas suivie d'une classification et d'une structuration supplémentaire de l'information.

### 3.1.2 Observation de la production de documents audiovisuels

Le suivi du stage de formation JRI devait nous permettre d'essayer de mettre à jour les *constantes* de prise de vue et de montage, constantes sur lesquelles pourrait s'appuyer une modélisation de liens entre des données physiques, des données liées à la forme du document et des données sémantiques (voir section 2.1). Nous avons assisté à quelques séances de dérushage et de montage de reportages correspondant à des journaux télévisés. Notre observation a semblé confirmer en effet l'existence de constantes. À titre d'exemple, les interviews paraissent être souvent précédées d'un court plan de mise en situation de la personne dans son cadre d'intervention. De même, les reporters semblent être incités à filmer un certain nombre de plans des objets auxquels ils s'intéressent

2. Objectif 1.5 du *Contrat d'objectifs et de moyens 2000-2003*, signé par l'État et l'INA le jeudi 27 avril 2000.

3. Ainsi, il nous a été rapporté l'exemple d'un documentaire tourné après-guerre en noir et blanc sur le peintre Modigliani. Ce documentaire réalisé par Jean-Marie Drot dans le cadre de la série *Les Arts et les hommes* présentait, outre des œuvres du peintre, de nombreuses scènes tournées dans le Montparnasse de l'époque. L'indexation originelle mettait l'accent sur les œuvres présentées et non sur le contexte parisien contemporain considéré comme un décor habituel. Aujourd'hui, tout l'intérêt de ce document réside au contraire dans ces images du *Vieux Paris*, tandis qu'il existe d'autres images (en couleur) des peintures de Modigliani.

et notamment des plans larges permettant une contextualisation des images montrées par la suite. Autre exemple concernant le montage, une certaine importance est attachée à la cohérence du mouvement dans la continuité des différents plans. Ainsi, il est déconseillé aux caméramen de faire des champs-contre-champs à 180° lorsqu'ils suivent un événement en mouvement (une manifestation par exemple).

Malheureusement, au-delà de constantes de production éparses et disparates, nous n'avons pu fixer de réelles contraintes. Celles-ci nous ont paru largement insuffisantes pour mettre en œuvre des démarches du type de celles préconisées dans [Bimbo 00]. De plus, il nous a semblé que ces constantes présentes dans une pratique professionnelle n'étaient absolument pas théorisées. Le discours premier de nos interlocuteurs était d'infirmer l'existence de ces constantes avant de reconnaître, confrontés au point de vue de l'observateur extérieur, qu'il y avait évidemment un regard à exercer de préférence à tout autre, des images à rapporter obligatoirement, ou qu'il n'était pas conseillé de filmer une conférence de presse n'importe comment. Les constantes sont donc intériorisées par ceux qui les pratiquent plutôt que théorisées. Sans chercher à tirer des conclusions trop définitives, il nous fut donc impossible, dans le cadre de notre étude, de nous appuyer sur des règles issues d'une éventuelle théorie du montage ou du tournage. De nouveau, il faudrait procéder sur ces pratiques professionnelles à une étude de plus large envergure, ce qui ne rentrait pas dans le champ de nos travaux.

Enfin, nous nous sommes intéressés aux contraintes techniques (par exemple, le positionnement des caméras sur un terrain de sport) qui pouvaient elles aussi induire des constantes de production des documents. Une fois de plus, nous avons pu observer certaines de ces constantes. Par exemple, notre intuition que les images extraites d'un match de football correspondent à un certain nombre de plans-types est confirmée par des études sur le placement des caméras lors de la coupe du monde de 1998 [Gourdon 99]. Toutefois, une étude un peu plus précise nous montre qu'il reste encore trop de variabilité dans ces constantes. Ainsi, le traitement des épreuves d'athlétisme dans deux des documents que nous avons dans notre corpus (document *talence* et *munich*, voir annexe A.1 pour plus de détails) laisse, au-delà d'une certaine régularité des plans (comme les travelling latéraux lors des courses de vitesse), trop de place aux variations (par exemple, le placement des caméras lors des épreuves de saut à la perche). C'est pourquoi nous n'avons pu mettre en œuvre les pistes suggérées par les tenants de la "grammaire de l'audiovisuel", en partie reprises dans [Bimbo 00].

### 3.1.3 Une étude de cas

Enfin, nous avons étudié un cas pratique de correspondance entre éléments physiques, de forme du document, et sémantiques. Nous avons développé une librairie permettant d'extraire les descripteurs de couleur les plus utilisés (voir sous-section 2.3.1), ainsi que les mesures de similarité associées. Nous nous sommes aussi dotés d'une base de 310 images fixes issues de nos corpus[4] pour lesquelles nous avons tenté de définir différents niveaux de similarité liés à des notions sémantiques et/ou de forme du document[5].

Nous nous sommes d'abord heurtés au point suivant : non seulement une lecture sémantique des images nous a menés hors de portée des capacités des descripteurs physiques, tant une même notion se trouvait mise en image avec beaucoup trop de variabilité, mais, de plus, aucune grille de lecture liée à la forme du document ne nous a paru faire pertinemment le lien entre le physique et le sémantique. Même lorsque nous nous sommes posés la question des éléments de forme intervenant dans la similarité physique des images, nous avons été réduits à lister vainement une série

4. notamment les documents *aim1mb05*, *aim1mb08*, *albertville*, *munich1*, *munich2*, *talence1*, cf. annexe A.1.

5. pour plus de détails, le lecteur pourra se référer à [Veneau 01].

non exhaustive d'éléments : l'angle de vue, l'éclairage, la valeur de plan[6], etc. Nous avons exploré différents éléments parmi ceux donnés dans les manuels techniques et de production [Duhen ] sans parvenir à proposer des niveaux de similarité s'appuyant sur d'éventuels liens physique-forme-sens qui nous auraient permis de franchir ce fameux saut qualitatif.

Nous nous sommes rabattus sur des définitions relativement généralistes de nos niveaux de similarité, mais nous n'avons pu ensuite associer clairement ceux-ci aux différents niveaux d'information extraits par les couples descripteur/similarité mis en œuvre. Il nous a simplement semblé que tous nos descripteurs rendaient compte d'une même information physique, de manière plus ou moins subtile ou riche, nous permettant d'obtenir sur un corpus donné une simple hiérarchisation des performances.

## 3.2 En guise de conclusion à nos réflexions préliminaires

L'ensemble de ces études théoriques et de ces expérimentations pratiques avait pour objectif de mieux nous permettre d'appréhender notre contexte de travail. Afin de fixer un cadre à nos recherches, nous avons été amenés à mettre en exergue quelques idées principales, au final, très nuancées :

- il n'y a pas d'indexation objective. L'analyse d'un document doit être guidée par un usage. Toutefois, il est nécessaire de garder à l'esprit la relativité de cet usage en fonction des prescripteurs, dans le temps et en fonction des évolutions techniques ;

- les rapports entre les éléments d'expression et le sens contenu dans un document audiovisuel sont d'autant plus complexes qu'il n'y a pas à proprement parler de langage audiovisuel. L'étude des liens entre données physiques, de forme, et sémantiques, que nous avons menée dans un cadre général, ne nous a pas permis de mettre à jour des constantes suffisantes pour être modélisées. Par ailleurs, la pratique documentaire telle qu'elle est appliquée à l'INA exige une formalisation suffisante des informations extraites automatiquement ;

- la question du saut à franchir qualitativement entre les descripteurs issus du flux vidéo et le niveau interprétatif est donc particulièrement cruciale. Seules les méthodes utilisant de l'information *a priori* semblent capables de franchir ce pas. Cela implique entre autres de réduire le champ de nos investigations à un problème précis où une modélisation s'avère possible.

## 3.3 Problèmes abordés et articulation avec les besoins de l'INA

Dans [Auffret 00], il est indiqué trois opérations principales pour l'indexation : le repérage des segments documentaires, la caractérisation du contenu de ces segments, la structuration de cette information.

Constatant que les résultats de la segmentation en plans par les algorithmes déjà développés semblent satisfaisants mais que les besoins de l'INA en terme de segmentation se situent davantage au niveau de séquences plus longues, nous avons décidé de nous intéresser à la segmentation des documents en séquences (autrement dénommée macro-segmentation de la vidéo).

6. La *valeur de plan* est une indication liée au cadrage, les valeurs de plan les plus courantes sont le gros plan, le plan américain, le plan d'ensemble, etc.

De plus, il nous a semblé pertinent de prolonger ce travail par la mise en œuvre d'une caractérisation des contenus segmentés. Nous nous sommes ainsi intéressés à des algorithmes de caractérisation fondés sur le mouvement. Dans ce cadre, nous nous sommes focalisés sur des séquences courtes [7]. Nous nous sommes de plus intéressés aux méthodes par apprentissage qui présentent le double intérêt d'utiliser de l'information *a priori* et de ne pas nécessiter de modélisation poussée de cette information. Nous avons dû toutefois identifier un champ d'investigation délimité. Notre choix s'est porté sur les documents sportifs, en raison de l'information de mouvement censément riche, et parce que le corpus du projet AGIR s'y prêtait.

Par ailleurs, la numérisation des documents en cours à l'INA se faisant au format MPEG, nous avions la contrainte technique de proposer des algorithmes pouvant extraire l'information d'images reconstruites par décompression du flux MPEG.

Nous nous sommes intéressés d'abord à la segmentation puis à la caractérisation du flux, respectant ainsi la pratique actuelle, ce qui nous a paru plus naturel. Toutefois, les tâches de segmentation et de caractérisation sont bien évidemment fortement imbriquées. Bien souvent, segmenter consiste plus ou moins implicitement à caractériser. Nous allons aborder successivement dans les deux prochains chapitres les problèmes de macro-segmentation et de caractérisation du contenu par le mouvement.

7. Il s'agit en fait de "blocs temporels" d'une durée inférieure à la seconde.

# Deuxième partie

# Macro-segmentation d'un document audiovisuel

# Introduction

L'objectif d'une segmentation temporelle en séquences, ou *macro-segmentation*, est de permettre de retrouver automatiquement des unités documentaires semblables à celles obtenues manuellement par les documentalistes de l'INA (voir section 1.3). L'automatisation de cette tâche s'inscrit dans les problématiques que nous avons évoquées à la section 2.1. En effet, nous souhaitons retrouver une segmentation temporelle fondée sur une compréhension globale du document à partir d'informations extraites du flux : les unités temporelles recherchées doivent exhiber une cohérence "sémantique".

Dans le chapitre 4, nous tenterons de définir le contexte d'utilisation des outils de macro-segmentation, et nous nous interesserons, dans un état de l'art spécifique sur cette question, aux méthodes proposées pour automatiser le découpage en séquences. Nous présenterons la solution que nous avons définie, et les choix algorithmiques qui nous ont guidés au chapitre 5.

Le problème de l'évaluation des performances se posant de manière aiguë pour les algorithmes de macro-segmentation, nous détaillerons, dans le chapitre 6, la démarche méthodologique que nous nous sommes donnée ainsi que ses limites, avant de présenter et de commenter les résultats obtenus.

# Chapitre 4

# Contexte de l'étude

## 4.1 De la segmentation temporelle d'un document audiovisuel

Obtenir une segmentation temporelle des documents audiovisuels semble être une tâche incontournable dans le cadre d'un procédé d'indexation audiovisuelle, notamment lorsque sont traités des documents numériques. En effet, l'indexation d'un document audiovisuel consiste à repérer des entités (pouvant être des événements, personnes, concepts, etc.) qui ont une existence sur un intervalle temporel qu'il s'agira de délimiter. Comme évoqué précédemment, le traitement numérique d'un document audiovisuel permet en outre d'attacher à une ligne temporelle des descriptions du document et d'accéder de manière non linéaire à celui-ci. Ceci suppose l'existence d'"objets" temporels et donc d'une segmentation temporelle du document.

### 4.1.1 Stratification ou structuration hiérarchique?

Deux approches concurrentes proposent un découpage temporel des documents audiovisuels [Prie 98]. Si nous donnons la prééminence aux objets au sein de la vidéo, les segments temporels considérés sont alors liés à la durée de vie de l'objet et il s'agit d'une *stratification* [Smith 92]. La segmentation obtenue peut être alors temporellement redondante (différents segments temporels pouvant se chevaucher) et lacunaire (certains instants de la vidéo pouvant n'appartenir à aucun segment). Si, par contre, nous nous attachons en premier lieu à la structuration temporelle du document avant de considérer les objects présents au sein des entités temporelles, il est alors question de *structuration hiérarchique* du document audiovisuel. Une structuration est constituée de différents niveaux de découpage dont il est généralement admis qu'ils peuvent s'emboîter. Contrairement à la stratification, tout instant de la vidéo appartient, pour un niveau de segmentation donné, à un unique segment temporel.

En recherchant des macro-segments, nous nous sommes inscrits, comme nous allons le voir, dans une démarche de structuration hiérarchique, et c'est donc plus particulièrement de celle-ci qu'il sera question dans les sous-sections suivantes.

### 4.1.2 Les différents niveaux de granularité d'une structuration hiérarchique

Dans le cadre de la structuration hiérarchique d'un document audiovisuel, nous allons nous intéresser aux différents niveaux possibles de granularité du découpage temporel, et nous tenterons de définir succinctement les principales entités temporelles retenues.

#### *4.1.2.1 Le plan, une unité de référence limitée*

Le plan correspond à l'unité audiovisuelle élémentaire. Il peut être défini *physiquement* comme la "plus petite unité d'une continuité d'un film ou d'un produit vidéo sans coupure de caméra ou de raccord" [Chanal 93]. Cette définition liée à la prise de vue et au montage permet de définir le plan comme une entité facilement identifiable dans un document audiovisuel. En effet, la recherche d'un découpage en plans s'apparente par définition à la recherche de discontinuités physiques dans le flux vidéo. Le problème de la segmentation automatique du document audiovisuel en plans a donc été largement étudié. De nombreuses méthodes sont disponibles et ont donné des résultats tout à fait acceptables en dépit de l'utilisation de plus en plus fréquente de techniques d'édition diverses et de la détection souvent délicate de certaines transitions progressives contenant des effets spéciaux (voir sous-section 2.3.2).

Cependant, certains auteurs ont contesté la pertinence de l'utilisation du plan comme unité de référence pour l'indexation de documents audiovisuels.

Du point de vue de la compréhension globale et de l'analyse documentaire d'un document audiovisuel, la segmentation en plans est un niveau de découpage discutable. Les mouvements de caméra, l'extraction d'objets en mouvement, etc., sont des informations pertinentes pour l'indexation souvent associées à des segments temporels contenus dans le plan [Cherfaoui 95,Joly 96b]. Inversement, pour rendre compte d'une action, d'un décor, d'un dialogue, d'une progression narrative, il est nécessaire de replacer le plan dans un contexte plus large que celui d'un plan isolé [Joly 96a, Chap. IV].

D'un point de vue pratique, le découpage en plans fournit bien souvent un nombre de segments trop abondant, ce qui rend difficile son utilisation dans le contexte de l'indexation semi-automatique, ou de la navigation par le contenu. De plus, lorsqu'un plan est représenté par des images-clefs, il peut y avoir plusieurs images-clefs par plan afin de rendre compte des éventuelles variations des caractéristiques du plan [Dimitrova 95]. Par conséquent, la segmentation en plans n'offre pas une représentation facilement utilisable du document audiovisuel. À titre d'exemple, pour la retransmission sportive des championnats d'athlétisme de Munich[1] de 57 minutes et 25 secondes, la segmentation manuelle indique 550 plans environ et la segmentation automatique effectuée avec l'algorithme *MD_Shots* [Bouthemy 99b] développé par l'équipe VISTA à l'INRIA Rennes fournit un découpage en 780 plans[2].

#### *4.1.2.2 Définition des principaux niveaux de segmentation*

La segmentation liée à une analyse vidéo à l'intérieur du plan est appelée *micro-segmentation* et consiste principalement à retrouver les mouvements de caméras constitutifs du plan [Joly 96b, Bouthemy 99b]. La détection de segments regroupant des plans, que l'on nomme parfois "séquences", constitue la *macro-segmentation*[3].

Ces différents niveaux de segmentation emboîtés induisent une structuration hiérarchique du document audiovisuel comme présenté à la figure 4.1.

1. document *munich2*, voir annexe A.1 pour la description des documents audiovisuels cités.
2. Yeung *et al.* trouvent même un rapport de plus grand ordre dans [Yeung 96] avec 300 plans pour 15 minutes extraites du film *Terminator 2*.
3. Ce terme a été initialement proposé par P. Joly dans [Joly 96a].

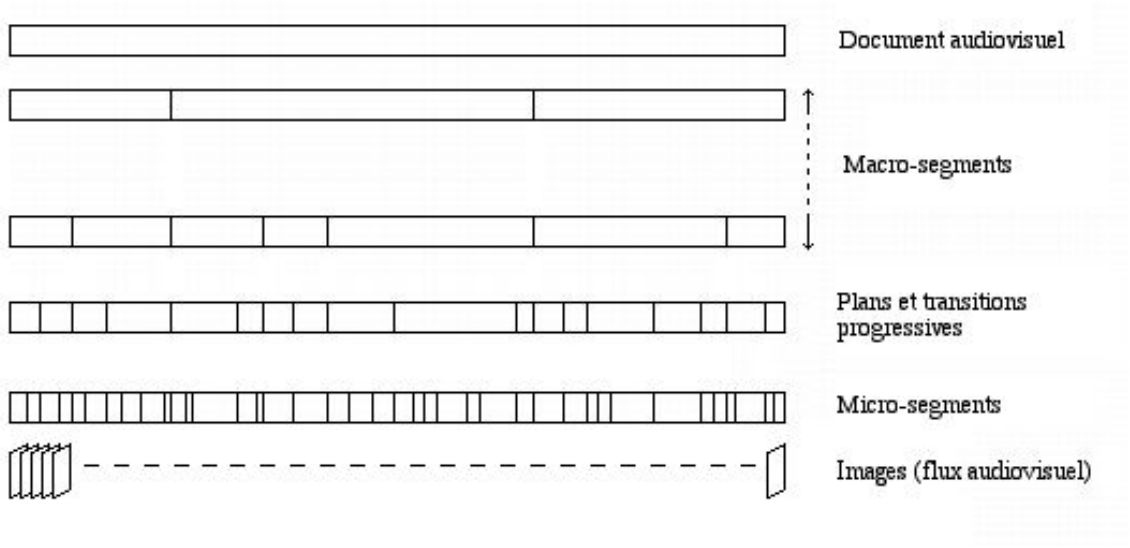



FIG. 4.1: *Résultat de la structuration d'un document audiovisuel en différents niveaux d'analyse*

### 4.1.3 Diversité de la notion de macro-segment

Le macro-segment ayant été défini comme un segment s'insérant entre le plan et le document dans la structuration hiérarchique de ce dernier, il convient d'essayer de préciser ce concept dans la mesure où les niveaux de granularité et les approches possibles restent nombreux et divers.

#### *4.1.3.1 Une notion controversée*

**De l'utilisation d'un vocabulaire ambigu...** Les travaux traitant de l'entité temporelle qu'est le macro-segment utilisent de nombreux termes parfois réducteurs, souvent ambigüs, preuve de la difficulté d'appréhender l'unité produite par la macro-segmentation. Les vocables utilisés peuvent être, en anglais : *act*, *episode*, *high-level segment*, *logical story unit*, *news article*, *news story*, *news unit*, *scene*, *sequence*, *shot cluster*, *story segment*, *story unit*, *theme*, *video paragraph*, etc.

**... lié à des contextes multiples...** L'utilisation d'un vocabulaire diversifié est sans aucun doute la marque de la difficulté des auteurs à cerner précisément ces objets temporels. Pour poser leur définition, ils ont ainsi recours à des références ou à des analogies à des domaines variés comme la musique (*theme* [Kender 98]), le théâtre (*act* [Aoki 96]) ou le texte (*video paragraph* [Hauptmann 95]), ou à des usages liés à d'autres pratiques du domaine audiovisuel (écriture du script [Hanjalic 99], production [Joly 96a], montage [Yeung 98], etc.)

**... d'un concept variable.** La diversité des champs lexicaux et des notions auxquelles se réfèrent les définitions du macro-segment n'est que l'expression de la diversité-même du concept à la fois selon les types de documents audiovisuels et au cours de l'histoire de la production audiovisuelle. En effet, la construction narrative des documents audiovisuels et les techniques de montage et de post-production ont évolué de manière notable depuis l'invention du cinéma muet jusqu'aux productions contemporaines, modifiant profondément les usages liés à la narration et à la structuration des

différents types de documents étudiés[4]. La diversité est aussi dans la variété des formes et des contenus des documents audiovisuels (journaux télévisés, documentaires, films de fiction, magazines, talk-shows, retransmissions sportives, émissions de variété, clips musicaux, etc.), ainsi que dans les styles différents, propres aux auteurs (notamment dans les documents de fiction). Cette variabilité du concept de macro-segment rend difficile, voire impossible, toute définition à la fois générale et suffisamment descriptive pour guider *a priori* la définition d'algorithmes de macro-segmentation.

#### *4.1.3.2 Tentative de définition* a minima

A défaut d'apporter une réponse définitive, nous nous contenterons pour l'instant d'une définition très générale en attendant de proposer, selon les types de documents étudiés, ce que pourraient être les macro-segments pertinents (voir sous-section 6.2.1). Ainsi, nous nous rallions provisoirement au consensus *a minima* qui semble se faire autour d'une définition proche de celle de la scène pour le théâtre donnée dans [Chanal 93] : "subdivision d'une continuité audiovisuelle qui est définie, le plus souvent, par unité de lieu et de temps". À la lecture de cette définition, il semble généralement admis que le macro-segment correspond soit à une unité de lieu (liée éventuellement à une homogénéité du décor), soit à une unité d'action (liée par exemple au montage ou à des primitives de mouvement).

Quant au choix du vocabulaire utilisé, nous utiliserons quant à nous les termes de *macro-segment* ou de *séquence* lorsque nous souhaiterons insister sur l'aspect temporel de l'unité.

#### *4.1.3.3 Difficultés spécifiques de la macro-segmentation*

L'absence de définition universelle du macro-segment, la diversité des notions qui lui sont liées, et le niveau sémantique auquel le concept se réfère font que la mise en œuvre d'outils automatiques de macro-segmentation est un problème intrinsèquement difficile. À titre d'exemple, citons deux difficultés classiques liées à la macro-segmentation dont la résolution est beaucoup plus cruciale que dans le cas d'une segmentation en plans.

La macro-segmentation est un exemple-type de tâche d'indexation automatique où le saut qualitatif à franchir entre le niveau physique du signal et le niveau sémantique du concept recherché est pour le moins problématique (cf. section 2.2). Ainsi, le niveau d'abstraction définissant le macro-segment peut être tel que certains éléments le caractérisant ne sont pas directement accessibles, ou simplement ne sont pas continûment présents dans la séquence temporelle.

Autre exemple, comment définir les entités liées à la notion de macro-segment, et notamment comment appréhender la notion de transition progressive entre deux macro-segments ? (Sur ce sujet, on pourra se référer aux réflexions menées dans [Kender 98]).

### 4.1.4 Applications de la macro-segmentation

Outre l'intérêt initial qu'il y aurait pour la chaîne de documentation de l'INA, les principales applications de la macro-segmentation, au travers de la mise en œuvre d'une structuration hiérarchique du document audiovisuel à laquelle elle participe, sont notamment : la navigation et l'accès non linéaire au contenu [Yeung 96], la représentation des documents [Aoki 96,Yeung 97], la création de résumés de vidéos (ou bande-annonces) [Saarela 99,Lacoste 02], la génération d'index ou de tables des matières [Rui 99b,Llach 99], la recherche d'événements notables, la classification des séquences

4. Ainsi, des auteurs se référant à des théories audiovisuelles datées, comme dans [Yeung 98] où est évoquée la théorie du montage d'Eisenstein, aboutissent à des algorithmes limités lorsqu'il s'agit de structures complexes ou de montages moins dogmatiques, ou simplement plus récents.

selon leur genre [Vasconcelos 98a], ou la requête par le contenu à différents niveaux de granularité [Vasconcelos 97b].

Une structuration hiérachique implique la détermination de *plusieurs* niveaux de macro-segments emboîtés. Certains algorithmes proposent de tels résultats soit en fournissant directement plusieurs niveaux de macro-segmentation [Kender 98,Hammoud 98], soit en utilisant les paramètres disponibles pour spécifier un niveau de granularité donnée (voir sous-section 4.2.5).

Par ailleurs, certaines des méthodes développées ont donné lieu à la création d'interfaces visuelles [Yeung 96,Aoki 96,Vasconcelos 98b,Kender 98], d'autres ont été conçues comme des outils algorithmiques au sein de projets ou systèmes plus vastes, comme par exemple les projets *Informedia* [Hauptmann 95] et *Movie Content Analysis* (MoCA) [Lienhart 99], le système *Broadcast News Editor and Navigator* (BNE-BNN) [Merlino 97].

Dans le cadre de l'INA (voir section 1.3), outre une proposition de découpage en séquences éventuellement intégrable à *Médiascope*, la navigation pourrait être facilitée dans *Thesaurus Rex* par la visualisation d'images-clefs au niveau des séquences permettant une identification plus facile des extraits d'intérêt. Les macro-segmentations proposées pourraient aussi permettre d'accéder à des requêtes de type "factuel" (recherche d'événements, comme par exemple les épreuves dans les retransmissions sportives), et de type "individu" (par exemple, les interviews ou les interprétations dans les émissions de variété). Toujours à titre d'exemple, l'usage scientifique des documents pourrait être facilité par le découpage des journaux télévisés en reportages permettant la comparaison de traitement d'une information dans différents contextes. Enfin, dans le cadre d'une consultation d'une partie du fonds sur l'internet, le résumé de vidéos souvent construit à partir des décors, acteurs et moments principaux du document, apparaît comme une application naturelle de la segmentation proposée pour la fiction.

## 4.2 État de l'art des techniques de macro-segmentation

Compte tenu de la diversité de la notion de macro-segment, il est peu aisé de proposer une vue synthétique des méthodes de macro-segmentation. Différentes grilles de lecture ont été proposées, certains différencient les méthodes liées aux domaines compressé ou décompressé, d'autres distinguent des méthodes dites "sémantiques" et celles dites "syntaxiques" [Gunsel 98a].

Dans le cadre de la structuration hiérarchique, nous avons souhaité mettre l'accent sur les principales approches soutendant les méthodes mises en œuvre, et nous avons ainsi identifié trois grandes familles de techniques :

1. méthodes proposant un regroupement des plans fondé sur une similarité à la fois physique et temporelle de ceux-ci ;
2. méthodes fondées sur l'utilisation d'informations *a priori* ;
3. méthodes fondées sur une utilisation conjointe de différents types d'informations présents dans le document audiovisuel.

Par ailleurs, nous avons souhaité intégrer aussi quelques-unes des méthodes liées à une approche par stratification dans la mesure où ces méthodes aboutissaient à l'obtention de séquences temporelles d'un niveau d'abstraction assez proche de celui des macro-segments.

Bien évidemment, certaines méthodes procèdent de plusieurs de ces approches. Il nous a paru cependant pertinent de structurer ainsi notre état de l'art dans la mesure où chaque approche propose une appréhension particulière des problématiques soulevées. Nous verrons aussi dans les

commentaires que nous pourrons apporter que, derrière des formalisations différentes, ce sont bien souvent des intuitions semblables qui sont mises en œuvre, et que les méthodes tendent de plus en plus à proposer des synthèses de ces différentes approches.

### 4.2.1 Approche par stratification

Cette approche est défendue par ceux qui s'intéressent à des applications (comme la surveillance ou la production) où l'importance de l'objet vidéo est primordiale. L'idée est alors de substituer aux notions de plans et de séquences la notion de "durée de vie" de l'objet (*lifespan of the video object* [Gunsel 98b]). La méthode CBSD (*component-based scene description* [Shibata 92]) propose de partir d'une stratification, constituant un *script* de la présence d'objets, personnes ou actions, effectuée manuellement. La construction des séquences se fait alors par regroupement hiérarchique des segments temporels adjacents fondées sur l'indication de la présence ou de l'absence des objets retenus lors de la description. Cette configuration organisant la présence de descripteurs le long d'axes temporels est aussi adoptée dans [Vasconcelos 98b], où des descripteurs plus ou moins abstraits sont extraits automatiquement par une méthode bayésienne, mais où la segmentation en séquences n'est que suggérée et laissée à l'appréciation de l'utilisateur.

### 4.2.2 Approche par regroupement de plans fondée sur une similarité contrainte temporellement

Afin de regrouper les plans en séquences, des travaux ont d'abord envisagé une procédure manuelle [Cherfaoui 94], puis des études de la similarité physique des plans permettant leur regroupement sans prise en compte de leur agencement temporel ont été menées [Yeung 95].

Les travaux qui nous intéressent plus particulièrement reposent sur les idées suivantes : (i) l'hypothèse d'une forte corrélation entre le contenu physique et le contexte sémantique des plans motive le regroupement des plans selon des critères (descripteurs et distances) physiques ; (ii) la disposition temporelle des plans est indispensable à la compréhension et à l'appréhension de l'organisation des plans et devra être prise en compte par la méthode ; (iii) la mise en œuvre des algorithmes ne nécessite pas d'informations *a priori*.

Plusieurs stratégies de classification des plans ont été étudiées. Elles utilisent surtout des primitives simples de couleurs, mais aussi, pour certains, de mouvement. Les principales méthodes de classification se fondent sur la construction d'un hiérarchie ascendante binaire [Yeung 98,Veneau 00], sur un indice de similarité évolué [Aoki 96], sur le seuillage adaptatif d'une similarité entre les images-clefs représentatives des plans [Hanjalic 99], ou sur un indice de similarité composite et une classification de type *k-moyennes* [Rui 98].

Notons que l'information temporelle est présente dès ce stade à travers l'utilisation d'une fenêtre temporelle. La similarité physique des plans n'est supposée avoir un sens que si les plans sont suffisamment proches temporellement. Dans [Yeung 98,Aoki 96,Hanjalic 99], au-delà d'une fenêtre temporelle donnée, la similarité entre plans devient automatiquement nulle. Dans [Rui 98], une contrainte temporelle plus douce est introduite par le calcul d'une attraction temporelle entre plans. En outre, dans certains travaux, la durée des plans est considérée comme une primitive en tant que telle.

Les classes de plans sont ensuite regroupées afin de former les macro-segments. Les stratégies sont à nouveau diversifiées : simple regroupement des classes alternées dans [Aoki 96,Hanjalic 99], seuillage d'un indice de similarité et classification de type nuées dynamiques [Rui 98], ou construction d'un graphe spatio-temporel (STG) dont les nœuds sont les classes et les arcs orientés la

succession temporelle des plans et dont on recherche les composantes connexes [Yeung 98]. Dans tous les cas de figure étudiés, les segments obtenus sont temporellement continus par construction.

Une autre méthode liée à des modèles psychologiques a été proposée dans [Kender 98], afin de pallier les limites de la méthode présentée dans [Yeung 96]. Les ruptures entre plans sont associées au calcul d'un critère de cohérence entre plans fondé sur leur similarité physique, leur longueur et l'éloignement des images qui les constituent. Un seuillage des *minima* locaux sur une fenêtre temporelle permet alors de déterminer des segments temporels continus de "haut-niveau". En dépit de son atypisme, cette méthode trouve naturellement sa place dans ce paragraphe, les hypothèses et les principes étant similaires : similarité physique contrainte par le temps.

### 4.2.3 Approche liée à l'utilisation d'informations *a priori*

Les méthodes fondées sur l'utilisation d'informations *a priori* requièrent de mener, au préalable, une analyse formelle des documents audiovisuels étudiés. Nous pouvons distinguer deux formalisations sensiblement équivalentes selon que l'information *a priori* est utilisée sous forme de modèle, ou *via* un ensemble de règles. Les règles sont issues d'une analyse de la construction des documents, et rendent compte à travers des relations entre objets audiovisuels d'une sémantique du document. Les modèles formalisent la typologie des documents que l'utilisateur est susceptible de rencontrer, dans un domaine parfois restreint.

Cette approche est fréquemment utilisée lorsque l'utilisateur est confronté à ce qu'il est convenu d'appeler une collection de documents [Carrive 00], c'est-à-dire des documents correspondant à une typologie précise, dont la structure est globalement constante d'un document à l'autre et peut être aisément modélisable. Le journal télévisé étant à la fois un domaine commercialement porteur et un exemple parfait de ce que peut être une collection de documents, les travaux proposant d'en retrouver automatiquement la structure ont été particulièrement nombreux.

Les objets utilisés pour la modélisation et extraits automatiquement sont souvent assez simples : détection des images de transition de couleur noire, des logos et éventuellement des visages pour les objets visuels et détection des silences et éventuellement de la parole pour le son. De plus, certaines méthodes utilisent surtout du texte au travers des transcriptions, des sous-titres ou des textes incrustés. La principale difficulté est alors de réaliser une transcription même approximative de la bande-son, ou l'alignement temporel du texte, s'il est obtenu par un autre moyen, sur le flux audiovisuel.

Les modèles et règles proposés sont de complexité diverse et reposent généralement sur l'alternance des plans de studio, de publicité et de reportage. Les méthodes les plus simples mettent en correspondance certaines séquences et différents modèles simples des plans de studio et des publicités. D'abord mis en œuvre par [Swanberg 93,Zhang 94], ces modèles ont été largement repris et adaptés, parfois pour effectuer de simples labellisations de plans [Low 96,Ide 98,Merialdo 98]. Lorsque les plans de studio sont détectés et regroupés, les reportages sont définis comme le regroupement des plans intermédiaires [Ariki 96], à moins que des règles de regroupement relativement simples ne soient proposées [Gunsel 98a,Hauptmann 98]. Dans [Merlino 97], l'information *a priori* est modélisée par six états, et la macro-segmentation est effectuée par une machine à états finis. Ce type de méthodes a aussi inspiré quelques travaux dont une adaptation pour les documents sportifs [Gong 95], et une tentative de généralisation aux documents de fiction [Yoshitaka 97].

D'autres travaux plus ambitieux ont proposé des formalisations de l'information *a priori* plus génériques [Joly 96a] ou des cadres théoriques d'utilisation de l'information moins *ad hoc* [Carrive 00].

Dans [Joly 96a], la méthode de macro-segmentation est fondée sur des règles et nécessite l'ex-

traction de primitives des flux audio et vidéo, ainsi que des formalisations de techniques de montage. Neuf règles, issues de l'étude approfondie de documents audiovisuels, de la théorie des films, et de discussions avec des professionnels ont finalement été définies dans [Aigrain 97]. Trois règles sont liées aux effets de transition, trois aux similarités entre plans, une au montage, une à la bande son, une aux mouvements de caméra. L'application de ces règles permet de determiner :

- un ensemble de limites précises des séquences ;
- des regroupements de plans supposés être dans une même séquence ;
- un ensemble de plans considérés comme importants .

L'étape finale consiste alors en :

- la synthèse des indications compatibles de début et fin de séquence ;
- le traitement des conflits ;
- le traitement des segments où aucune règle n'a pu s'appliquer ;
- le choix des images représentatives.

La méthode étudiée dans [Carrive 98] se situe dans un contexte plus général que celui de la macro-segmentation. Il s'agit de présenter un cadre de coopération entre des informations de niveaux différents dans l'optique, notamment, de permettre le pilotage des algorithmes d'analyse automatique des documents audiovisuels [Carrive 00]. La macro-segmentation, à proprement parler, repose sur une taxonomie des éléments de la production filmique, notamment des éléments liés à la prise de vue et au montage, ainsi que sur une taxonomie des types de documents, liée à l'existence de canevas relativement précis (comme dans les actualités, émissions de variété, émissions sur le cinéma). Ces règles structurelles et les connaissances spécifiques associées à une collection de documents permettent de définir un modèle de document et de dégager des régularités dans les structures. Un formalisme de logiques de descriptions étendu, notamment, aux relations de Allen et à un opérateur d'itération permet une structuration du document à partir du modèle construit et de primitives de niveaux informationnels variables extraites du flux.

Notons, enfin, la méthode proposée dans [Hammoud 98]. Une première étape permet le regroupement des plans similaires en classes à partir de descripteurs physiques et d'une contrainte temporelle liée à la présence des fondus. Un graphe est alors construit dont les nœuds sont les classes et les arcs orientés sont les relations temporelles entre classes, formalisés par les relations de Allen étendues (*meet*, *before*, *overlap*, *during*). Notons que cette étape est assez semblable à la constrution du STG dans [Yeung 96], un STG où les relations entre classes seraient enrichies de la relation *before*. Des règles de segmentation sont ensuite associées à chacune des relations temporelles permettant la structuration hiérarchique du document en *scènes*, puis en *séquences*.

### 4.2.4 Approche fondée sur une coopération entre primitives extraites

Les méthodes présentées dans cette sous-section utilisent *conjointement* différentes primitives issues principalement des domaines audio et visuel. L'objectif est d'enrichir la description obtenue afin de produire une segmentation plus robuste et plus pertinente d'un point de vue sémantique. Certaines de ces méthodes suivent une approche incrémentale, dans la mesure où des entités liées à des niveaux d'abstraction plus élevés que ceux de simples descripteurs sont extraites avant de déterminer les macro-segments.

Les méthodes les plus simples fusionnent, après repèrage des ruptures dans les flux audio et vidéo, les segmentations de ces flux [Hauptmann 95,Nam 97]. Dans [Saraceno 98], l'auteur recherche des similarités entre plans fondées sur le contenu visuel, mais l'étude de la succession des labels des classes de plans, en vue de la construction de séquences de dialogue, d'action, narratives ou génériques, est contrôlée par la caractérisation de la bande sonore. Ces méthodes donnent souvent l'avantage au contenu audio dans la détermination des séquences, puisqu'il semble qu'en pratique les ruptures de plans vidéo deviennent des ruptures de séquences s'il y a concomitamment une discontinuité dans la bande sonore. Notons que l'idée d'une forte corrélation entre image et son sur laquelle repose ces méthodes est elle-même contestée dans certains cas [Saraceno 97].

D'autres méthodes, comme celle décrite dans [Lienhart 99], consistent à regrouper, dans un premier temps, les plans selon différents critères. Des séquences audios à partir de primitives sonores, des séquences de dialogues construites d'après la détection et la classification de visages ainsi que de règles, et des séquences vidéos liées au décor sont déterminées indépendamment les unes des autres. Ces regroupements sont ensuite utilisés conjointement pour construire des séquences par fusion des segments temporels non disjoints. Une contrainte temporelle est introduite lors des regroupements de plans, ainsi qu'en introduisant des ruptures de séquence au niveau de chaque fondu. Ces approches incrémentales sont semblables aux techniques par stratification, par certains aspects, tels que l'utilisation d'objets vidéos ou de dialogues.

Dans [Adams 00], l'auteur suggère l'utilisation conjointe, au sein d'un critère nommé *tempo*, de l'information de mouvement et du rythme d'édition (la longueur des plans) lié au montage. La fusion des informations est faite par simple pondération additive et la segmentation par un seuillage du critère ainsi défini.

### 4.2.5 Gestion de la paramétrisation des algorithmes

Nous allons brièvement évoquer la signification et la gestion des différents paramètres présents dans les outils de macro-segmentation.

Pour les méthodes fondées sur la similarité physique contrainte par le temps, il y a deux paramètres principaux, le premier lié au niveau de similarité, le second à la contrainte temporelle. Dans [Yeung 98], $\delta$ est le seuil de dissimilarité maximale au-dessus duquel les plans ne pourront être regroupés dans la hiérarchie, et $T$ est la taille de la fenêtre temporelle. $T$ peut être interprété comme le diamètre temporel maximal d'une classe au sein de la hiérarchie, mais il n'indique pas une taille maximale de macro-segments. *A contrario*, $T$ doit être obligatoirement inférieur à l'éloignement temporel de deux plans physiquement similaires n'appartenant pourtant pas au même macro-segment. Les autres méthodes de cette famille fonctionnent globalement sur le même modèle. Dans [Aoki 96], il y a quatre paramètres liés à la similarité et un à la contrainte temporelle. Dans [Rui 98] quatre paramètres apparaissent dans le critère de similarité (dont deux calculés automatiquement) et un, quoiqu'en dise l'auteur, lié à la contrainte temporelle. Dans [Hanjalic 99], deux paramètres relèvent de la contrainte temporelle, et deux de la similarité. Enfin, dans [Kender 98], il y a deux paramètres : la taille de la mémoire des similarités entre plans, et la taille de la fenêtre temporelle utilisée pour la recherche des *minima* locaux du critère de cohérence.

Lorsque les paramètres ne peuvent être calculés automatiquement, il convient de pouvoir les fixer intuitivement afin que l'interaction avec l'utilisateur ou le pilotage d'algorithmes soit rendu aisé. Les deux principaux paramètres utilisés semblent assez intuitifs. Liés aux similarités physique et temporelle, ils influent sur la granularité de la segmentation obtenue. Toutefois, cette influence n'est pas toujours facile à gérer dynamiquement. L'utilisation des paramètres dans [Yeung 98] a notamment été largement discutée : les critiques se focalisant sur la nature binaire de la contrainte

introduite par $T$ et le manque de synergie dans l'utilisation de $\delta$ [Kender 98]. Des efforts ont été faits, notamment dans [Rui 98], afin soit de rendre l'influence de ces paramètres plus continue, soit de proposer des formulations plus généralistes, le paramètre devenant la combinaison d'une constante générique et de données liées au document et calculées automatiquement.

Pour les autres familles de méthodes, il n'y a pas à proprement parler de paramétrage. Toutefois, il est tentant de considérer que dans le cas de l'utilisation d'informations *a priori*, la définition du modèle ou des règles n'est finalement qu'un ajustement plus ou moins complexe de nombreux paramètres, et que, dans les cas de l'utilisation conjointe de primitives et de la stratification, cette difficulté est repoussée au niveau de l'extraction paramétrée des primitives.

La gestion des paramètres est une question cruciale et sensible dans le développement d'algorithmes d'indexation automatique sur laquelle nous reviendrons, pour ce qui concerne la macro-segmentation, lors de la présentation des résultats (voir sous-section 6.3.1).

### 4.2.6 Comparaison et évaluation des différentes approches

Lorsqu'elles existent, les expérimentations effectuées à partir des méthodes précédemment décrites sont de deux sortes : soit quelques expérimentations ont été réalisées sur des extraits de vidéos limités et disparates, parfois de manière strictement qualitative, soit des résultats obtenus sur des heures de vidéo sont communiqués, sans toutefois que des conclusions claires et objectives puissent en être tirées.

La variété des approches augmente la difficulté d'évaluer les différentes méthodes ; nous allons cependant essayer d'apporter quelques éléments de comparaison.

#### *4.2.6.1 Difficultés d'une évaluation des outils automatiques de macro-segmentation*

Commençons par une remarque un peu provocatrice. Les techniques existantes s'intéressant à des objets qui ne sont pas strictement identiques, et ce dans des contextes différents, est-il pertinent de vouloir comparer leurs performances ? Ainsi, quel intérêt y aurait-il à comparer un outil dédié qui s'apparente à un parseur de reportages d'actualité et un outil de macro-segmentation à visée générale ?

La deuxième difficulté rencontrée lors de l'évaluation de ces outils est la définition d'un corpus. Il n'y a pas de corpus commun disponible et les corpus utilisés sont très disparates qu'il s'agisse de la variété ou de la longueur des documents traités. Ainsi, la longueur des documents considérés peut varier entre quelques dizaines de secondes [Nam 97] et une heure [Kender 98]. De même, certains auteurs ont mené des expérimentations sur un extrait de vidéo, d'autres sur des corpus dédiés (notamment des journaux télévisés de CNN, NHK ou SBC), ou des corpus généralistes (films d'action, de science-fiction, comédies, extraits musicaux) [Rui 99b]. Enfin, lorsque des expérimentations ont été réalisées sur des corpus conséquents, l'absence d'information concernant la constitution de la macro-segmentation de référence empêche toute reproduction des tests ou toute comparaison des résultats.

Dernière difficulté notable, les indicateurs de performance ne sont pas toujours homogènes (même s'il s'agit bien souvent d'indicateurs d'oublis ou de fausses détections), ce qui ne facilite pas une étude comparative des différents outils.

Une évaluation comparative objective se heurte ainsi au niveau d'abstraction de l'objet *macro-segment*, à la difficulté même de la constitution d'une annotation manuelle de référence, à la subjectivité de la notion de macro-segment. Ceci implique d'avoir recours à des analyses qualitatives difficile à mettre en œuvre, et à des indicateurs numériques forcément insatisfaisants. Dans le cadre de nos travaux, nous aborderons ces problématiques plus en détail à la section 6.2.

#### *4.2.6.2 Éléments de comparaison*

Les méthodes utilisant une similarité entre plans contrainte temporellement sont présentées comme des méthodes génériques, indépendantes des domaines d'utilisation et ne nécessitant aucune information *a priori*. Cette assertion doit être tempérée par les difficultés liées au réglage des paramètres. Les valeurs des paramètres sont en fait dépendantes du domaine, voire du document. Ainsi, la paramétrisation peut être une manière de tenir compte d'éventuelles informations *a priori*. Autre inconvénient, les coûts de calcul des similarités entre plans et de la classification s'avèrent souvent élevés. Ces algorithmes sont par ailleurs peu robustes aux erreurs de classification, notamment dans le cas de plans faussement regroupés. Enfin, certains cas de figure sont plus difficilement pris en compte par ces méthodes, comme les séquences montées en parallèle, les transitions entre macro-segments, les plans similaires proches temporellement et non liés sémantiquement. *A contrario*, les séquences alternées (dialogue, champ contre-champ, etc.) semblent bien détectées.

Les outils incorporant des informations *a priori* sont pour la plupart dédiés à un domaine d'application particulier, ce qui constitue à la fois leur efficacité et leur limite. L'utilisation de formalisations d'un niveau d'abstraction élevé laisse espérer des analyses plus fines et plus appropriée des documents. Dans le cas de collections de documents décrites par un modèle commun, ce qui est le cas à l'INA, ces méthodes peuvent être séduisantes. Toutefois, la construction des formalisations n'est pas toujours évidente et induit de nombreux travaux préliminaires. Pour les méthodes visant une certaine généricité, on peut notamment s'interroger sur la réelle généralité des modèles, et la flexibilité de ceux-ci pour les différents types de documents, ainsi que face à l'ajout ou à l'évolution des contraintes considérées. Enfin, la dépendance des règles vis-à-vis des performances des traitements préalables du document audiovisuel peut rapidement devenir un inconvénient majeur, compte tenu notamment du niveau d'abstraction de certains des éléments requis.

Les méthodes procédant par stratification ou utilisation conjointe de primitives semblent prometteuses par la richesse de l'information rendue disponible. Toutefois, le niveau d'abstraction des éléments extraits et la fiabilité de leur détection peuvent être des difficultés à prendre en compte. Ainsi certains auteurs, en l'absence d'extraction automatique disponible des descripteurs utilisés, ont eu recours à des annotations manuelles afin de démontrer la validité de leurs méthodes [Shibata 92]. La mise en œuvre de la coopération entre primitives peut s'avérer également problématique : les solutions simples comme la fusion des différentes segmentations ne sont pas toujours convaincantes, et le recours à des distances pondérées crée souvent des difficultés.

Il convient, toutefois, de relativiser la classification proposée des méthodes de macro-segmentation en remarquant que les principales notions sont finalement reprises par la plupart des auteurs sous différentes formalisations. Nous avons vu plus haut que l'utilisation de l'information *a priori* était en fait "généralisée" par l'utilisation de paramètres. Il est intéressant de noter également que les règles proposées dans [Aigrain 97] s'appuient aussi sur le principe du regroupement des plans similaires (règles 6 et 9) et ce, dans une certaine limite temporelle (règle 4). Dans le même ordre d'idée, Yeung *et al.* ont appliqué dans [Yeung 97] leur méthode de classification des plans contrainte temporellement, à la recherche de regroupements particuliers de plans comme les dialogues ou les actions, dépassant ainsi le cadre un peu rigide de leurs unités narratives, et se rapprochant ainsi des démarches par stratification. Cette approche a aussi été adoptée par [Saraceno 98] dans des travaux intégrant le traitement du flux audio. La tendance générale, dans les travaux récents, est d'utiliser des primitives plus diversifiées ou issues de niveaux d'abstractions plus élevés.

# Chapitre 5

# Définition d'une méthode de macro-segmentation

Après ce tour d'horizon des méthodes de macro-segmentation, nous exposons dans ce chapitre la stratégie que nous avons choisie. Une description globale est donnée ci-après, et les choix effectués sont justifiés. Les différents modules sont ensuite détaillés dans les sections suivantes.

## 5.1 Analyse fonctionnelle du problème

La chaîne de traitement que nous avons définie pour la macro-segmentation est synthétisée par le schéma synoptique de la figure 5.1.

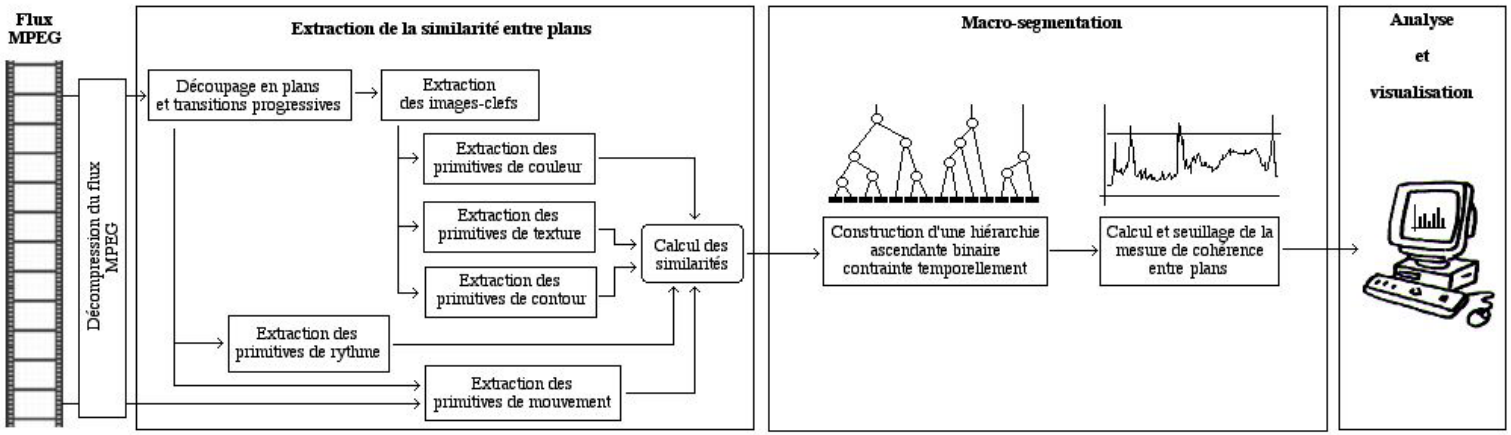
Nous avons inscrit notre démarche dans le cadre de la structuration hiérarchique, et nous avons fondé notre approche sur la connaissance des similarités entre plans. Afin de rendre compte au mieux de l'information présente dans le flux vidéo, notamment des notions de décors, d'activité et de montage, nous avons utilisé des primitives qui leur sont associées : couleur, mouvement et rythme. Dans une optique plus expérimentale, nous avons enrichi l'information disponible avec des descripteurs de texture et de contour. Par contre, pour des raisons pratiques, nous n'avons eu recours à aucun descripteur du flux audio.

Concernant l'algorithme de macro-segmentation, la hiérarchie ascendante proposée dans [Yeung 96] nous a semblé à même de donner une bonne représentation des similarités entre plans et ainsi constituer une base d'étude intéressante. Nous avons développé une méthode s'inspirant de cet algorithme, en y apportant un certain nombre d'améliorations.

## 5.2 Détermination de la similarité entre plans

### 5.2.1 Découpage en plans et extraction d'images-clefs

Nous avons choisi de travailler à partir de deux types de segmentation en plans et transitions graduelles : (i) une segmentation réalisée manuellement afin de pouvoir développer notre outil sans trop se préoccuper dans un premier temps de la fiabilité des prétraitements ; (ii) une segmentation fournie automatiquement nous permettant de réaliser la chaîne complète de traitements, et d'évaluer la robustesse de notre outil aux erreurs de découpage en plans. Les segmentations manuelles des plans utilisées proviennent d'un groupe de travail de l'Action Indexation Multimédia (AIM) du GDR-PRC ISIS [Ruiloba 99], ou ont été effectuées en interne au GRAMM à l'INA. L'ou-




TAB. 5.1: *Présentation synthétique de la méthode de macro-segmentation, comprenant trois étapes principales : (i) l'extraction des primitives ; (ii) la classification des plans et le calcul d'un critère de cohérence entre ceux-ci ; (iii) la visualisation des résultats.*

til de découpage automatique en plans et transitions progressives utilisé est le logiciel *MD_Shots*, développé par l'équipe VISTA à l'INRIA Rennes [Bouthemy 99b].

Considérant que l'information est principalement contenue dans les plans, et que les transitions graduelles donnent plutôt des informations sur la forme du document audiovisuel, seuls les plans seront pris en compte lors du calcul des similarités. Les transitions graduelles seront à nouveau considérées, dans la dernière phase, lors de la construction des macro-segments (voir sous-section 5.3.3).

L'extraction des images-clefs représentatives des plans est effectuée par un algorithme développé au sein du GRAMM de l'INA (voir annexe E). Elle s'appuie sur une méthode de classification des images par un algorithme de k-moyennes. Un plan peut être représenté par plusieurs images-clefs si la diversité de son contenu le justifie. Chaque image-clef extraite possède un cœfficient de pondération lié à sa représentativité au sein du plan. La réduction d'information ainsi obtenue est de 5 images-clefs pour 400 images en moyenne.

### 5.2.2 Extractions des primitives

Les primitives extraites sont liées à des informations de couleur, de mouvement, de rythme d'édition, de texture et de coutour. Ce sont, pour la plupart, des descripteurs globaux de l'image censés synthétiser l'information présente.

#### *5.2.2.1 Descripteurs de couleur*

L'information de couleur dépend des caractéristiques du décor et des objets présents dans l'image. L'extraction de descripteurs de couleur, globaux ou locaux, a été très largement étudiée (voir section 2.3.1). Nous avons développé une librairie permettant le calcul d'un certain nombre de ces primitives et, après une étude préliminaire (voir sous-section 3.1), nous avons retenu les signatures suivantes :

- l'histogramme global des couleurs tel qu'il est défini dans [Swain 91], noté $H_{col}$, calculé dans l'espace $RGB$. Le nombre de valeurs de quantification a été fixé à 8 par coordonnées dans l'espace des couleurs ;
- l'histogramme localisé par région, noté $H_{reg}$, correspondant à la concaténation des histogrammes de couleurs sur 12 régions (4 horizontalement et 3 verticalement) rectangulaires et disjointes ;
- le vecteur des couleurs cohérentes (*Color Coherent Vector*), noté $CCV$, qui précise l'information véhiculée par l'histogramme de couleur en séparant les points situés dans des régions de couleur homogène et les points isolés [Pass 96]. Le seuillage utilisé dans le regroupement des points a été fixé à la valeur 5 ;
- l'auto-corrélogramme des couleurs qui introduit une notion de texture [Huang 97], noté $AutoCrlg$. Les paramètres de cette signature sont les espacements de points considérés, par défaut nous avons retenu 3 valeurs : $\{1, 3, 5\}$ ;
- les couleurs dominantes, proposées par [Hsieh 00], notées $DomCol$. L'extraction de cette signature a été mise en œuvre par la méthode des nuées dynamiques et le nombre maximal de couleurs retenues a été fixé à 10.

L'ensemble de ces signatures ont été rendues indépendantes du nombre de points présents dans une image et normalisées dans $[0, 1]$. Les primitives de couleurs sont calculées sur les images-clefs extraites précédemment. La signature au niveau du plan est une simple moyenne pondérée par le cœfficient de représentativité associé à chacune des images-clefs représentant le plan.

#### *5.2.2.2 Descripteurs du mouvement*

L'information de mouvement est importante pour rendre compte des différents types d'activités présents à l'image, qu'il s'agisse de mouvements relatifs à la prise de vue (mouvements de la caméra), ou d'informations liées au contenu de la scène (objets en mouvement). Nous avons utilisé le logiciel *RMR*, dédié à l'estimation robuste de modèles paramétriques 2D de mouvement, et développé par l'équipe VISTA de l'INRIA Rennes [Odobez 95]. À partir de l'estimation du mouvement dominant entre images supposé être dû au mouvement de la caméra, un traitement ultérieur permet de caractériser le mouvement de caméra et l'activité des objets dans le plan [Bouthemy 99b]. Ainsi, les outils *C_Motion* et *S_Activity* extraient les informations suivantes : (i) une classe[1] de mouvements de caméra, déduite de l'estimation des paramètres du mouvement dominant ; (ii) un indicateur numérique d'activité, calculé à partir de statistiques simples sur la carte des "outliers" issue du calcul du mouvement dominant. À partir de ces mesures et d'une segmentation en plans, nous avons extrait les deux descripteurs suivants :

- l'histogramme des mouvements de caméra, noté $H_{cam}$, aussi utilisé par [Llach 99] et défini pour un plan $S_n$ par $H_{cam}(S_n)[i] = \frac{N_i}{N_f(S_n)}$, où $i \in \{\text{statique, complexe, zoom, panoramique}\}$, $N_i$ est le nombre d'images associées à un mouvement de type $i$ et $N_f(S_n)$ le nombre d'images dans le plan. Ce descripteur est censé rendre compte de la nature du mouvement de la caméra et est normalisé dans $[0, 1[$ ;
- l'activité du plan, notée $Act_{Sh}$ et définie simplement pour un plan $S_n$ par $Act_{Sh}(S_n) = \frac{\sum_{i=1}^{N_f(S_n)} Act(i)}{N_f(S_n)}$, où $Act(i)$ est l'indicateur d'activité évoqué plus haut pour l'image $i$. Cette primitive reflète le "niveau d'activité" des éléments de la scène, et est normalisée dans $[0, 1[$.

#### *5.2.2.3 Descripteurs du rythme d'édition*

Afin de prendre en compte un descripteur temporel lié au montage et à la construction des documents audiovisuels, nous avons considéré, comme [Adams 00], un descripteur noté *RythmEdit* et défini pour un plan $S_n$ par $RythmEdit(S_n) = N_f(S_n)$.

#### *5.2.2.4 Descripteurs de texture*

Pour l'extraction d'information de texture, un algorithme fondé sur la méthode décrite dans [Manjunath 96] a été développé au GRAMM de l'INA. Nous avons simplifié cette méthode utilisant des filtres de Gabor bidimensionnels, en réduisant le banc de filtres à quatre filtres correspondant à $\sigma_g = 4.0$ et $\omega_s = 0.5$ pour les quatre orientations suivantes : $\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$. Nous avons ensuite pris en compte les deux descripteurs suivants :

- les moments principaux issus des quatre filtres retenus comme indiqué dans [Manjunath 96], notés $Stat_{gt}$ et définis, pour une image $I$, par le vecteur des moyennes et des écarts-types, $Stat_{gt}(I) = \{\mu_0, \sigma_0, \mu_{\frac{\pi}{4}}, \sigma_{\frac{\pi}{4}}, \mu_{\frac{\pi}{2}}, \sigma_{\frac{\pi}{2}}, \mu_{\frac{3\pi}{4}}, \sigma_{\frac{3\pi}{4}}\}$ ;

1. les huit classes de mouvements de caméra introduites sont : statique, zoom ou travelling avant, zoom ou travelling arrière, panoramique droit, panoramique gauche, panoramique haut, panoramique bas, complexe.

– l'histogramme des orientations, noté $H_{gt}$. En chaque point, l'orientation retenue est celle du filtre ayant répondu le plus fortement. Un histogramme est ensuite construit sur les quatre orientations.

Comme pour les descripteurs de couleur, les descripteurs de texture sont calculés sur les images-clefs, moyennés et normalisés.

#### *5.2.2.5 Descripteurs de contours*

Nous avons utilisé le détecteur de contours de Canny [Canny 86], dont une version simple a été développée au GRAMM de l'INA. Le descripteur de contour est la carte des éléments de contour dans l'image, et est noté *Edge*. Calculé sur les images-clefs, il est défini au niveau du plan comme l'union des différentes cartes de contour obtenues sur celles-ci.

### 5.2.3 Calcul des similarités

Afin de calculer les similarités entre plans, $d(S_i, S_j)$, nous avons retenu, pour chacun des descripteurs, des mesures de similarité classiques.

Pour la signature $H_{col}$, nous avons retenu, comme mesure de similarité, l'intersection d'histogrammes équivalente à la norme $L_1$ [Swain 91], et le test du $\chi_2$ tel qu'il est défini dans [Brunelli 99]. Pour la signature $H_{reg}$, nous avons gardé la norme $L_1$ et adopté la mesure dite *Earth Mover's distance* (EMD) [Rubner 98]. Celle-ci nous a servi aussi pour la signature *DomCol*. Pour comparer les images binarisées des contours, nous avons repris la méthode proposée dans [Zabih 96], fondée sur le calcul des apparitions et disparitions des éléments de contour entre deux images.

Pour les autres signatures $H_{cam}$, $Act_{Sh}$, *RythmEdit*, $H_{gt}$, $Stat_{gt}$, *CCV*, *AutoCrlg*, nous avons uniquement considéré la distance $L_1$ [Pass 96,Huang 97].

### 5.2.4 Utilisation conjointe de plusieurs primitives

L'utilisation conjointe de signatures de natures diverses renvoie à la difficulté de fusionner des sources d'information non homogènes. Dans le cadre de nos travaux, nous avons opté pour des solutions simples et directes.

Afin d'*homogénéiser* les valeurs issues du calcul des différentes mesures de similarités, nous avons normalisé dans $[0, 100]$ les valeurs obtenues : $d_{norm}(S_i, S_j) = 100 * \frac{d(S_i,S_j)-\min_d}{\max_d - \min_d}$, où $\min_d$ et $\max_d$ sont respectivement les valeurs minimales et maximales obtenues pour la distance $d$ entre deux plans.

Deux modes de fusion de ces distances ont été considérés :

– $d_{glob}(S_i, S_j) = f(\{d_k(S_i, S_j),\ k \in \mathcal{S}\})$, où $\mathcal{S}$ est l'ensemble des signatures considérées et $f$ représente les fonctions *min* ou *max* ;

– $d_{glob}(S_i, S_j) = \sum_{k\in\mathcal{S}} \alpha_k d_k(S_i, S_j)$, où les $\alpha_k$ sont des cœfficients de pondération tels que $\sum_{k\in\mathcal{S}} \alpha_k = 1$.

$d_k$ est une mesure de similarité normalisée pour le descripteur $k$ et $d_{glob}$ est la mesure globale de similarité sur l'ensemble des signatures considérées.

La premier mode présente l'avantage de ne pas introduire de paramètres supplémentaires à gérer. Il implique que deux plans sont similaires dès qu'une des caractéristiques considérées présente des similarités (choix du *min*), ou que la similarité de deux plans est définie par la primitive la plus discriminante (choix du *max*). Le second mode introduit un ensemble de paramètres supplémentaires, mais permet de règler plus finement l'importance relative des différentes sources d'information.

## 5.3 Mise en œuvre de la macro-segmentation

Les mesures de similarité entre plans étant construites, l'algorithme de macro-segmentation peut être mis en œuvre. Les trois modules qui le constituent sont présentés ci-dessous.

### 5.3.1 Construction d'une hiérarchie ascendante binaire contrainte temporellement

La construction d'une hiérarchie binaire ascendante est une technique classique de classification, pour laquelle il est nécessaire de définir une mesure de similarité $d$ entre éléments et d'un indice de similarité $\delta$ entre classes [Zupan 82]. Le lecteur trouvera, sur ce sujet, un complément d'informations dans l'annexe B où sont présentés le pseudo-code de l'algorithme, le traitement complet d'un exemple simple et quelques commentaires proposant une vision plus intuitive du procédé.

#### *5.3.1.1 Définition de la contrainte temporelle*

Nous inspirant des travaux de [Yeung 96], nous avons substitué à la mesure de similarité[2] $d$, une mesure de similarité contrainte temporellement $\tilde{d}$ définie par l'équation suivante :

$$\tilde{d}(S_i, S_j) = \begin{cases} [100 \times (1 - W(d_t(S_i, S_j)))] + [d(S_i, S_j) \times W(d_t(S_i, S_j))] & \text{si } d_t(S_i, S_j) \leq \Delta T \\ \infty & \text{sinon} \end{cases} \tag{5.1}$$

où :

- $S_i$ et $S_j$ sont deux plans quelconques ;
- $\Delta T$ est l'intervalle temporel maximal pour la prise en compte de l'interaction entre deux plans ;
- $d_t$ est la distance temporelle entre deux plans, définie par $d_t(S_i, S_j) = \max(Ff_{S_j} - Lf_{S_i}, Ff_{S_i} - Lf_{S_j})$, avec $Ff_S$ et $Lf_S$ notant les première et dernière images du plan $S$ ;
- $W(d_t(S_i, S_j))$ est une fonction de pondération temporelle reflétant l'éloignement entre les plans $S_i$ et $S_j$.

La fonction $W$ est définie sur $[0, \Delta T]$ et trois versions en ont été explorées :

- fonction constante ; $W(l) = 1 \ \forall l \in [0, \Delta T]$. Dans ce cas, la relation 5.1 est strictement celle donnée dans [Yeung 96] ;
- fonction linéaire ; $W(l) = 1 - \frac{l}{\Delta T} \ \forall l \in [0, \Delta T]$. On retrouve le type de contrainte temporelle utilisée par [Rui 98] ;
- fonction quadratique : $W(l) = 1 - (\frac{l}{\Delta T})^2 \ \forall l \in [0, \Delta T]$.

2. par exemple une des mesures normalisées décrites à la sous-section 5.2.3, ou une des mesures globales proposées à la sous-section 5.2.4.

#### *5.3.1.2 Construction de la hiérarchie*

Á l'initialisation, chaque plan forme une classe, et l'indice de similarité contraint par le temps $\tilde{\delta}$ entre deux classes $C_i$ et $C_j$ se confond avec la mesure de similarité contrainte temporellement $\tilde{d}$ entre les plans $S_i$ et $S_j$ qui les constituent, soit $\tilde{\delta}(C_i, C_j) = \tilde{d}(S_i, S_j)$.

La matrice $\tilde{\mathcal{D}} = [\tilde{\delta}(i,j)]$ est alors construite à partir des $N$ classes initiales [Jain 88]. Elle est symétrique et de taille $N \times N$. Une hiérarchie binaire ascendante et contrainte temporellement peut alors être construite de manière itérative, en regroupant à chaque étape les deux classes les plus proches au sens de $\tilde{\delta}$. La matrice $\tilde{\mathcal{D}}$ est remise à jour après chaque regroupement de classes afin de tenir compte de la nouvelle classe créée. La procédure se poursuit par itérations successives jusqu'à ce que tous les cœfficients de $\tilde{\mathcal{D}}$ soient de valeur infinie.

#### *5.3.1.3 Mise à jour de la matrice $\tilde{\mathcal{D}}$*

Pour remettre à jour la matrice $\tilde{\mathcal{D}}$, nous utilisons la formule de Lance et William. Elle est donnée par :

$$\tilde{\delta}(A \cup B, C) = a_1\tilde{\delta}(A,C) + a_2\tilde{\delta}(B,C) + a_3\tilde{\delta}(A,B) + a_4|\tilde{\delta}(A,C) - \tilde{\delta}(B,C)| \qquad (5.2)$$

Il existe plusieurs procédures correspondant à diverses valeurs de $a_1$, $a_2$, $a_3$ et $a_4$ [Zupan 82]. Nous avons retenu les deux suivantes :

- la méthode du lien maximal (*complete link method*), utilisée dans [Yeung 96]. L'indice de dissimilarité entre classes est alors défini par :

  $$\tilde{\delta}(C_p, C_q) = \max_{(S_i,S_j)\in C_p\times C_q} \{\tilde{d}(S_i, S_j)\}$$

  et les cœfficients de la relation (5.2) prennent les valeurs $a_1 = a_2 = a_4 = \frac{1}{2}$ et $a_3 = 0$ ;

- le méthode de Ward. L'indice de dissimilarité entre classes est alors défini par :

  $$\tilde{\delta}(C_p, C_q) = \frac{n_{C_p}.n_{C_q}}{n_{C_p} + n_{C_q}} \tilde{d}(G_{C_p}, G_{C_q}),$$

  où $G_{C_i}$ est le centre de gravité de la classe $C_i$ et où $n_{C_i}$ peut représenter, au choix, $Cardinal(C_i)$ ou $Duration(C_i) = \sum_{S_n \in C_i} N_f(S_n)$. Les cœfficients de la relation (5.2) prennent alors les valeurs $a_1 = \frac{n_A + n_C}{n_{A\cup B} + n_C}$, $a_2 = \frac{n_B + n_C}{n_{A\cup B} + n_C}$, $a_3 = 0$, $a_4 = \frac{n_C}{n_{A\cup B} + n_C}$.

La méthode du lien maximal introduit une contrainte forte sur les classes, puisqu'elle assure qu'à chaque niveau de la hiérarchie il existe une valeur $\tilde{\delta}_h$ telle que la dissimilarité entre tout couple de classes distinctes est supérieure à $\tilde{\delta}_h$, et telle que la dissimilarité entre tout couple de plans regroupés dans une même classe est inférieure à cette même valeur. Nous obtenons alors des classes homogènes et bien séparées. La méthode de Ward, de nature barycentrique, permet d'obtenir une représentation plus fidèle du contenu des classes lors de leurs regroupements successifs.

Ainsi, nous obtenons, par la construction d'une hiérarchie ascendante binaire contrainte par le temps, la description de la proximité physique et temporelle des plans extraits.

### 5.3.2 Calcul d'une mesure de la cohérence entre plans successifs

La hiérarchie construite peut être décrite grâce à une autre matrice de dissimilarité, dite *matrice cophénétique* et notée $\tilde{\mathcal{D}}_c$ [Jain 88]. Celle-ci est définie par $\tilde{\mathcal{D}}_c = [\tilde{d}_c(S_i, S_j)]$, où $\tilde{d}_c$ est la *distance cophénétique* donnée par :

$$\tilde{d}_c(S_i, S_j) = \max_{p \neq q / (S_i, S_j) \in C_p \times C_q} \{\tilde{\delta}(C_p, C_q)\} \tag{5.3}$$

La distance cophénétique entre deux plans $S_i$ et $S_j$ est tout simplement la distance intra de la classe où ils sont regroupés pour la première fois (voir figure 5.1 où les plans sont notés *shot_cls_i*). Ainsi, à titre d'exemple, d'après la hiérarchie présentée sur cette figure, nous avons $\tilde{d}_c(S_{192}, S_{199}) = 22.34$, $\tilde{d}_c(S_{195}, S_{199}) = 31.33$ et $\tilde{d}_c(S_{186}, S_{195}) = 47.52$.

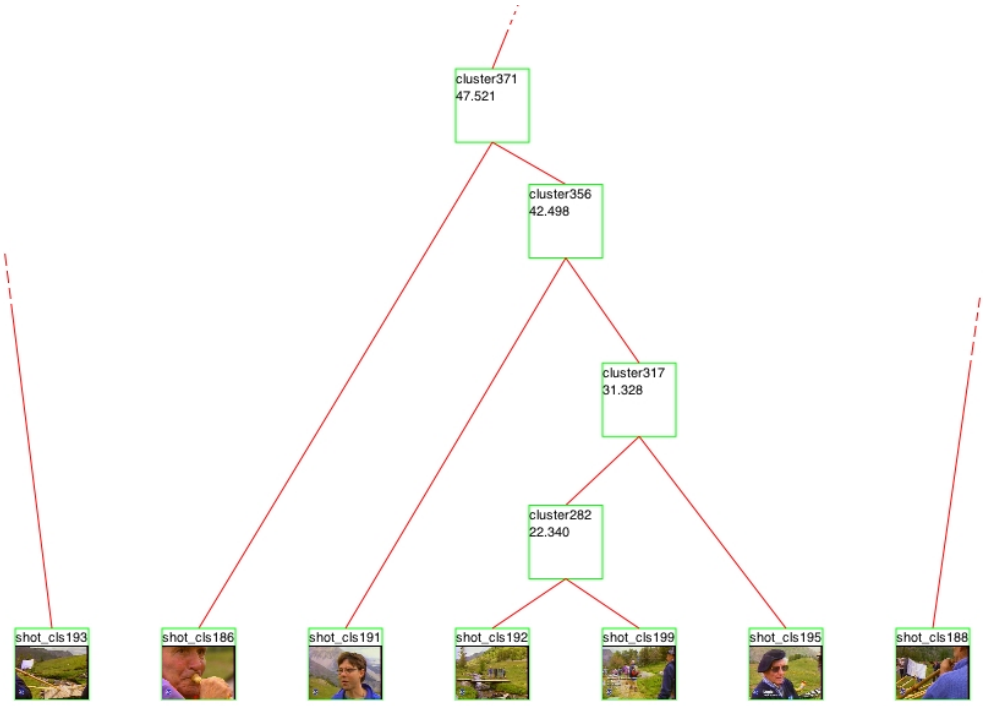



FIG. 5.1: *Exemple d'un extrait d'une hiérarchie ascendante binaire contrainte par le temps (document* aim1mb05 - *reportage* "Les sonneurs de cor de Briançon"*)*

En supposant que les indices des plans reflètent leur ordre temporel, il est alors possible de calculer, entre deux plans consécutifs, une mesure de cohérence, notée $d_m$ et définie par :

$$\tilde{d}_m(S_i, S_{i+1}) = Q_\alpha(\{\tilde{\mathcal{D}}_c(k, l) / l \leq i < k\}), \tag{5.4}$$

où $Q_\alpha$ désigne le quantile d'ordre $\alpha$. Dans l'expression (5.4), on considère l'ensemble des distances cophénétiques $\{\tilde{\mathcal{D}}_c(k, l)\}$ de part et d'autre de la frontière entre les plans $S_i$ et $S_{i+1}$ comme une population statistique dont $d_m$ est le quantile (ou fractile) d'ordre $\alpha$. Si $\alpha = 0$, $Q_\alpha$ est la fonction *min* et nous retrouvons la méthode mise en œuvre dans [Yeung 96], si $\alpha = 1$ $Q_\alpha$ est la fonction *max*, et si $\alpha = 0.5$, la valeur de $d_m$ sera la médiane.

On obtient ainsi un critère, calculé à chaque frontière de plans, permettant de représenter la cohérence physique et temporelle entre plans.

### 5.3.3 Construction de la macro-segmentation

Un simple seuillage absolu de la mesure $\tilde{d}_m$ donnée par la relation (5.4) permet de déterminer les frontières entre plans qui seront considérées comme des ruptures entre macro-segments. Les transitions graduelles entre deux plans d'une même séquence seront intégrées à celle-ci, les autres seront conservées comme transition entre macro-segments. Enfin, si plusieurs transitions de plans successives sont détectées entre des macro-segments, elles sont arbitrairement regroupées en une unique transition au niveau de ceux-ci.

## 5.4 Description des outils d'analyse et de visualisation utilisés

Les segmentations résultantes sont exportées au format XML, ce qui permet leur visualisation dans l'outil *Content Provider Application* (COPA) développé au sein du GRAMM de l'INA [Agir 01]. Une copie d'écran de cet outil est proposée sur la figure 6.7.

Par ailleurs, dans le cadre d'une étude quantitative, nous avons utilisé, afin de comparer les macro-segmentations obtenues à une macro-segmentation de référence, les sept indicateurs proposés par [Ruiloba 99] ainsi que le critère utilisé dans [Hanjalic 99]. La constitution d'une macro-segmentation manuelle de référence sera présentée à la sous-section 6.2.1 et les indicateurs statistiques seront détaillés à la sous-section 6.2.2.

# Chapitre 6

# Expérimentations

## 6.1 Objectifs des expérimentations

L'évaluation des outils d'indexation automatique, et notamment des outils de macro-segmentation, est un point délicat pour lequel il n'existe pas de réponse satisfaisante (voir sous-section 4.2.6).

Dans le cadre de notre étude, sans prétendre apporter une réponse complète aux difficultés soulevées, nous avons souhaité aborder cette question au travers des deux points suivants : comment définir une macro-segmentation de référence et comment juger de l'adéquation de nos résultats avec l'annotation de référence ?

Le premier point renvoie aux interrogations sur l'objectif qui est donné aux outils d'indexation automatique et s'inscrit dans la réflexion à mener sur les usages (voir sous-section 1.2.3). Le second renvoie à la fois aux problèmes de subjectivité de la lecture d'un document et à la difficulté de trouver des critères d'évaluation objectifs et appropriés (voir sous-section 2.4).

À défaut de pouvoir proposer une réponse exhaustive à ces problèmes sensibles, nous proposons, dans les sections suivantes, un premier cadre méthodologique d'évaluation de nos outils.

## 6.2 Méthodologie d'évaluation qualitative et quantitative

### 6.2.1 Constitution d'une annotation manuelle de référence

Dans un premier temps, nous avons cherché à définir un corpus représentatif de la variété des documents audiovisuels, et notamment de ceux présents à l'INA, sur lequel nous pourrions évaluer notre outil de macro-segmentation. Ensuite, une segmentation manuelle a été réalisée conjointement par des chercheurs et des professionnels de la documentation au sein du GRAMM de l'INA.

#### *6.2.1.1 Un corpus expérimental*

La construction d'une indexation de référence passe par l'élaboration d'un corpus expérimental comportant des documents de nature et de structure différentes. Nous avons donc constitué un corpus qui, à défaut d'être volumineux (environ trois heures), se veut diversifié. Nous avons souhaité nous fonder à la fois sur la constitution du fonds d'archives présent à l'INA (sans toutefois prétendre à l'exhaustivité), sur les grilles de programmes, et sur les travaux déjà effectués précédemment dans le cadre de constitutions de corpus à des fins de recherche (AIM) ou dans le cadre de projets européens (DiVAN) ou RNRT (AGIR). Le lecteur pourra se référer à l'annexe A.1 pour la description

détaillée des documents audiovisuels cités. Le corpus finalement retenu est constitué de :

- un journal télévisé (*aim1mb05*) ; ce genre télévisuel est relativement stable depuis sa création (à quelques exceptions près), caractérisé par l'alternance de situations de *plateau* et de *reportages* ;
- un magazine sportif (*munich2*) sélectionné pour la diversité des épreuves sportives entrecoupées de séquences de commentaires sur site et en plateau, de publicité, de bandes annonces promotionnelles du diffuseur, etc. ;
- une fiction (*aim1mb08*) permettant d'étudier les différents niveaux de structuration, plus ou moins implicites, à l'œuvre dans une fiction ;
- une émission de variétés (*topa_gainsbourg*) où les éléments peuvent varier fortement autour d'un canevas narratif stable (alternance d'interviews et d'interprétations d'artistes).

Notons, dès à présent, que, si le journal télévisé est un genre relativement stable et si les retransmissions sportives présentent différents types d'épreuves, il aurait fallu dans l'absolu se donner un corpus plus large comprenant davantage de documents afin de ne pas introduire de biais lié à l'éventuelle spécificité d'un document. Ainsi, nous devons reconnaître que la fiction choisie ne saurait évidemment représenter toutes les fictions, et que la forme et le rythme des émissions de variété ont quelque peu évolué depuis la diffusion des *Top à* en 1974. Toutefois, pour des considérations de droits et de temps, nous avons été amenés à formuler cette première proposition de corpus pour nos expérimentations.

#### *6.2.1.2 Définition d'un guide d'indexation*

Le visionnage des documents sélectionnés a confirmé toute la difficulté qu'il y avait à définir, *a priori* et en toute généralité, la notion de macro-segment. En effet, cette démarche supposait que nous étions capables de construire des unités sémantiques standardisées quelle que soit la nature du document, une sorte de grille de lecture commune, ce qui nous a rapidement paru inaccessible. La notion de décor, par exemple, semble structurer la construction d'un journal télévisé au travers de la récurrence des séquences de plateau, mais cette notion devient plus lâche pour d'autres types de documents. Les épreuves d'athlétisme ont lieu, par exemple, dans différents lieux d'un décor unique (le stade de Munich, en l'occurence).

Par conséquent, nous avons essayé de définir, pour chacun des documents, des niveaux de macro-segmentation spécifiques. Nous avons fait, de plus, l'hypothèse de niveaux de macro-segmentation emboîtés, et tenté d'obtenir une structuration hiérarchique des documents (voir figure 4.1). Ces hypothèses, énoncées après un premier visionnage des documents, ont ensuite été confrontées à une pratique d'indexation manuelle sur l'ensemble du corpus.

Nous avons effectué la structuration de chacun des documents du corpus expérimental en fonction d'hypothèses préétablies fondées sur des usages. L'expérience de mise en œuvre d'une indexation référentielle fut double: elle nous a permis de valider l'existence de niveaux de macro-segmentation pertinents d'une part, et la réalisation d'un modèle d'évaluation pour l'indexation automatique d'autre part.

#### *6.2.1.3 Une proposition d'annotation manuelle*

Au cours de la réalisation de la structuration manuelle des documents, nous avons pu être amenés à modifier nos schémas de structuration *a priori*. Nous avons notamment remis en cause

une de nos hypothèses concernant l'existence d'une structuration des documents par emboîtements successifs des différents niveaux de segmentation temporelle.

En effet, dans la pratique, nous n'avons pas toujours pu obtenir des segmentations emboîtées. Citons l'exemple du document *topa_gainsbourg* où le passage du générique à la première chorégraphie se fait sans changement de plans alors qu'il s'agit de deux séquences distinctes. En outre, lorsqu'il a été possible de suggérer différents niveaux de similarité fondés sur différents critères (par exemple, dans la fiction *aim1mb08*, les segmentations fondées respectivement sur le déroulement des actions, le changement de lieux, et la succession de personnages), nous n'avons pas obtenu des segmentations hiérarchisées et emboîtées correspondant à différents niveaux de granularité, mais des segmentations se chevauchant liées à des grilles de lecture parallèles.

Par ailleurs, suivant en cela le point de vue du documentaliste, nous avons extrait un certain nombre de séquences correspondant à des niveaux d'intérêt précis. Ainsi, les interviews forment, au sein de macro-segments de type plateau ou reportage, des séquences qu'il est intéressant d'identifier en tant que telles. Évidemment, les niveaux de ce type sont lacunaires par construction.

Ainsi, la création d'une annotation manuelle nous a permis de remettre partiellement en cause notre approche, et de proposer une description manuelle hybride entre les approches par *structuration* et par *stratification* définies dans [Prie 98, Sec. 4.2]. Notons que les niveaux d'annotation lacunaires ne sauraient, par construction, être retrouvés par notre algorithme.

Le détail des annotations obtenues est donné dans l'annexe A.2, nous n'évoquerons ici que les éléments nécessaires à l'évaluation de notre outil.

**Journal télévisé** La macro-segmentation a été effectuée sans difficulté majeure conformément aux prescriptions établies. Les niveaux retenus pour l'évaluation sont {*habillage graphique*, *studio* et *reportage*}, reprenant l'articulation classique de la structuration des journaux télévisés et noté *aim1mb05n1*, ainsi qu'un second niveau de segmentation tenant compte, en plus, des principaux changements de décors dans les reportages, noté *aim1mb05n2*.

**Émission de variété** L'alternance des prestations des artistes a servi de guide pour la constitution d'un niveau de segmentation principal, organisé selon les critères {*générique*, *interprétation*, *interview*, *archive*}, noté *topa_gainsbourg*.

**Magazine sportif** Comme pour les journaux télévisés, nous avons mis en évidence la structure des émissions *Sport Dimanche* étudiées. Ainsi, le niveau de segmentation retenu pour l'évaluation est le découpage {*lancement sujet*, *brèves*, *sommaire*, *épreuve*, *commentaire sur site*, *interview*, *divers* (regroupant les vues générales récurrentes du site et les tableaux de résultats)}, noté *munich2*.

**Fiction** C'est sans conteste le type de document qui a posé le plus de problèmes en matière de repérage d'unités sémantiques. Comme expliqué en annexe, nous avons retenu pour l'évaluation un découpage en actions, noté *aim1mb08*.

Précisons une fois de plus que cette annotation manuelle ne prétend pas donner une vérité absolue d'une macro-segmentation de référence, mais est une proposition fondée sur des considérations d'usages tels que nous les avons perçus, devant permettre une démarche d'évaluation pour notre outil. Une fois élaborés le corpus expérimental et l'annotation manuelle, notre objectif est d'évaluer la pertinence des points suivants :

- le traitement automatique proposé par notre outil de macro-segmentation a-t-il une validité pratique ? Peut-il, en particulier, intégrer une richesse informationnelle au travers de la prise en

compte de plusieurs critères physiques et être capable d'appréhender des informations censées être fortement corrélées à des notions de décor (primitives de couleur), d'action (primitives de mouvement) ou de montage (rythme d'édition, calcul d'un critère de cohérence entre plans) ?

- le paramétrage de l'algorithme est-il aisé, intuitif? L'optimisation est-elle possible, de manière générale ou locale en fonction des genres des documents? Peut-on introduire de l'information *a priori* par ce biais et ainsi piloter notre outil en fonction du contexte?
- obtenons-nous une évaluation différenciée de notre outil selon la typologie des documents étudiés? Est-il générique, ou sinon quelles sont ses capacités de généralisation?

L'appréhension de ces différents points nécessite, aussi bien au niveau quantitatif que qualitatif, la définition d'indicateurs appropriés pour l'évaluation de la macro-segmentation réalisée.

### 6.2.2 Utilisation d'indicateurs statistiques

Afin d'évaluer les performances de la méthode de macro-segmentation sur les différents documents avec divers paramétrages, la première piste est de considérer des indicateurs numériques permettant d'exprimer l'adéquation de la segmentation obtenue automatiquement à la segmentation manuelle de référence associée. Sept indicateurs ont été proposés dans le cadre des travaux du groupe AIM, afin de comparer deux segmentations en plans [Ruiloba 99]. Nous nous sommes aussi intéressés au critère utilisé dans [Hanjalic 99].

Notons $N_t$ le nombre de transitions abruptes ou graduelles entre macro-segments dans la segmentation de référence. $N_s$ est le nombre total de ruptures entre plans dans le document. Les $N_i$ fausses détections (ou fausses alarmes ou encore transitions insérées) et les $N_d$ oublis (ou transitions supprimées) étant connues, il est proposé, dans [Ruiloba 99], les sept indicateurs suivants pour comparer les segmentations :

- le taux de correction (ou rappel) : $T_{cor} = \frac{N_t - N_d}{N_t}$
- le taux de suppression (ou oubli) : $T_{del} = \frac{N_d}{N_t}$
- le taux d'insertion (ou fausses détections) : $T_{ins} = \frac{N_i}{N_t}$
- le taux d'erreur : $T_{err} = \frac{N_d + N_i}{N_t}$
- l'indice de qualité (pondéré afin de pénaliser davantage les suppressions que les insertions[1]) : $T_{wqual} = \frac{N_t - N_d - (\frac{N_i}{3})}{N_t}$
- la probabilité de correction : $T_{cpr} = 1 - \frac{1}{2}(\frac{Ni}{(N_s - N_t)} + \frac{N_d}{N_t})$
- la précision : $T_{pre} = \frac{N_t - N_d}{N_t - N_d + N_i}$

Nous avons ajouté un critère supplémentaire utilisé, dans [Hanjalic 99], pour évaluer des macro-segmentations :

- l'indice de qualité normalisé : $T_{nqual} = \frac{N_t - N_d}{N_t + N_i}$

1. Cela se justifie en effet du point de vue de l'indexation dans la mesure où une sur-segmentation est moins coûteuse à corriger qu'un oubli pour les documentalistes.

Notons que la plupart de ces indicateurs sont calculés à partir des données $N_i$, $N_d$ et $N_t$ et que certains sont redondants (ainsi $T_{err} = T_{ins} + T_{del}$ et $T_{wqual} = T_{cor} - \frac{T_{ins}}{3}$). Par conséquent, nous ne retiendrons par la suite que quatre indicateurs: $T_{cor}$ qui indique parmi les transitions véritables celles qui ont été repérées expérimentalement, $T_{pre}$ qui indique parmi les transitions repérées expérimentalement celles qui en sont vraiment, et $T_{del}$ et $T_{ins}$ donnant le nombre d'oublis et de fausses alarmes ramenés au nombre de transitions véritables. Lorsque les segmentations expérimentales et de référence sont identiques, nous avons $T_{cor} = T_{pre} = 1$ et $T_{del} = T_{ins} = 0$.

Compte tenu de l'ensemble des paramètres et options disponibles pour l'outil de macro-segmentation, il aurait fallu mener plus de 120 tests[2] sur chaque document, sans compter les trois paramètres à valeur continue $\alpha$, $N_{req}$, $\Delta T$, et les nombreuses combinaisons possibles de la mise en œuvre d'une segmentation multi-critère. Nous nous sommes limités à une quarantaine d'expérimentations par document, ce qui représente déjà plus de 200 tests.

Afin d'étudier, et éventuellement de régler, nos paramètres, nous avons opté pour une stratégie progressive, faisant varier nos paramètres l'un après l'autre, et gardant à chaque étape la meilleure valeur obtenue. À défaut d'une hypothétique étude exhaustive, il nous a semblé que cela devait nous permettre un certain nombre d'observations, même si nous sommes pleinement conscients qu'une telle heuristique ne peut nous mener qu'à des optimisations "locales" des paramètres.

Sauf mention contraire, le seuillage du critère $d_m$ est fixé par un nombre de séquences requises $N_{req}$ égal au nombre de séquences dans la segmentation de référence[3] $N_{th}$.

### 6.2.3 Étude qualitative

Nous avons visualisé sous COPA quelques-unes des segmentations obtenues avec l'aide d'une documentaliste. Nous avons pu ainsi évaluer qualitativement la cohérence de nos segmentations expérimentales et la gravité des erreurs (oublis ou fausses détections).

Enfin, des résultats complémentaires et partiels ont été obtenus afin d'évaluer la robustesse aux erreurs de découpage en plans, et d'illustrer quelques pistes de réflexions apparues lors des expérimentations.

## 6.3 Résultats obtenus

Suivant la démarche décrite plus haut, nous présentons et commentons ci-dessous quelques résultats quantitatifs généraux, les expérimentations menées sur la paramétrisation, les tests faisant intervenir les différentes primitives en mono-critère, puis quelques tentatives d'utilisation multi-critère. Les segmentations les plus intéressantes feront ensuite l'objet de commentaires qualitatifs. Enfin, avant de tirer une conclusion sur cette partie de nos travaux, nous présenterons quelques expérimentations complémentaires.

Sauf mention contraire, les résultats présentés ont été obtenus à partir de la segmentation en plans manuelle de référence.

2. trois valeurs pour $W$, trois variantes de la formule de Lance & William, un choix parmi les valeurs possibles, dénombrables ou non, pour $\alpha$, $N_{req}$, et $\Delta T$, treize couples signature/distance, trois familles de méthodes en multi-critère permettant de combiner ces treize couples, et faisant intervenir pour l'une d'elles des paramètres supplémentaires de pondération à valeur continue.

3. l'évaluation de l'algorithme en "supervisé" sera justifié ultérieurement au paragraphe 6.3.3.3. Par ailleurs, ce protocole opératoire explique que, lors des évaluations, le nombre d'oublis et de fausses détections soit sensiblement égal.

### 6.3.1 Évaluations quantitatives

Le tableau 6.1 contient les meilleurs résultats obtenus pour chacun des documents étudiés[4], et ce avec des jeux de paramètres différents (nous préciserons cela ci-après). Les résultats nous semblent tout à fait honorables dans la mesure où $T_{cor}$ varie entre 39% et 73% pour une tâche de type "sémantique", sachant que pour une segmentation en plans, réputée plus facile, ce critère est évalué entre 54% et 95% [Dailianas 95]. Toutefois, les résultats sont largement perfectibles dans la mesure où les oublis et fausses détections cumulés donnent un taux d'erreur $T_{err}$ variant entre 64% et 123%, dans le meilleur des cas (pour la détection des plans, $T_{ins}$ est annoncé entre 56% et 175% [Dailianas 95]).

| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| aim1mb05n1 | 0.55 | 0.52 | 0.45 | 0.52 |
| aim1mb05n2 | 0.59 | 0.60 | 0.41 | 0.39 |
| aim1mb08 | 0.39 | 0.39 | 0.61 | 0.61 |
| munich2 | 0.54 | 0.53 | 0.46 | 0.48 |
| topa_gainsbourg | 0.73 | 0.67 | 0.27 | 0.36 |

TAB. 6.1: *Meilleurs résultats obtenus sur l'ensemble des expérimentations*

Nous remarquons aussi une certaine disparité des résultats entre les documents, sur laquelle nous reviendrons. Comme nous l'avons précisé au paragraphe 6.2.1, il ne nous sera pas possible en toute rigueur de généraliser les résultats obtenus sur nos documents spécifiques aux genres qu'ils représentent (journaux, fictions, magazines de sport ou de variété).

#### *6.3.1.1 Influence de la fonction de pondération temporelle W*

Afin de tenir compte des critiques formulées sur la méthode de Yeung *et al.* [Yeung 98] concernant la non continuité de la contrainte temporelle (voir [Kender 98]), nous avons, à la suite d'autres auteurs (comme [Rui 98] avec son attraction temporelle), introduit des versions non constantes de la fonction $W$. Les résultats rassemblés dans le tableau 6.2 montrent qu'une telle démarche était justifiée puisque, dans trois cas sur cinq, les versions non constantes améliorent les résultats obtenus avec la méthode de Yeung *et al.*

| fonction $W$ | constante | | | | linéaire | | | | quadratique | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| aim1mb05n1 | 0.26 | 0.22 | 0.74 | 0.94 | 0.52 | 0.49 | 0.48 | 0.55 | 0.39 | 0.36 | 0.61 | 0.68 |
| aim1mb05n2 | 0.20 | 0.19 | 0.80 | 0.83 | 0.50 | 0.51 | 0.50 | 0.48 | 0.44 | 0.43 | 0.57 | 0.59 |
| aim1mb08 | 0.32 | 0.32 | 0.68 | 0.68 | 0.26 | 0.26 | 0.74 | 0.74 | 0.39 | 0.39 | 0.61 | 0.61 |
| munich2 | 0.48 | 0.47 | 0.53 | 0.54 | 0.42 | 0.42 | 0.58 | 0.59 | 0.44 | 0.43 | 0.56 | 0.58 |
| topa_gainsbourg | 0.64 | 0.58 | 0.36 | 0.46 | 0.50 | 0.46 | 0.50 | 0.59 | 0.46 | 0.42 | 0.55 | 0.64 |

TAB. 6.2: *Influence de la fonction de la contrainte temporelle W*

Néanmoins, si l'intérêt d'une version plus évoluée de la fonction $W$ est confirmé, l'usage de la version constante n'est pas à proprement parler disqualifié. La difficulté de choisir une option par défaut est sensible puisque les écarts de performance sont importants (jusqu'à 30% pour $T_{cor}$). Nous

4. Le nombre de macro-segments recherché automatiquement est celui de la macro-segmentation manuelle de référence, et est indiqué dans le tableau 6.8 par le paramètre $N_{req}$.

pouvons observer que la constance des performances sur les deux niveaux considérés du document *aim1mb05* laisse espérer la définition d'un profil de paramétrisation par genre. Notons enfin que les deux meilleures versions correspondent aux options mises en œuvre respectivement par [Yeung 98] et [Rui 98] (voir section 5.3).

#### *6.3.1.2 Influence de la méthode de mise à jour de la matrice $\tilde{\mathcal{D}}$*

Bien que disparates, les résultats donnés dans le tableau 6.3 semblent plus lisibles.

| méthode | lien maximal | | | | Ward (durée) | | | | Ward (cardinal) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| aim1mb05n1 | 0.52 | 0.49 | 0.48 | 0.55 | 0.55 | 0.50 | 0.45 | 0.55 | 0.45 | 0.42 | 0.55 | 0.61 |
| aim1mb05n2 | 0.50 | 0.51 | 0.50 | 0.48 | 0.50 | 0.49 | 0.50 | 0.52 | 0.50 | 0.51 | 0.50 | 0.48 |
| aim1mb08 | 0.39 | 0.39 | 0.61 | 0.61 | 0.26 | 0.26 | 0.74 | 0.74 | 0.39 | 0.39 | 0.61 | 0.61 |
| munich2 | 0.48 | 0.47 | 0.53 | 0.54 | 0.49 | 0.48 | 0.51 | 0.53 | 0.49 | 0.48 | 0.51 | 0.53 |
| topa_gainsbourg | 0.64 | 0.58 | 0.36 | 0.46 | 0.59 | 0.54 | 0.41 | 0.50 | 0.64 | 0.58 | 0.36 | 0.46 |

TAB. 6.3: *Influence de la formule de Lance & William retenue*

Les écarts sur $T_{cor}$ sont, en effet, moins sensibles à cet aspect (l'écart est de 6% en moyenne) et, pour toutes les expérimentations menées, l'une des méthodes de Ward égale ou améliore les résultats obtenus avec la méthode du lien maximal. Dans quatre cas sur cinq, nous pouvons choisir la méthode de Ward fondée sur la cardinalité pour un résultat optimal.

Il semble ainsi possible de suggérer la méthode de Ward fondée sur la cardinalité comme choix par défaut pour cette option.

#### *6.3.1.3 Influence de l'ordre $\alpha$ du quantile dans le calcul de $d_m$*

Le choix de l'ordre $\alpha$ du quantile nous permet, *a priori*, de mieux prendre en compte la distribution des distances cophénétiques pour décider de la présence d'une rupture. Malheureusement, le tableau 6.4 infirme cette intuition; tout autre choix que le minimum dégrade considérablement les performances.

| quantile | *minima* $\alpha = 0$ | | | | premier décile $\alpha = 0.1$ | | | | dernier décile $\alpha = 0.9$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| aim1mb05n1 | 0.55 | 0.50 | 0.45 | 0.55 | 0.55 | 0.21 | 0.45 | 2.13 | 1.00 | 0.15 | 0.00 | 5.77 |
| aim1mb05n2 | 0.50 | 0.51 | 0.50 | 0.48 | 0.57 | 0.30 | 0.44 | 1.30 | 1.00 | 0.21 | 0.00 | 3.78 |
| aim1mb08 | 0.39 | 0.39 | 0.61 | 0.61 | 0.65 | 0.08 | 0.36 | 7.00 | 1.00 | 0.05 | 0.00 | 21.45 |
| munich2 | 0.49 | 0.48 | 0.51 | 0.53 | 0.61 | 0.12 | 0.39 | 4.51 | 1.00 | 0.13 | 0.00 | 6.86 |
| topa_gainsbourg | 0.64 | 0.58 | 0.36 | 0.46 | 0.55 | 0.39 | 0.46 | 0.86 | 0.86 | 0.10 | 0.14 | 7.96 |

TAB. 6.4: *Influence de l'ordre $\alpha$ du quantile*

Si le nombre d'oublis diminue sensiblement, c'est au prix d'un nombre de fausses détections rhédibitoire, même si, du point de vue de la documentation, il est préférable d'obtenir une sur-segmentation qu'une sous-segmentation. De fait, pour $\alpha = 0.9$, on retrouve presque la segmentation en plans. Ainsi, comme le montre la figure 6.1, lorsque $\alpha = 0.9$, le critère reste contant et égal à la valeur maximale.

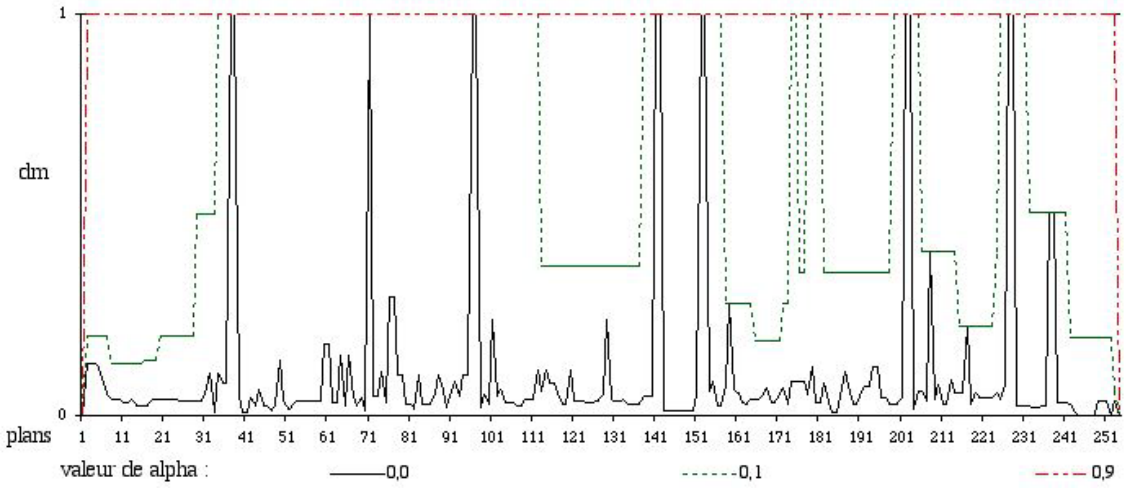



FIG. 6.1: *Influence de l'ordre $\alpha$ du quantile sur le critère $d_m$ de cohérence entre plans*

#### *6.3.1.4 Influence d'une sur-segmentation*

Considérant que les oublis étaient plus graves du point de vue de l'indexation que les fausses détections, nous avons essayé de sur-segmenter les documents de 50%. Rappelons que le critère de cohérence $d_m$ est seuillé à l'aide du nombre de macro-segments souhaités $N_{req}$ et que, sauf mention contraire, $N_{req}$ est fixé dans nos expérimentations au nombre théorique de macro-segments $N_{th}$ obtenus dans la macro-segmentation manuelle de référence. Nous espérions en sur-segmentant diminuer sensiblement le nombre d'oublis avec un coût raisonnable de fausses détections supplémentaires. Le bilan résumé dans le tableau 6.5 est peu probant.

| Segmentation | $N_{req} = N_{th}$ | | | | $N_{req} = N_{th}+50\%$ | | | |
|---|---|---|---|---|---|---|---|---|
| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| aim1mb05n1 | 0.55 | 0.50 | 0.45 | 0.55 | 0.55 | 0.34 | 0.45 | 1.07 |
| aim1mb05n2 | 0.50 | 0.51 | 0.50 | 0.48 | 0.59 | 0.40 | 0.41 | 0.89 |
| aim1mb08 | 0.39 | 0.39 | 0.61 | 0.61 | 0.45 | 0.30 | 0.55 | 1.07 |
| munich2 | 0.49 | 0.48 | 0.51 | 0.53 | 0.58 | 0.37 | 0.42 | 0.97 |
| topa_gainsbourg | 0.64 | 0.58 | 0.36 | 0.46 | 0.73 | 0.44 | 0.27 | 0.91 |

TAB. 6.5: *Influence de la sur-segmentation*

Si le taux de fausses détections $T_{ins}$ double pour l'ensemble des documents, le gain sur $T_{del}$ est faible (6% en moyenne). Une explication probable est qu'une partie des oublis ne peut être corrigée par un simple abaissement du seuillage de $d_m$. Cette hypothèse semble justifiée et confortée par les observations qualitatives présentées au paragraphe 6.3.2.1.

#### *6.3.1.5 Influence de la fenêtre temporelle $\Delta T$*

Afin d'étudier l'influence de la longueur de la fenêtre temporelle, nous l'avons dans un premier temps fixée à la valeur généralement admise[5] de 3000 images (ou 2 minutes). Ensuite, nous

5. Dans [Yeung 98], $\Delta T = 3000$ pour des documents à 30 images par seconde, ce qui correspond à une fenêtre d'une minute quarante secondes.

l'avons doublée. Puis, tenant compte du fait que deux plans physiquement identiques non liés sémantiquement ne doivent pas se situer à moins de $\Delta T$ (cf. sous-section 4.2.5), nous avons examiné les différents documents dans cette optique.

| fenêtre temporelle | $\Delta T = 1200$ | | | | $\Delta T = 3000$ | | | | $\Delta T = 6000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| aim1mb05n1 | 0.45 | 0.42 | 0.55 | 0.61 | 0.55 | 0.50 | 0.45 | 0.55 | 0.45 | 0.39 | 0.55 | 0.71 |
| aim1mb05n2 | 0.48 | 0.49 | 0.52 | 0.50 | 0.50 | 0.51 | 0.50 | 0.48 | 0.44 | 0.44 | 0.57 | 0.54 |
| aim1mb08 | 0.52 | 0.32 | 0.48 | 1.10 | 0.39 | 0.39 | 0.61 | 0.61 | 0.32 | 0.32 | 0.68 | 0.68 |
| munich2 | 0.32 | 0.32 | 0.68 | 0.70 | 0.49 | 0.48 | 0.51 | 0.53 | 0.41 | 0.38 | 0.59 | 0.68 |
| topa_gainsbourg | 0.50 | 0.37 | 0.50 | 0.86 | 0.64 | 0.58 | 0.36 | 0.46 | 0.64 | 0.58 | 0.36 | 0.46 |

TAB. 6.6: *Influence de la longueur de la fenêtre temporelle* $\Delta T$

Pour *aim1mb05*, nous avons évalué la durée des reportages séparés par des plans en studio identiques (cf. figure 6.2), ce qui nous a conduits à retenir un $\Delta T$ de 1200 images. Puis, nous avons procédé à l'étude des nouvelles brèves séparées par des "jingles" graphiques semblables (cf. figure 6.3), ce qui nous a amenés à prendre $\Delta T = 450$. Pour *topa_gainsbourg* et *aim1mb08*, nous avons considéré les longueurs des macro-segments annotés manuellement (cf. figures 6.6 et 6.4). Enfin, pour *munich2*, nous avons compté les intervalles entre deux vues générales du stade qui constituent des plans de coupe récurrents (cf. figure 6.5). Pour ces dernières études, un $\Delta T$ de 1200 images nous a aussi semblé être une valeur convenable.

| fenêtre temporelle | $\Delta T = 450$ | | | |
|---|---|---|---|---|
| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| aim1mb05n1 | 0.55 | 0.43 | 0.45 | 0.74 |
| aim1mb05n2 | 0.54 | 0.56 | 0.46 | 0.44 |

TAB. 6.7: *Influence de la longueur de la fenêtre temporelle* $\Delta T$ *(suite)*

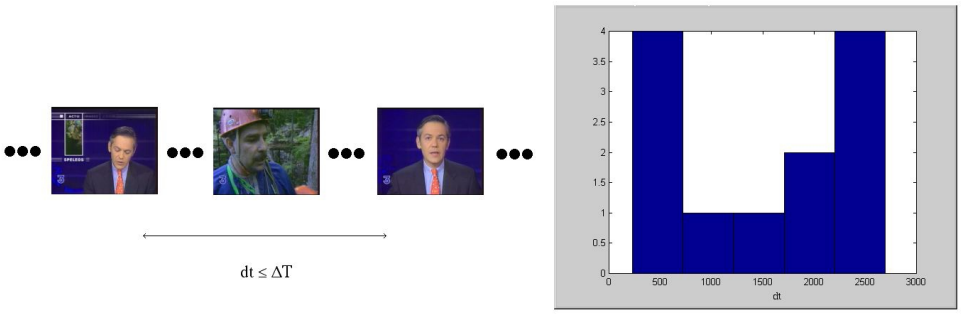



FIG. 6.2: *Répartition des durées des reportages dans le document* aim1mb05

Les valeurs des différents taux des tableaux 6.6 et 6.7 sont significatifs de la difficulté à fixer le paramètre $\Delta T$. Hormis le cas de *aim1mb05n2* où un $\Delta T$ de 450 images semble en adéquation avec le niveau de granularité recherché, pour toutes les autres segmentations nous n'avons pu améliorer les résultats obtenus à $\Delta T = 3000$. Alors que d'après les graphiques 6.2 à 6.6, pour cette valeur, des

plans similaires appartenant à des macro-segments différents risquent fortement d'être regroupés dans la hiérarchie.

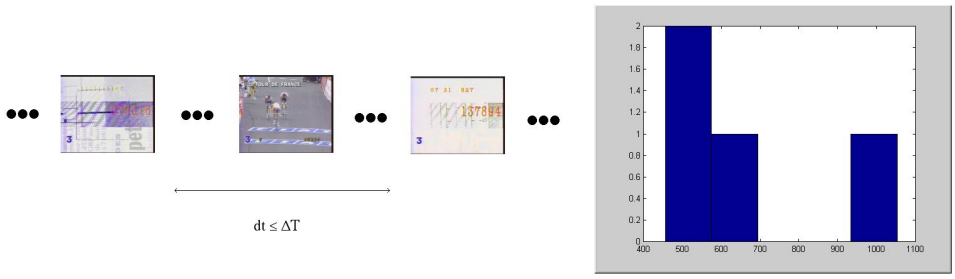



FIG. 6.3: *Répartition des durées des nouvelles brèves dans le document* aim1mb05

En effet, lorsque $\Delta T$ augmente, la similarité physique entre des plans éloignés temporellement est prise en compte, ce qui entraîne un regroupement précoce au sein de la hiérarchie de plans sans nécessairement de lien sémantique, et par suite une augmentation des oublis. Inversement, lorsque $\Delta T$ diminue, la similarité de plans sémantiquement liés et distants n'est plus prise en compte. Une conséquence en est la sur-segmentation du document, la hiérarchie se construisant sur des classes en fait moins homogènes, car plus localisées temporellement. L'étude menée (figures 6.2 à 6.6) n'a pas permis de dégager une valeur plus à même d'équilibrer ces deux comportements contradictoires. Il n'est d'ailleurs pas sûr qu'il existe une valeur capable de tenir compte de cette double contrainte. Dans l'immédiat, $\Delta T = 3000$ semble être un paramétrage par défaut raisonnable.


FIG. 6.4: *Répartition des durées des séquences dans le document* aim1mb08

Notons que, dans [Yeung 98], une étude sur le choix de $\Delta T$ fait émerger des conclusions similaires. Le $\Delta T$ idéal se situe pour les auteurs aux alentours d'une minute, sachant qu'ils proposent une macro-segmentation en deux passages. Ils procèdent dans un premier temps à une sur-segmentation, puis calculent un $\Delta T$ variable fondé sur la longueur des macro-segments obtenus, avant de raffiner leur segmentation par l'utilisation de cette nouvelle valeur de $\Delta T$. Dans notre contexte, nous n'avons pas jugé utile de mettre en œuvre une telle stratégie, dans la mesure où les auteurs évoquent la possiblité de regroupements abusifs de séquences.

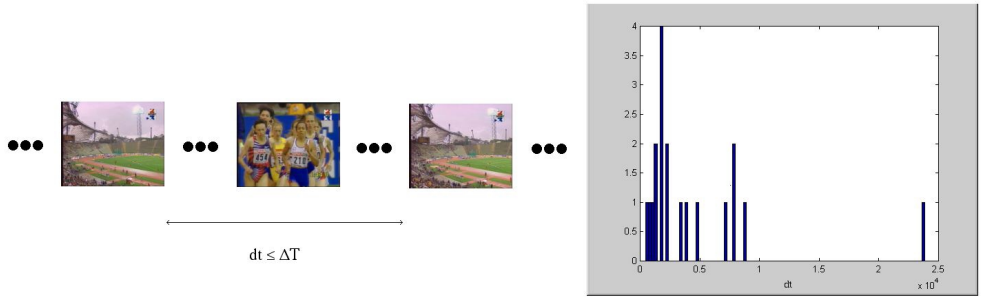



FIG. 6.5: *Répartition des durées des séquences entre deux plans fixes du stade dans le document* munich2


FIG. 6.6: *Répartition des durées des séquences dans le document* topa_gainsbourg

#### *6.3.1.6 Paramétrisation de l'algorithme en fonction du document*

Comme indiqué à la sous-section 6.2.2, les expérimentations menées nous ont permis de fixer un ensemble de paramètres "localement optimaux" pour chaque document. Cette paramétrisation sera utilisé (sauf mention contraire) dans les expériences ultérieures.

| documents | $W$ | Lance & William | $\alpha$ | $N_{req}$ | $\Delta T$ |
|---|---|---|---|---|---|
| aim1mb05n1 | linéaire | Ward (durée) | 0.0 | 34 | 3000 |
| aim1mb05n2 | linéaire | Ward (cardinal) | 0.0 | 46 | 450 |
| aim1mb08 | quadratique | Ward (cardinal) | 0.0 | 32 | 3000 |
| munich2 | constante | Ward (durée) | 0.0 | 60 | 3000 |
| topa_gainsbourg | constante | Ward (cardinal) | 0.0 | 23 | 3000 |

TAB. 6.8: *Choix des paramètres pour les segmentations mono- et multi-critères*

Si ce mode d'utilisation pourrait augurer d'une gestion des paramètres fondé sur les genres des documents, rappelons toutefois que notre corpus ne nous permet pas de valider une telle hypothèse. Le choix final pour les paramètres considérés est indiqué par document dans le tableau 6.8.

#### *6.3.1.7 Influence des données extraites (primitives et distances)*

Ayant réglé les paramètres expérimentalement sur chacun des documents, nous avons testé treize couples signatures/distances dans le cadre d'une macro-segmentation mono-critère. Les résultats

pour chacun des documents sont présentés dans les tableaux 6.9 à 6.13. Les meilleurs résultats obtenus y sont indiqués en italique.

| signature et distance | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $H_{col}$ et $L_1$ | *0.55* | *0.50* | *0.45* | *0.55* |
| $H_{col}$ et $\chi_2$ | 0.45 | 0.42 | 0.55 | 0.61 |
| $H_{reg}$ et $L_1$ | 0.42 | 0.39 | 0.58 | 0.65 |
| $H_{reg}$ et $EMD$ | 0.45 | 0.42 | 0.55 | 0.61 |
| $CCV$ et $L_1$ | *0.55* | *0.52* | *0.45* | *0.52* |
| $AutoCrlg$ et $L_1$ | 0.48 | 0.46 | 0.52 | 0.58 |
| $DomCol$ et $EMD$ | *0.55* | *0.52* | *0.45* | *0.52* |
| $RythmEdit$ et $L_1$ | 0.27 | 0.26 | 0.73 | 0.76 |
| $Act_{Sh}$ et $L_1$ | 0.29 | 0.27 | 0.71 | 0.77 |
| $H_{cam}$ et $L_1$ | 0.26 | 0.24 | 0.74 | 0.81 |
| $H_{gt}$ et $L_1$ | 0.32 | 0.30 | 0.68 | 0.74 |
| $Stat_{gt}$ et $L_1$ | 0.36 | 0.33 | 0.65 | 0.71 |
| $Edge$ et $d_{Zabih}$ | 0.36 | 0.33 | 0.65 | 0.71 |

TAB. 6.9: *Résultats pour la segmentation mono-critère de* aim1mb05n1

Le constat que nous pouvons en tirer apparaît assez déconcertant et difficile à analyser. Nous observons en effet beaucoup de disparités dans les comparaisons entre les choix de primitives en fonction des documents et des granularités de segmentation (document *aim1mb05*). En toute rigueur, il serait nécessaire de mener des expérimentations sur un corpus plus large afin de s'affranchir d'éventuelles particularités liées à certains documents, expérimentations supplémentaires que nous n'avons pu mener compte tenu de la lourdeur de la mise en œuvre nécessaire. Nous allons cependant essayer de dégager quelques points importants.

| signature et distance | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $H_{col}$ et $L_1$ | 0.54 | 0.56 | 0.46 | 0.44 |
| $H_{col}$ et $\chi_2$ | *0.57* | *0.58* | *0.44* | *0.41* |
| $H_{reg}$ et $L_1$ | 0.52 | 0.53 | 0.48 | 0.46 |
| $H_{reg}$ et $EMD$ | 0.52 | 0.53 | 0.48 | 0.46 |
| $CCV$ et $L_1$ | 0.52 | 0.53 | 0.48 | 0.46 |
| $AutoCrlg$ et $L_1$ | 0.52 | 0.53 | 0.48 | 0.46 |
| $DomCol$ et $EMD$ | 0.48 | 0.49 | 0.52 | 0.50 |
| $RythmEdit$ et $L_1$ | 0.39 | 0.40 | 0.61 | 0.59 |
| $Act_{Sh}$ et $L_1$ | 0.39 | 0.40 | 0.61 | 0.59 |
| $H_{cam}$ et $L_1$ | 0.44 | 0.44 | 0.57 | 0.54 |
| $H_{gt}$ et $L_1$ | 0.37 | 0.38 | 0.63 | 0.61 |
| $Stat_{gt}$ et $L_1$ | 0.39 | 0.40 | 0.61 | 0.59 |
| $Edge$ et $d_{Zabih}$ | 0.35 | 0.36 | 0.65 | 0.63 |

TAB. 6.10: *Résultats pour la segmentation mono-critère de* aim1mb05n2

Lorsqu'on considère les cinq groupes de primitives utilisées (couleur, mouvement, édition, texture, contour), les primitives de couleur donnent, à une exception près, des résultats bien meilleurs que n'importe quelle autre primitive. L'écart moyen entre les performances atteintes pour la primitive de couleur la moins efficace et celles associées à la meilleure des primitives restantes est de

3% en moyenne et peut s'élever jusqu'à 10%.

Lorsqu'on essaie de globaliser [6] les résultats sur l'ensemble des expérimentations, les meilleures performances sont obtenues, en dehors des primitives de couleur, avec les primitives de texture, de contour, puis d'édition, et enfin de mouvement. Si l'on considère les motivations qui ont guidé le choix des primitives extraites (voir section 5.1 et sous-section 5.2.2), les résultats, notamment sur les primitives d'édition et de mouvement, pourraient apparaître comme décevants.

| signature et distance | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $H_{col}$ et $L_1$ | *0.39* | *0.39* | *0.61* | *0.61* |
| $H_{col}$ et $\chi_2$ | 0.26 | 0.26 | 0.74 | 0.74 |
| $H_{reg}$ et $L_1$ | *0.39* | *0.39* | *0.61* | *0.61* |
| $H_{reg}$ et $EMD$ | *0.39* | *0.39* | *0.61* | *0.61* |
| $CCV$ et $L_1$ | 0.36 | 0.36 | 0.65 | 0.65 |
| $AutoCrlg$ et $L_1$ | 0.36 | 0.36 | 0.65 | 0.65 |
| $DomCol$ et $EMD$ | *0.39* | *0.39* | *0.61* | *0.61* |
| $RythmEdit$ et $L_1$ | 0.10 | 0.10 | 0.90 | 0.90 |
| $Act_{Sh}$ et $L_1$ | 0.19 | 0.19 | 0.81 | 0.81 |
| $H_{cam}$ et $L_1$ | 0.13 | 0.13 | 0.87 | 0.87 |
| $H_{gt}$ et $L_1$ | 0.36 | 0.36 | 0.65 | 0.65 |
| $Stat_{gt}$ et $L_1$ | 0.16 | 0.16 | 0.84 | 0.84 |
| $Edge$ et $d_{Zabih}$ | 0.23 | 0.23 | 0.77 | 0.77 |

TAB. 6.11: *Résultats pour la segmentation mono-critère de* aim1mb08

Toutefois, nous proposons quelques pistes, explications probables de cet état de fait.

1. L'optimisation des paramètres des signatures extraites. L'extraction des primitives de couleur a fait l'objet au début de nos travaux d'une attention particulière dans le cadre d'expérimentations préliminaires (décrites à la sous-section 3.1.3). Ces travaux nous ont permis de ne garder que les signatures les plus efficaces parmi de nombreuses autres, et surtout d'en optimiser partiellement les paramètres. Pour les autres primitives, nous avons intégré, adapté des outils ou partiellement développé des méthodes dont nous n'avons pas évalué la paramétrisation, reprenant en général les valeurs indiquées par l'auteur. Nous n'avons donc pas la certitude que l'emploi de ces primitives soit le plus adapté.

2. La question des faux positifs. Les travaux portant sur l'extraction des primitives de couleur a, à de nombreuses reprises, mis en évidence que si deux images identiques ont des signatures similaires, *a contrario* des images dissimilaires pouvaient tout à fait avoir des signatures identiques. Ce problème a été pris en compte par la proposition de signatures de couleur plus riches en information. Les histogrammes par régions $H_{reg}$, l'auto-corrélogramme $AutoCrlg$ ou les vecteurs de couleurs cohérentes $CCV$, en font notamment partie. Le problème se pose aussi et de manière cruciale pour les autres primitives, d'autant que nous avons mis en œuvre des primitives souvent simples (histogramme à quatre ou cinq bins ou moments statistiques des premier et second ordres pour les mouvements et textures), et par conséquent particulièrement sujettes au phénomène des faux positifs. Nous aurons l'occasion d'illustrer cette hypothèse par quelques exemples à la sous-section 6.3.2.1.

6. globalisation dont la validité se trouve limitée par les disparités précédemment évoquées.

3. Réduction ou extraction de l'information. Il nous semble que nous pourrions séparer nos primitives en deux classes. Les signatures qui réduisent simplement l'information disponible, et celles qui transforment l'information présente dans le flux. Dans la première catégorie, on trouve la plupart des primitives de couleur que nous utilisons. Par exemple, un histogramme de couleurs n'est jamais inexact, au pire l'information peut y être trop réduite pour être utilisable si le nombre de bins n'est pas suffisant. Dans la seconde catégorie, on trouve les primitives de mouvement, celles de contour et dans une moindre mesure celles de texture. La donnée du mouvement de la caméra, du support du mouvement dominant ou de la présence d'un contour est elle-même le résultat, plus ou moins exact, d'une analyse automatique du flux audiovisuel. Ce sont sur ces données, incluant une part d'imprécision, que nous calculons nos signatures (histogramme, moments statistiques, etc.). Notons que cette dernière catégorie est, par conséquent, d'autant plus sensible à l'adéquation du paramétrage de l'extraction des signatures.

| signature et distance | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $H_{col}$ et $L_1$ | 0.49 | 0.48 | 0.51 | 0.53 |
| $H_{col}$ et $\chi_2$ | 0.48 | 0.47 | 0.53 | 0.54 |
| $H_{reg}$ et $L_1$ | 0.53 | 0.52 | 0.48 | 0.49 |
| $H_{reg}$ et $EMD$ | 0.49 | 0.48 | 0.51 | 0.53 |
| $CCV$ et $L_1$ | 0.49 | 0.48 | 0.51 | 0.53 |
| $AutoCrlg$ et $L_1$ | *0.54* | *0.53* | *0.46* | *0.49* |
| $DomCol$ et $EMD$ | 0.48 | 0.47 | 0.53 | 0.54 |
| $RythmEdit$ et $L_1$ | 0.27 | 0.26 | 0.73 | 0.76 |
| $Act_{Sh}$ et $L_1$ | 0.25 | 0.24 | 0.75 | 0.80 |
| $H_{cam}$ et $L_1$ | 0.15 | 0.14 | 0.85 | 0.92 |
| $H_{gt}$ et $L_1$ | 0.44 | 0.43 | 0.56 | 0.58 |
| $Stat_{gt}$ et $L_1$ | 0.37 | 0.35 | 0.63 | 0.70 |
| $Edge$ et $d_{Zabih}$ | 0.36 | 0.30 | 0.64 | 0.83 |

TAB. 6.12: *Résultats pour la segmentation mono-critère de* munich2

Enfin, une étude succincte des différentes familles de signatures utilisées semble suggérer que $Stat_{gt}$ est préférable à $H_{gt}$, que $Act_{Sh}$ est préférable à $H_{cam}$, et que les primitives de couleur se valent globalement. Les résultats sur les primitives de couleur sont à la fois trop proches (moins de 10% d'écart au pire sur un même test) et trop disparates (d'un test sur l'autre) pour pouvoir conclure sérieusement sans expérimentations supplémentaires.

#### *6.3.1.8 Apport d'un traitement multi-critère*

Pour la macro-segmentation multi-critère, nous avons mené une série d'expérimentations afin d'évaluer les différentes stratégies proposées (voir sous-section 5.2.4). À nouveau, notre démarche n'a pas été exhaustive compte tenu du nombre de variantes possibles. Lorsqu'une expérimentation multi-critère présente une amélioration au regard des tests mono-critère, nous l'avons indiquée en italique dans les tableaux 6.14 à 6.18.

Pour chaque document, nous avons utilisé d'une part une primitive de couleur et une primitive de mouvement et d'autre part une primitive de couleur et la primitive d'édition. Nous avons cherché à garder une démarche cohérente par rapport aux objectifs annoncés à la sous-section 5.1, c'est-à-dire appréhender principalement des notions de décor, d'activité et de montage. Les primitives de

| signature et distance | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $H_{col}$ et $L_1$ | 0.64 | 0.58 | 0.36 | 0.46 |
| $H_{col}$ et $\chi_2$ | 0.64 | 0.58 | 0.36 | 0.46 |
| $H_{reg}$ et $L_1$ | *0.68* | *0.60* | *0.32* | *0.46* |
| $H_{reg}$ et $EMD$ | 0.59 | 0.54 | 0.41 | 0.50 |
| $CCV$ et $L_1$ | 0.59 | 0.54 | 0.41 | 0.50 |
| $AutoCrlg$ et $L_1$ | 0.59 | 0.54 | 0.41 | 0.50 |
| $DomCol$ et $EMD$ | 0.55 | 0.50 | 0.46 | 0.55 |
| $RythmEdit$ et $L_1$ | 0.46 | 0.42 | 0.55 | 0.64 |
| $Act_{Sh}$ et $L_1$ | 0.27 | 0.25 | 0.73 | 0.82 |
| $H_{cam}$ et $L_1$ | 0.41 | 0.38 | 0.59 | 0.68 |
| $H_{gt}$ et $L_1$ | 0.41 | 0.38 | 0.59 | 0.68 |
| $Stat_{gt}$ et $L_1$ | 0.50 | 0.42 | 0.50 | 0.68 |
| $Edge$ et $d_{Zabih}$ | 0.46 | 0.42 | 0.55 | 0.64 |

TAB. 6.13: *Résultats pour la segmentation mono-critère de* topa_gainsbourg

| critères et distances | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $\min\{(CCV, L_1); (Act_{Sh}, L_1)\}$ | 0.29 | 0.27 | 0.71 | 0.77 |
| $\max\{(CCV, L_1); (Act_{Sh}, L_1)\}$ | 0.48 | 0.46 | 0.52 | 0.58 |
| $\frac{1}{2}(CCV, L_1); \frac{1}{2}(Act_{Sh}, L_1)$ | 0.52 | 0.49 | 0.48 | 0.55 |
| $\frac{3}{4}(CCV, L_1); \frac{1}{4}(Act_{Sh}, L_1)$ | 0.55 | 0.50 | 0.45 | 0.55 |
| $\min\{(CCV, L_1); (RythmEdit, L_1)\}$ | 0.52 | 0.46 | 0.48 | 0.61 |
| $\max\{(CCV, L_1); (RythmEdit, L_1)\}$ | 0.55 | 0.52 | 0.45 | 0.52 |
| $\frac{1}{2}(CCV, L_1); \frac{1}{2}(RythmEdit, L_1)$ | 0.45 | 0.42 | 0.55 | 0.61 |
| $\frac{3}{4}(CCV, L_1); \frac{1}{4}(RythmEdit, L_1)$ | 0.48 | 0.44 | 0.52 | 0.61 |
| $\min\{(CCV, L_1); (RythmEdit, L_1); (Act_{Sh}, L_1); (Stat_{gt}; L_1); (Edge, d_{Zabih}\}$ | 0.36 | 0.33 | 0.65 | 0.71 |
| $\max\{(CCV, L_1); (RythmEdit, L_1); (Act_{Sh}, L_1); (Stat_{gt}; L_1); (Edge, d_{Zabih}\}$ | 0.48 | 0.46 | 0.52 | 0.58 |
| $\frac{1}{5}(CCV, L_1); \frac{1}{5}(RythmEdit, L_1); \frac{1}{5}(Act_{Sh}, L_1); \frac{1}{5}(Stat_{gt}; L_1); \frac{1}{5}(Edge, d_{Zabih})$ | 0.45 | 0.42 | 0.55 | 0.61 |
| $\frac{1}{2}(CCV, L_1); \frac{1}{5}(RythmEdit, L_1); \frac{1}{20}(Act_{Sh}, L_1); \frac{1}{20}(Stat_{gt}; L_1); \frac{1}{5}(Edge, d_{Zabih})$ | 0.48 | 0.44 | 0.52 | 0.61 |

TAB. 6.14: *Résultats pour la segmentation multi-critère de* aim1mb05n1

texture et de contour sont prises en compte dans une troisième expérimentation fusionnant les cinq familles de primitives retenues.

Dans chaque cas, quatre méthodes de fusion des données (cf. sous-section 5.2.4) sont utilisées : le minimum, le maximum, une pondération uniforme, une pondération différenciée. Pour chaque document, le choix des primitives au sein des différentes familles, ainsi que la pondération différenciée sont guidés par les résultats correspondants des primitives pour le cas mono-critère (tableaux 6.9 à 6.13). Les autres paramètres sont fixés selon le tableau 6.8.

Avec ces douze expérimentations multi-critères par document, nous ne pouvons prétendre avoir épuisé les possibilités qu'offre l'utilisation de l'algorithme de macro-segmentation multi-critère. Nous pouvons toutefois en tirer quelques constatations.

La plupart des tests effectués donnent des résultats inférieurs au meilleur des résultats obtenus en mono-critère. À quelques rares exceptions près, les performances se situent dans la même fourchette que celles correspondant au mono-critère sur les primitives utilisées. La première impression est donc que les méthodes mises en œuvre pour fusionner les informations conduisent à une représentation "moyenne", ce qui constitue, dans une certaine mesure, un échec. Cette affirmation peut être tempérée par une étude plus détaillée des résultats qui montre que les indicateurs semblent majoritairement tirés vers le haut de cette fourchette.

Notons néanmoins que huit expérimentations donnent des résultats légèrement meilleurs que le

| critères et distances | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $\min\{(H_{col},\chi_2);(H_{cam},L_1)\}$ | 0.44 | 0.44 | 0.57 | 0.54 |
| $\max\{(H_{col},\chi_2);(H_{cam},L_1)\}$ | 0.52 | 0.53 | 0.48 | 0.46 |
| $\frac{1}{2}(H_{col},\chi_2)+\frac{1}{2}(H_{cam},L_1)$ | 0.54 | 0.56 | 0.46 | 0.44 |
| $\frac{3}{4}(H_{col},\chi_2)+\frac{1}{4}(H_{cam},L_1)$ | 0.52 | 0.53 | 0.48 | 0.46 |
| $\min\{(H_{col},\chi_2);(RythmEdit,L_1)\}$ | 0.44 | 0.44 | 0.57 | 0.54 |
| $\max\{(H_{col},\chi_2);(RythmEdit,L_1)\}$ | 0.50 | 0.51 | 0.50 | 0.48 |
| $\frac{1}{2}(H_{col},\chi_2);\frac{1}{2}(RythmEdit,L_1)$ | 0.48 | 0.49 | 0.52 | 0.50 |
| $\frac{3}{4}(H_{col},\chi_2);\frac{1}{4}(RythmEdit,L_1)$ | 0.50 | 0.51 | 0.50 | 0.48 |
| $\min\{(H_{col},\chi_2);(RythmEdit,L_1);(H_{cam},L_1);(Stat_{gt};L_1);(Edge,d_{Zabih}\}$ | 0.44 | 0.44 | 0.57 | 0.54 |
| $\max\{(H_{col},\chi_2);(RythmEdit,L_1);(H_{cam},L_1);(Stat_{gt};L_1);(Edge,d_{Zabih}\}$ | 0.54 | 0.56 | 0.46 | 0.44 |
| $\frac{1}{5}(H_{col},\chi_2);\frac{1}{5}(RythmEdit,L_1);\frac{1}{5}(H_{cam},L_1);\frac{1}{5}(Stat_{gt};L_1);\frac{1}{5}(Edge,d_{Zabih})$ | 0.48 | 0.49 | 0.52 | 0.50 |
| $\frac{1}{2}(H_{col},\chi_2);\frac{1}{5}(RythmEdit,L_1);\frac{1}{5}(H_{cam},L_1);\frac{1}{20}(Stat_{gt};L_1);\frac{1}{20}(Edge,d_{Zabih})$ | *0.59* | *0.60* | *0.41* | *0.39* |

TAB. 6.15: *Résultats pour la segmentation multi-critère de* aim1mb05n2

| critères et distances | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $\min\{(H_{col},L_1);(Act_{Sh},L_1)\}$ | 0.26 | 0.25 | 0.74 | 0.77 |
| $\max\{(H_{col},L_1);(Act_{Sh},L_1)\}$ | 0.39 | 0.38 | 0.61 | 0.65 |
| $\frac{1}{2}(H_{col},L_1);\frac{1}{2}(Act_{Sh},L_1)$ | 0.29 | 0.28 | 0.71 | 0.74 |
| $\frac{3}{4}(H_{col},L_1);\frac{1}{4}(Act_{Sh},L_1)$ | 0.39 | 0.38 | 0.61 | 0.65 |
| $\min\{(H_{col},L_1);(RythmEdit,L_1)\}$ | 0.19 | 0.19 | 0.81 | 0.84 |
| $\max\{(H_{col},L_1);(RythmEdit,L_1)\}$ | 0.29 | 0.28 | 0.71 | 0.74 |
| $\frac{1}{2}(H_{col},L_1);\frac{1}{2}(RythmEdit,L_1)$ | 0.29 | 0.28 | 0.71 | 0.74 |
| $\frac{3}{4}(H_{col},L_1);\frac{1}{4}(RythmEdit,L_1)$ | 0.32 | 0.31 | 0.68 | 0.71 |
| $\min\{(H_{col},L_1);(RythmEdit,L_1);(Act_{Sh},L_1);(H_{gt};L_1);(Edge,d_{Zabih}\}$ | 0.16 | 0.16 | 0.84 | 0.87 |
| $\max\{(H_{col},L_1);(RythmEdit,L_1);(Act_{Sh},L_1);(H_{gt};L_1);(Edge,d_{Zabih}\}$ | 0.23 | 0.22 | 0.77 | 0.81 |
| $\frac{1}{5}(H_{col},L_1);\frac{1}{5}(RythmEdit,L_1);\frac{1}{5}(Act_{Sh},L_1);\frac{1}{5}(H_{gt};L_1);\frac{1}{5}(Edge,d_{Zabih})$ | 0.29 | 0.28 | 0.71 | 0.74 |
| $\frac{1}{2}(H_{col},L_1);\frac{1}{5}(RythmEdit,L_1);\frac{1}{20}(Act_{Sh},L_1);\frac{1}{20}(H_{gt};L_1);\frac{1}{5}(Edge,d_{Zabih})$ | 0.29 | 0.28 | 0.71 | 0.74 |

TAB. 6.16: *Résultats pour la segmentation multi-critère de* aim1mb08

mono-critère. Dans un unique cas, le taux de correction $T_{err}$ progresse de 2%. Il s'agit de la segmentation *aim1mb05n2* lorsque les cinq signatures sont utilisées avec une pondération différenciée (cf. tableau 6.15). Dans les autres cas, il s'agit principalement d'une légère amélioration du taux d'insertion $T_{ins}$ de 1% sur le document *munich2* (cf. tableau 6.17), et de 3% sur le document *topa_gainsbourg* (cf. tableau 6.18). Sur ce même document, signalons que l'utilisation du maximum sur les primitives de couleur et de mouvement augmente la précision $T_{pre}$ de 3% et le taux d'insertion de 5%.

| critères et distances | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $\min\{(AutoCrlg,L_1);(Act_{Sh},L_1)\}$ | 0.25 | 0.24 | 0.75 | 0.81 |
| $\max\{(AutoCrlg,L_1);(Act_{Sh},L_1)\}$ | 0.49 | 0.48 | 0.51 | 0.54 |
| $\frac{1}{2}(AutoCrlg,L_1);\frac{1}{2}(Act_{Sh},L_1)$ | 0.41 | 0.39 | 0.59 | 0.63 |
| $\frac{3}{4}(AutoCrlg,L_1);\frac{1}{4}(Act_{Sh},L_1)$ | 0.48 | 0.47 | 0.53 | 0.54 |
| $\min\{(AutoCrlg,L_1);(RythmEdit,L_1)\}$ | 0.27 | 0.26 | 0.73 | 0.76 |
| $\max\{(AutoCrlg,L_1);(RythmEdit,L_1)\}$ | *0.54* | *0.53* | *0.46* | *0.48* |
| $\frac{1}{2}(AutoCrlg,L_1);\frac{1}{2}(RythmEdit,L_1)$ | 0.41 | 0.39 | 0.59 | 0.63 |
| $\frac{3}{4}(AutoCrlg,L_1);\frac{1}{4}(RythmEdit,L_1)$ | 0.49 | 0.48 | 0.51 | 0.54 |
| $\min\{(AutoCrlg,L_1);(RythmEdit,L_1);(Act_{Sh},L_1);(Stat_{gt};L_1);(Edge,d_{Zabih}\}$ | 0.29 | 0.28 | 0.71 | 0.75 |
| $\max\{(AutoCrlg,L_1);(RythmEdit,L_1);(Act_{Sh},L_1);(Stat_{gt};L_1);(Edge,d_{Zabih}\}$ | 0.41 | 0.40 | 0.59 | 0.61 |
| $\frac{1}{5}(AutoCrlg,L_1);\frac{1}{5}(RythmEdit,L_1);\frac{1}{5}(Act_{Sh},L_1);\frac{1}{5}(H_{gt};L_1);\frac{1}{5}(Edge,d_{Zabih})$ | 0.36 | 0.35 | 0.64 | 0.66 |
| $\frac{1}{2}(AutoCrlg,L_1);\frac{1}{5}(RythmEdit,L_1);\frac{1}{20}(Act_{Sh},L_1);\frac{1}{5}(H_{gt};L_1);\frac{1}{20}(Edge,d_{Zabih})$ | 0.48 | 0.47 | 0.53 | 0.54 |

TAB. 6.17: *Résultats pour la segmentation multi-critère de* munich2

Les améliorations restent donc rares et assez peu significatives, en comparaison du coût supplémentaire de traitement que requiert le multi-critère tant au niveau de l'extraction des diverses primitives que

de la fusion de celles-ci. Il nous semble probable que le multi-critère a été handicapé par les performances résultantes des primitives autres que celles de couleur, dans la mesure où elles apportent, en même temps que leur information spécifique, un certain nombre de confusions préjudiciables. Outre l'optimisation et l'amélioration des primitives extraites, il nous semble nécessaire de s'intéresser à l'avenir à d'autres méthodes de fusion peut-être plus élaborées.

| critères et distances | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| $\min\{(H_{reg}, L_1); (H_{cam}, L_1)\}$ | 0.36 | 0.33 | 0.64 | 0.73 |
| $\max\{(H_{reg}, L_1); (H_{cam}, L_1)\}$ | *0.68* | *0.63* | *0.32* | *0.41* |
| $\frac{1}{2}(H_{reg}, L_1); \frac{1}{2}(H_{cam}, L_1)$ | 0.64 | 0.58 | 0.36 | 0.46 |
| $\frac{3}{4}(H_{reg}, L_1); \frac{1}{4}(H_{cam}, L_1)$ | *0.73* | *0.67* | *0.27* | *0.36* |
| $\min\{(H_{reg}, L_1); (RythmEdit, L_1)\}$ | 0.46 | 0.42 | 0.55 | 0.64 |
| $\max\{(H_{reg}, L_1); (RythmEdit, L_1)\}$ | *0.73* | *0.67* | *0.27* | *0.36* |
| $\frac{1}{2}(H_{reg}, L_1); \frac{1}{2}(RythmEdit, L_1)$ | *0.73* | *0.67* | *0.27* | *0.36* |
| $\frac{3}{4}(H_{reg}, L_1); \frac{1}{4}(RythmEdit, L_1)$ | *0.73* | *0.67* | *0.27* | *0.36* |
| $\min\{(H_{reg}, L_1); (RythmEdit, L_1); (H_{cam}, L_1); (Stat_{gt}; L_1); (Edge, d_{Zabih}\}$ | 0.41 | 0.38 | 0.59 | 0.68 |
| $\max\{(H_{reg}, L_1); (RythmEdit, L_1); (H_{cam}, L_1); (Stat_{gt}; L_1); (Edge, d_{Zabih}\}$ | *0.68* | *0.63* | *0.32* | *0.41* |
| $\frac{1}{5}(H_{reg}, L_1); \frac{1}{5}(RythmEdit, L_1); \frac{1}{5}(H_{cam}, L_1); \frac{1}{5}(H_{gt}; L_1); \frac{1}{5}(Edge, d_{Zabih})$ | 0.64 | 0.56 | 0.36 | 0.50 |
| $\frac{1}{2}(H_{reg}, L_1); \frac{1}{8}(RythmEdit, L_1); \frac{1}{8}(H_{cam}, L_1); \frac{1}{8}(H_{gt}; L_1); \frac{1}{8}(Edge, d_{Zabih})$ | *0.68* | *0.63* | *0.32* | *0.41* |

TAB. 6.18: *Résultats pour la segmentation multi-critère de* topa_gainsbourg

Pour conclure, nous avons cependant comparé les différentes méthodes de fusion entre elles. Les meilleures performances sont obtenues avec une pondération différenciée. Cette procédure présente toutefois l'inconvénient d'introduire de nouveaux paramètres (les poids), dont le réglage n'est pas automatique. Si l'utilisateur veut s'affranchir du choix de la pondération, la méthode du maximum semble intéressante (bien que donnant des résultats légèrement inférieurs), et celle du minimum doit être proscrite au regard de nos expérimentations. Il s'avère donc que la macro-segmentation est plus robuste lorsque la similarité entre plans prend en compte toutes les primitives considérées.

### 6.3.2 Étude qualitative

Les études qualitatives menées avaient un double objectif. Tout d'abord, nous avons souhaîté analyser finement les erreurs de macro-segmentation et essayer d'en comprendre la nature et les causes. C'est pourquoi, nous avons visualisé, outre les segmentations en séquences, la construction de la hiérarchie, et le critère $d_m$ le long du découpage en plans. Ensuite, nous avons soumis les résultats de la macro-segmentation aux utilisateurs potentiels afin d'obtenir une critique selon le point de vue d'un usage futur de cet outil d'aide à l'indexation. Nous avons mené ce travail en collaboration avec un documentaliste de l'INA, sur l'ensemble des documents, en visualisant pour chacun d'entre eux la meilleure macro-segmentation obtenue (cf. tableau 6.1). La visualisation a été menée avec l'outil *COPA*, dont la figure 6.7 présente une copie d'écran.

#### *6.3.2.1 Analyse des erreurs*

Lorsque nous avons visualisé les hiérarchies construites, nous avons été plutôt satisfaits de l'organisation de l'information qu'elles proposent. Il nous a d'abord semblé que globalement l'utilisation de la fenêtre temporelle était plutôt efficace pour contraindre l'évaluation de la similarité des plans, et que les niveaux de hiérarchisation étaient plutôt pertinents. Sur l'exemple présenté à la figure 6.19, par exemple, les plans similaires de troupeaux ou d'interview sont regroupés dès le début, et s'agglomèrent assez logiquement au fur et à mesure de la construction de la hiérarchie. Ces impressions ont été tempérées par la mise en évidence d'un certain nombre d'erreurs dans les

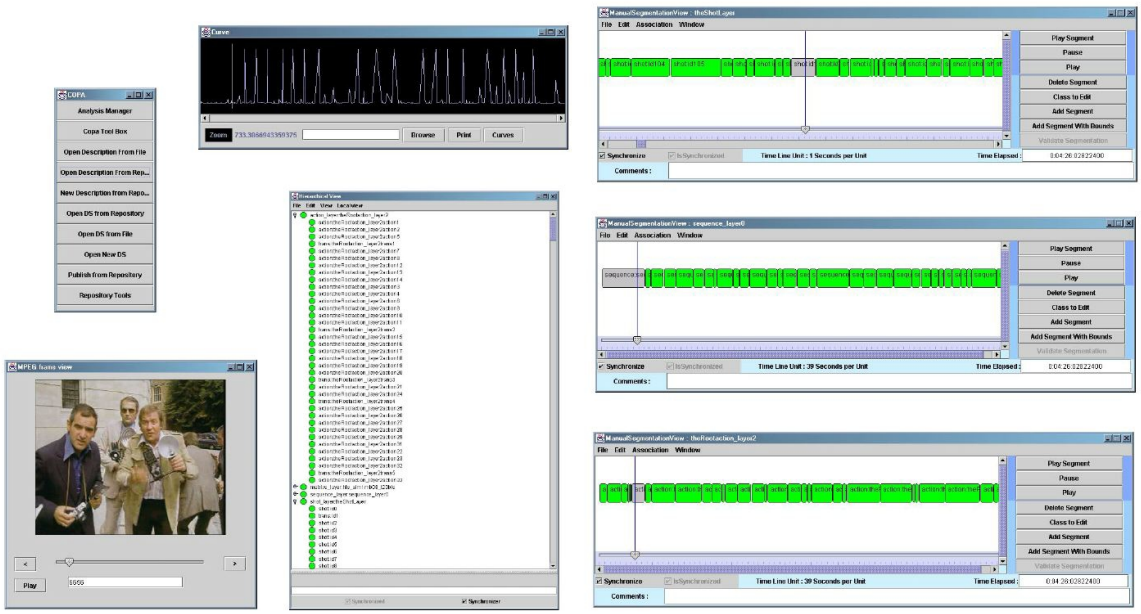


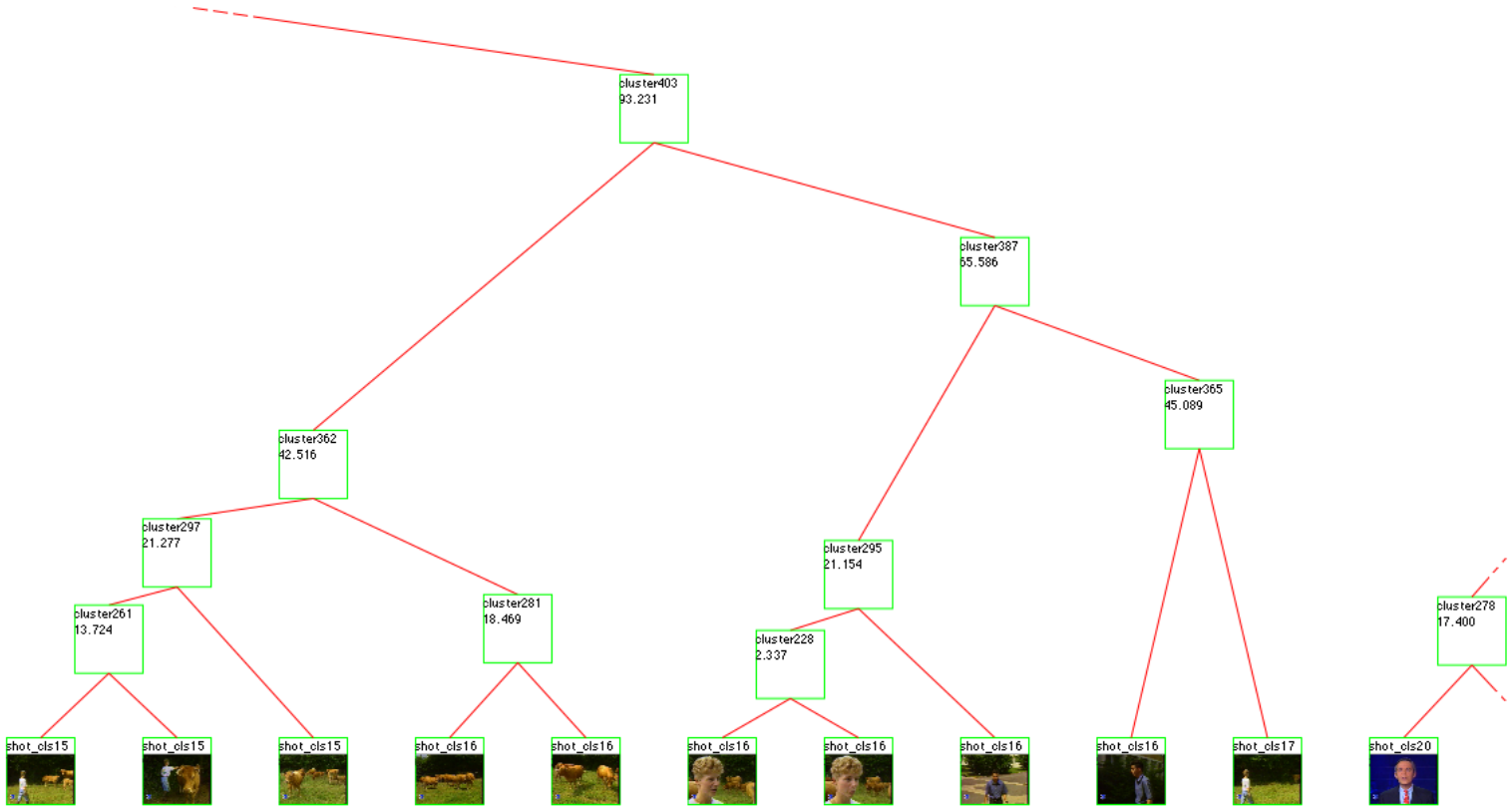
FIG. 6.7: *Copie d'écran de l'outil* Content Provider Application *(COPA). Sont visualisés les découpages en plans, en macro-segments de référence et la macro-segmentation expérimentale, le flux vidéo, le critère de cohérence $d_m$ et une liste des segments temporels*

regroupements que nous détaillons ci-dessous, et par la sensation que les niveaux supérieurs des hiérarchies donnaient lieu à des classifications moins bien maîtrisées.

Les erreurs de macro-segmentation, quelle que soit leur nature, semblent plus nombreuses dans certaines parties des documents. Les séquences contenant des publicités, des bandes-annonces, des génériques, et d'autres formes d'habillage graphique paraissent souvent pertuber l'analyse des macro-segments et le regroupement des plans dans la hiérarchie. Ces plans sont en effet généralement très courts, très variables dans leur contenu, et sans forte corrélation entre le contenu visuel et la sémantique sous-jacente. Lorsqu'une séquence de ce type (un générique, par exemple) contient des plans très différents entre eux, des sur-segmentations peuvent apparaître. Le corrolaire est que des plans des macro-segments environnants peuvent présenter des similarités fortuites, au sens des primitives considérées, avec l'un des plans du générique et entraîner des sous-segmentations irrécupérables, pour peu que le regroupement des plans ait eu lieu rapidement dans la hiérarchie.

Dans le document *topa_gainsbourg*, la plupart des erreurs sont localisées sur des séquences difficiles (effets de flou, incrustation d'image en noir et blanc). Certains changements de séquence sont repérés à un ou deux plans près, à cause du montage un peu particulier de cette émission où certains changements d'interprète se font en un même lieu, presque sans changement de plan (par exemple la fin du générique où il n'y a pas de changement de plan, ou entre l'interprétation *C'est la vie qui veut ça* par Jane Birkin et le sketch *Le tombeur de filles* de Guy Bedos).

Par ailleurs, nous avons pu noter dans *munich2* que les plans montés après une épreuve sportive étaient souvent sur-segmentés. Ces plans, comme ceux d'échauffement tournés avant le déroulement de l'épreuve, sont variés (gros plan sur le gagnant, plan général des athlètes se félicitant, travelling sur un sportif regagnant son banc, plan général de la foule, ralenti, insertion de tableaux de résultats




TAB. 6.19: *Présentation d'un extrait d'une hiérarchie ascendante binaire contrainte par le temps (document* aim1mb05 - *reportage* "Invités Élysée"*)*

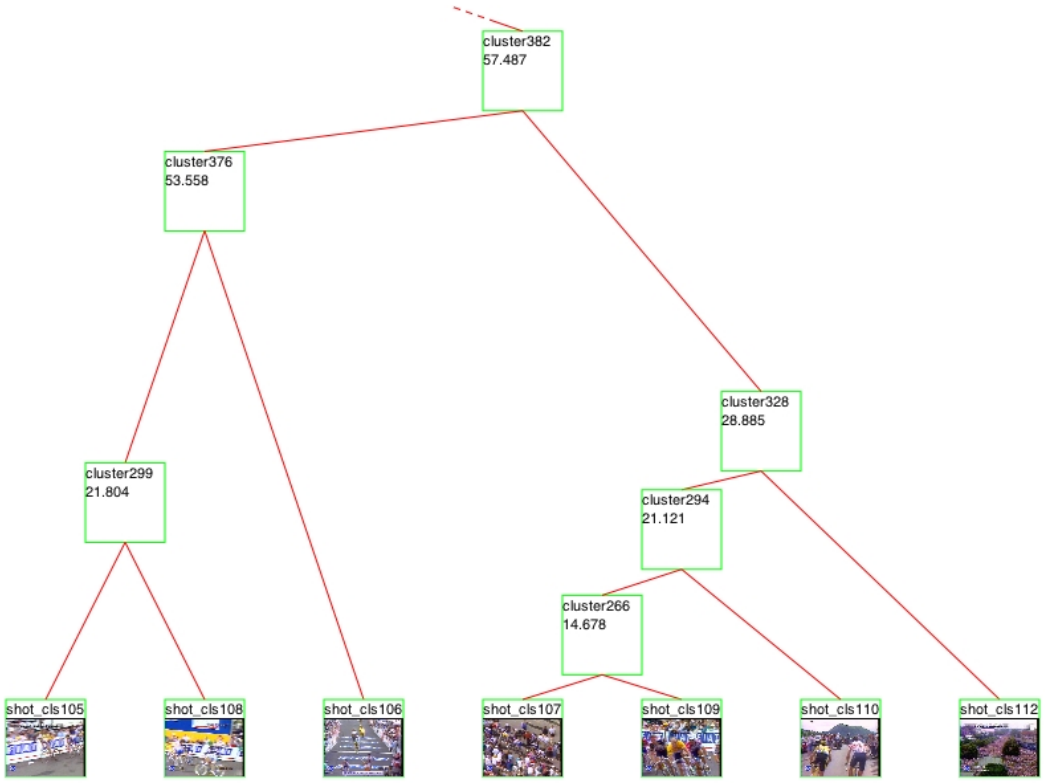



FIG. 6.8: *Présentation d'un extrait d'une hiérarchie ascendante binaire contrainte par le temps (document* aim1mb05 - *reportages* "Tour de France" *et* "Rave party à Berlin"*)*

sur une portion plus ou moins importante de l'image). Cette variété se retrouve dans la construction de la hiérarchie, puis dans le calcul du critère $d_m$, et induit une sur-segmentation.

De manière générale, sur l'ensemble des documents, de nombreuses erreurs sont dues au regroupement, dès les premières itérations, de plans similaires au sens des primitives considérées mais sans réel lien sémantique. Ce phénomène peut être en partie limité par le réglage du paramètre $\Delta T$, toutefois nous avons vu que la réduction de la valeur de ce paramètre peut induire d'autres effets néfastes. Quelques exemples parmi d'autres : dans *aim1mb05*, des plans issus des reportages sur la *Love Parade* et le cyclisme sont confondus très tôt dans la hiérarchie (voir figure 6.8), créant deux oublis et deux sur-segmentations, qui ne pourront être corrigés par un abaissement du seuil, sauf à revenir à un niveau de segmentation proche du plan à plan. Ces confusions existent pour les primitives de couleur, et sont extrêmement courantes pour les primitives plus synthétiques (primitives d'édition et de mouvement). Il est assez fréquent qu'en amont et en aval d'un plan de studio, se trouvent des plans de reportages différents présentant des similarité de durée ou ayant un histogramme de mouvement avec une forte composante *mouvement complexe*. Ainsi, toujours dans *aim1mb05*, l'utilisation de la primitive $Act_{Sh}$ conduit à regrouper des plans issus des reportages sur la voile et sur les sonneurs de cor (voir figure 6.9). Dans les autres documents, ce phénomène se retrouve également. Ainsi, dans *munich2*, les mouvements de la sauteuse en hauteur s'échauffant et du sauteur à la perche se relevant de son saut sont regroupés. Dans *aim1mb08*, cette situation est extrêmement fréquente ; les primitives de couleur confondent, par exemple, les décors plus ou moins brunâtres des différentes pièces, de la péniche, du couloir de l'hôtel particulier, etc. Pour ce

document, l'organisation de l'information dans la hiérarchie semble quelque peu confuse. Si nous pouvons espérer corriger une partie de ces erreurs par le recours à des primitives plus riches, plus discriminantes, cela ne saurait constituer une solution dans la mesure où des signatures trop discriminantes finiraient par rendre fortement dissimilaires des plans liés sémantiquement présentant une certaine variabilité des images. De plus, cette stratégie se heurterait aux remarques faites aux sections 2.1 et 3.1. S'il semble raisonnable de fonder, comme nous l'avons fait, les algorithmes sur l'hypothèse d'une certaine corrélation du signal physique, de la forme du document, et de l'information sémantique, nos travaux montrent qu'une partie du saut à franchir vers des concepts sémantiques restera hors de portée de nos algorithmes.

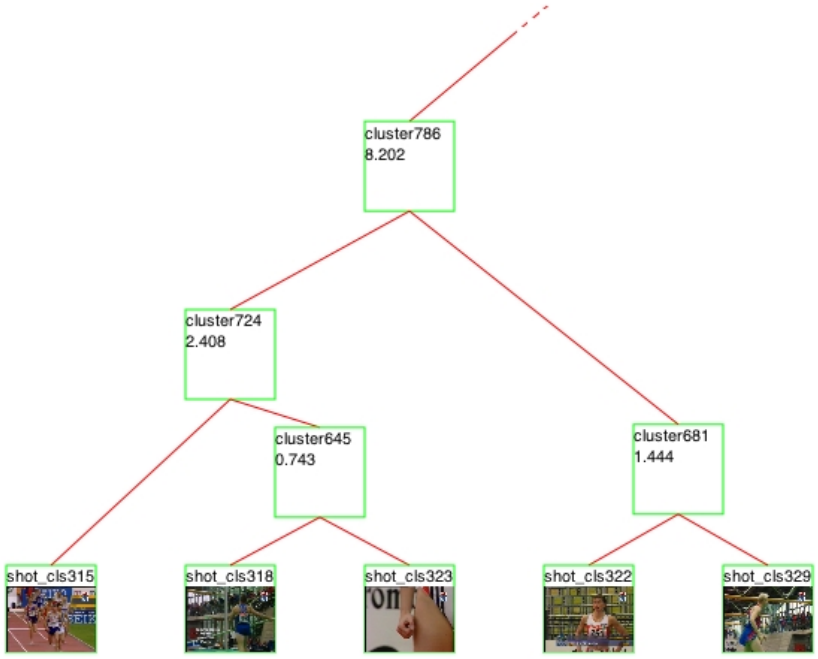



FIG. 6.9: *Présentation d'un extrait d'une hiérarchie ascendante binaire contrainte par le temps (document* munich2 *- épreuves de saut en hauteur et de saut à la perche)*

Inversement, nous avons aussi noté l'absence de regroupement, dans la hiérarchie, de plans appartenant à un même macro-segment. Soit ces derniers sont physiquement similaires mais trop éloignés temporellement (dans ce cas il faudrait accroître $\Delta T$). C'est le cas, par exemple, lors de longues séquences comme celles d'aviron dans *munich2*. Soit les plans sont physiquement différents alors qu'ils participent d'un même contexte sémantique. Nous avions évoqué cette possibilité à la sous-section 4.1.3.3. Nos observations expérimentales confirment la fréquence du phénomène. Nous avons noté que des changements d'éclairage, d'angle de vue, de valeurs de plan, pouvaient provoquer de fortes variations dans les similarités associées aux primitives de couleur comme de mouvement[7], là où le regard et l'interprétation humains voient une forte cohérence.

Lorsque le document possède des plans au contenu récurrent (journal télévisé) ou des plans de

7. Le constat de la sensibilité des signatures de couleur à ces changements a été souvent noté, et nous retrouvons des observations que nous avions pu faire au cours de travaux préliminaires [Veneau 01]. Pareillement, un même mouvement vu sous différents angles, en plan large ou en gros plan, pourra présenter des caractéristiques fort différentes. Nous retrouvons ce constat dans des travaux comme [Fablet 01, Chap. 7], où les résultats des recherches par l'exemple fondées sur des descripteurs globaux du mouvement sont fortement liés à des notions de valeurs de plans.

coupe (vue générale du stade pour la retransmission d'athlétisme), des regroupements de ceux-ci et de tous ceux compris dans l'intervalle apparaissent lorsque $\Delta T$ est trop petit. À l'inverse du cas précédent, il faudrait réduire le $\Delta T$, ce que nous avons tenté de faire sans grand succès (voir tableaux 6.6 et 6.7 et commentaires associés).

#### *6.3.2.2 Évaluation par un documentaliste*

Lors de la visualisation des segmentations obtenues, le documentaliste a été plus particulièrement attentif aux oublis et à la cohérence de lecture éventuellement proposée par la macro-segmentation automatique. L'attention portée aux oublis induit une lecture sévère de nos résultats. En effet, certains oublis peuvent être particulièrement pénalisant pour l'utilisateur. Celui-ci pourrait être amené à faire une mauvaise lecture de l'enchaînement des séquences proposées et omettrait à son tour de documenter une séquence oubliée par l'algorithme. De plus, en cas d'oubli, l'utilisateur n'a pas la possibilité d'identifier l'erreur commise à moins de reprendre en partie la tâche confiée à l'algorithme de macro-segmentation.

Au contraire, la prise en compte de la cohérence de la lecture permet de tolérer certaines erreurs qui ne pénaliseraient pas notablement la compréhension du document, le travail d'indexation à mener au niveau des séquences, et l'accès ultérieur aux informations indexées.

Sur les quatre documents étudiés, deux ont fait l'objet de vives critiques. Le découpage proposé pour la fiction (*aim1mb08*) n'a guère soulevé d'enthousiasme, comme les indicateurs objectifs le laissaient prévoir. Aucune cohérence de lecture n'a pu être dégagée, les macro-segments détectés étant tantôt trop proches des plans, tantôt trop longs. Le déroulement de l'intrigue n'est pas mis en valeur pas notre découpage, le document devient donc difficile à indexer. Plus surprenant, les deux segmentations proposées pour le journal télévisé (*aim1mb05*), pour lesquelles les performances semblaient honorables, ont été fortement critiquées. En effet, les erreurs évoquées plus haut entraînent un certain nombre d'oublis, soit sous forme de regroupement de type {studio-reportage-studio} soit de type {reportage-studio-reportage} qui nuisent au travail d'indexation : mauvaise compréhension du déroulement du journal, manque d'adéquation avec l'indexation très précise - reportage par reportage - effectuée à l'INA. Il semble que la déception de l'utilisateur était d'autant plus forte que le journal télévisé est une structure classique, connue, perçue comme simple par sa construction fondée sur une alternance de plans en studio et de reportage, et largement contrainte par la redondance des plans de studio fortement typés. À notre décharge, nous avons eu l'occasion d'expliquer, plus haut, un certain nombre de ces erreurs. Il convient, en outre, de rappeler que notre algorithme n'utilise aucune information *a priori*, et que la structure du document ou la typologie des plans rencontrés lui sont inaccessibles.

| évaluation sur *munich2* | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| sur la segmentation de référence | 0.54 | 0.53 | 0.46 | 0.48 |
| en différenciant les types d'erreur | 0.88 | 0.71 | 0.12 | 0.37 |
| évaluation sur *topa_gainsbourg* | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| sur la segmentation de référence | 0.73 | 0.67 | 0.27 | 0.36 |
| en différenciant les types d'erreur | 0.91 | 0.80 | 0.09 | 0.14 |

TAB. 6.20: *Résultats avec différenciation des erreurs sur* munich2 *et* topa_gainsbourg

Les deux derniers découpages ont, au contraire, largement séduit notre évaluateur. Dans le cas de l'émission de variété, comme l'indiquaient les indicateurs statistiques, l'adéquation de la segmentation proposée avec celle de référence a pu être confirmée. La visualisation du découpage

de l'émission sportive a montré une grande cohérence de lecture, davantage que ne le laissaient espérer les indicateurs objectifs. Une étude plus précise nous a permis de constater que, à l'inverse du journal télévisé, les erreurs dues à la récurrence de plans de coupe ne gênaient pas la lecture du document.

Nous avons alors demandé au documentaliste de répartir les erreurs en deux classes : graves et bénignes. Dans les sur-segmentations bénignes entrent les ruptures supplémentaires correspondant au passage de l'échauffement à l'épreuve, aux ralentis et photogrammes, aux faux départs, aux détections erronées au sein des publicités et bandes-annonces. Dans les oublis bénins, on trouvera le regroupement d'un plan de coupe avec une épreuve, ou l'ommission d'un tableau de résultats inséré dans une interview. Ne conservant que les erreurs graves (oublis d'épreuves, sur-segmentation des séquences d'aviron, etc.), nous avons recalculé les indicateurs objectifs pour la segmentation proposée. Ceux-ci sont alors sensiblement en amélioration (cf. tableau 6.20) et reflètent davantage le taux de satisfaction exprimé par l'utilisateur. Notons, en particulier, la réduction sensible du taux d'oublis $T_{del}$.

Ce même tableau présente les résultats avec une différenciation similaire des erreurs pour le document *topa_gainsbourg*. Nous avons qualifié de bégnines les sur-segmentations dues à l'utilisation d'effets de flou en marge des séquences, et les oublis et insertions dus à un décalage d'un plan lorsque le changement de séquence se fait dans un même décor.

### 6.3.3 Quelques résultats complémentaires

Outre l'étude de la robustesse aux erreurs de traitements préalables que nous comptions mener dès le début de nos travaux, l'analyse qualitative et quantitative des performances de l'algorithme de macro-segmentation nous a amenés à nous interroger sur des améliorations possibles de certaines parties des traitements. Quelques-unes de ces idées nées de discussions plus ou moins récentes sont présentées dans cette sous-section. Il s'agit notamment d'expérimentations issues de nos interrogations portant sur l'extraction de primitives de mouvement fiables et sur l'information présente dans la hiérarchie. Il s'agit davantage pour nous d'ouvrir des perspectives de recherche, et nous ne prétendons en aucun cas avoir épuisé les quelques pistes évoquées ci-dessous.

#### *6.3.3.1 Robustesse aux erreurs et insertion dans une chaîne de traitement automatique*

Nous avons effectué une segmentation automatique en plans et transitions progressives des documents (cf. sous-section 5.2.1), puis nous avons procédé à une macro-segmentation à partir de celle-ci. Ceci nous a permis d'évaluer le comportement de notre algorithme lorsqu'il est inséré dans une chaîne d'analyse automatique, ainsi que sa robustesse aux erreurs de traitement préalable. Les erreurs produites par l'algorithme de découpage en plans peuvent être des oublis ou des sur-segmentations, ces dernières pouvant être corrigées par notre algorithme de macro-segmentation, alors que les premiers sont sources d'erreurs supplémentaires irrécupérables.

Nous n'avons pas cherché à optimiser ou à modifier l'algorithme de segmentation automatique en plans et transitions progressives. Nous avons ainsi exploité une segmentation en plans et transitions progressives possèdant les caractéristiques suivantes par rapport à notre segmentation en plans manuelle de référence : $T_{cor} = 0.75$, $T_{pre} = 0.58$, $T_{del} = 0.25$ et $T_{ins} = 0.55$, ce qui semble être des valeurs habituelles d'après l'étude décrite dans [Dailianas 95].

Nous avons mené six expérimentations sur le document *munich2* qui sont résumées dans le tableau 6.21.

| Segmentation en plans | automatique | | | | manuelle | | | |
|---|---|---|---|---|---|---|---|---|
| critères et distances (macro-segmentation) | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
| *AutoCrlg* et $L_1$ | 0.36 | 0.35 | 0.64 | 0.66 | 0.54 | 0.53 | 0.46 | 0.49 |
| *RythmEdit* et $L_1$ | 0.17 | 0.16 | 0.83 | 0.92 | 0.27 | 0.26 | 0.73 | 0.76 |
| $Act_{Sh}$ et $L_1$ | 0.17 | 0.16 | 0.83 | 0.86 | 0.25 | 0.24 | 0.75 | 0.80 |
| $H_{gt}$ et $L_1$ | 0.24 | 0.23 | 0.76 | 0.80 | 0.44 | 0.43 | 0.56 | 0.48 |
| *Edge* et $d_{Zabih}$ | 0.27 | 0.27 | 0.73 | 0.75 | 0.36 | 0.30 | 0.64 | 0.83 |
| $\max\{(AutoCrlg, L_1); (RythmEdit, L_1)\}$ | 0.41 | 0.39 | 0.59 | 0.63 | 0.54 | 0.53 | 0.46 | 0.48 |

TAB. 6.21: *Influence des erreurs liées à la segmentation en plans sur la macro-segmentation*

Les résultats montrent, assez logiquement, que notre algorithme de macro-segmentation s'avère sensible aux erreurs de découpage en plans. En moyenne, sur les expérimentations menées, $T_{cor}$ chute de 11%, $T_{pre}$, $T_{del}$ de 12% et $T_{ins}$ de 13%.

Toutefois, si nous considèrons que la segmentation automatique en plans avait déjà conduit à une chute relative de ces indicateurs par rapport à la segmentation en plans manuelle, respectivement de 25%, 42%, 25% et 55%, alors nous pouvons constater qu'une partie des erreurs initiales est rattrapée par l'algorithme de macro-segmentation. Ceci est renforcé par le constat qu'aucune des baisses des indicateurs n'atteint la détérioration initiale observée sur les plans, que les oublis augmentent légèrement moins que les insertions, et que la précision est largement améliorée. Ainsi, particulièrement dans le cas d'une sur-segmentation des plans, notre algorithme de macro-segmentation semble à même de prendre en compte des erreurs sur les plans extraits sans subir une dégradation rédhibitoire de ses performances.

Dans l'hypothèse de son insertion dans une chaîne d'analyse automatique, avec un outil de segmentation en plans proche de l'état de l'art et correctement paramétré afin de fournir, à tout prendre, plutôt une sur- qu'une sous-segmentation des plans, nous retrouverions sans doute des performances proches de celles annoncées lors des expérimentations menées précédemment.

#### *6.3.3.2 Une piste pour améliorer la description du mouvement*

Les remarques sur les limites des primitives de mouvement que nous avions choisies nous ont amenés à nous intéresser au travail sur la reconnaissance du mouvement décrit dans [Fablet 01]. Ces travaux portent sur des modélisations statistiques non paramétriques du mouvement ou plus largement de l'activité contenue dans les scènes. Nous avons pu tester plusieurs variantes des primitives définies dans [Fablet 01], sur un court extrait de soixante-treize plans issus du document *aim1mb08*.

Cet extrait d'un document de fiction constitue un cas d'étude intéressant et difficile. L'analyse en macro-segments semble simple: une séquence d'espionnage, suivie d'une séquence de poursuite en voiture. Toutefois, l'hétérogénéité des images au sein de ces séquences rend le regroupement des plans et la macro-segmentation délicate à maîtriser. De plus, lors de la création de la macro-segmentation de référence, l'identification de la rupture de séquence n'a pas été aisée puisque l'action se déplace, en plusieurs plans, de l'intérieur de l'hôtel où se déroule la scène d'espionnage, à l'extérieur où commence la poursuite en voiture. Ces quelques plans pourraient d'ailleurs être un exemple de "macro-segment de transition" dont nous avions évoqué l'existence éventuelle à la sous-section 4.1.3.3. Cette "transition" entre macro-segments est, de plus, constituée de plans alternés ; le changement de séquences de la segmentation de référence se situe au milieu de ceux-ci. Par conséquent, notre algorithme de macro-segmentation est censé trouver une forte cohérence

au niveau de la rupture théorique et ne fournir au mieux qu'un découpage à la limite de cette "transition".

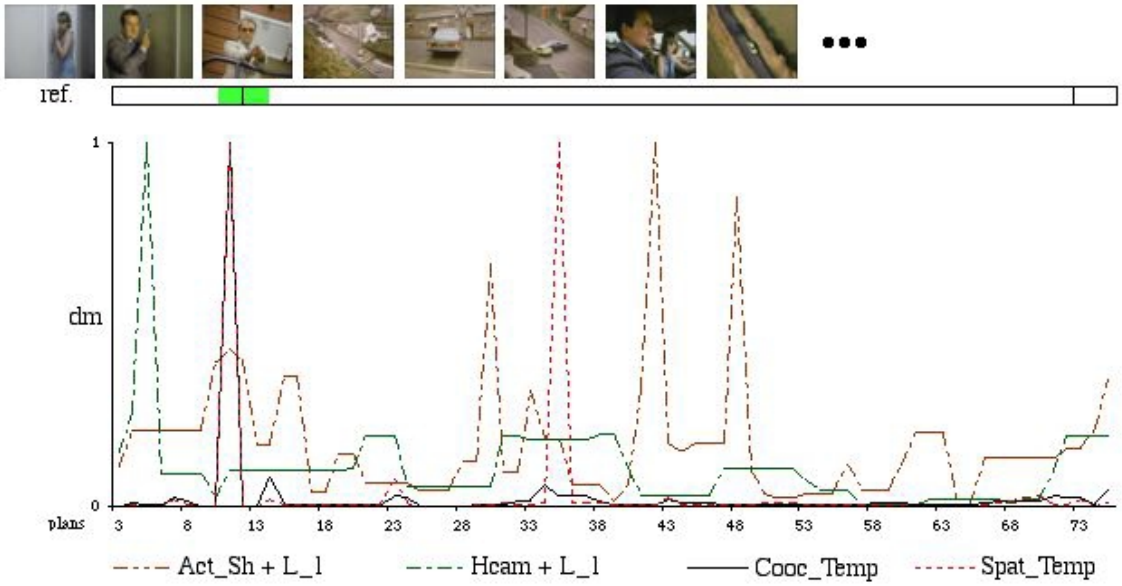



FIG. 6.10: *Visualisation du critère de cohérence $d_m$ calculé à partir de quatre primitives de mouvement différentes. En bas, l'évolution du critère $d_m$ sur les plans 3 à 73 du document* aim1mb08. *Au milieu, la macro-segmentation manuelle de référence : en trait plein les ruptures de séquences, en grisé la zone de transition possible. En haut, quelques images-clefs de la séquence considérée.*

La figure 6.10 présente la macro-segmentation de référence (en trait plein la rupture indiquée dans celle-ci, en grisé l'ensemble des plans formant la transition entre la scène d'espionnage et la poursuite en voiture), et les critères de cohérence $d_m$ obtenus à partir des deux primitives de mouvement $Act_{Sh}$ et $H_{cam}$ utilisées au cours de notre étude, ainsi que deux des primitives introduites dans [Fablet 01] : $Cooc_{Temp}$ et $Spat_{Temp}$. Par ailleurs, nos algorithmes ont été configurés comme indiqué dans le tableau 6.8.

L'étude de ce graphique suggère que les primitives proposées dans [Fablet 01] pourraient être plus pertinentes pour rendre compte de l'information de mouvement pour la macro-segmentation. En effet, le critère $d_m$ calculé à partir de $H_{cam}$ et $Act_{Sh}$ s'avère assez bruité et présente de nombreux pics significatifs en dehors de la zone de transition. Le critère $d_m$ pour $Cooc_{Temp}$ et $Spat_{Temp}$ apparaît, au contraire, moins bruité et présente un pic principal au niveau des plans 11-12, c'est-à-dire à la limite de la zone de transition.

Ces résultats constituent une piste de recherche intéressante qui nécessiterait d'être confirmée par une étude plus conséquente.

#### *6.3.3.3 Complément sur le multi-critère*

Les résultats mitigés concernant notre algorithme de macro-segmentation multi-critère ont conduit à quelques expérimentations supplémentaires. Nous avions constaté par la visualisation des différentes hiérarchies que celles-ci constituaient une bonne représentation de l'organisation de l'information contenue dans le document pour peu que les primitives utilisées soient suffisamment

pertinentes (cf. figure 6.19).

Une première extension envisagée est de faire varier les conditions de regroupement des plans, puis des classes. Ainsi, afin d'affiner la représentation de l'information, nous aimerions, par exemple, que les plans commencent par être regroupés en tenant compte surtout de l'information de couleur. Puis, une fois ces regroupements opérés, que nous puissions faire émerger des classes se fondant sur l'information de mouvement. Pour cela, nous avons modifié notre algorithme de macro-segmentation en faisant varier les poids respectifs des primitives au cours de la construction de la hiérarchie. Nous avons été cependant confrontés à de nombreux problèmes liés principalement aux paramètres supplémentaires nécessaires. Comment fixer les poids initiaux et finaux ? Comment guider l'évolution de ceux-ci au cours du temps ?

Nous avons opté pour une solution simple au travers d'une évolution linéaire des poids au cours du temps. Pour une expérimentation sur la vidéo *aim1mb05*, nous avons donné une pondération initiale plus forte à la primitive d'édition dans la mesure où les reportages sont généralement constitués de plans courts, et les séquences en studio d'un plan unique. Pour le document *aim1mb08*, nous avons pris en compte que la segmentation liée au décor était proche des plans (cf. annexe A.2), et nous avons donné une importance initiale plus forte à la couleur. De même, pour *munich2*, nous avons cherché à regrouper d'abord les plans liés à un même décor, avant de prendre en compte les mouvements spécifiques aux épreuves. Pour *topa_gainsbourg*, nous avons observé de nombreuses variations d'éclairage au sein des séquences. Nous avons donc tenté de nous affranchir des différences d'éclairage en donnant la priorité à l'information de mouvement. Les autres paramètres sont ceux donnés dans le tableau 6.8.

| documents et pondération des expérimentations | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| aim1mb05n1 [$CCV\ w_{init} = 0.25\ w_{fin} = 0.75$] [$RythmEdit\ w_{init} = 0.75\ w_{fin} = 0.25$] | 0.48 | 0.46 | 0.52 | 0.58 |
| aim1mb05n2 [$H_{col} + \chi_2\ w_{init} = 0.25\ w_{fin} = 0.75$] [$RythmEdit\ w_{init} = 0.75\ w_{fin} = 0.25$] | 0.48 | 0.49 | 0.52 | 0.50 |
| aim1mb08 [$H_{col} + L_1\ w_{init} = 0.75\ w_{fin} = 0.25$] [$Act_{Sh}\ w_{init} = 0.25\ w_{fin} = 0.75$] | 0.32 | 0.31 | 0.68 | 0.71 |
| munich2 [$AutoCrlg\ w_{init} = 0.75\ w_{fin} = 0.25$] [$Act_{Sh}\ w_{init} = 0.25\ w_{fin} = 0.75$] | 0.42 | 0.42 | 0.58 | 0.59 |
| topa_gainsbourg [$H_{reg} + L_1\ w_{init} = 0.25\ w_{fin} = 0.75$] [$H_{cam}\ w_{init} = 0.75\ w_{fin} = 0.25$] | 0.64 | 0.58 | 0.36 | 0.46 |

TAB. 6.22: *Résultats obtenus avec une variation de la pondération dans la macro-segmentation multi-critère*

Les résultats regroupés dans le tableau 6.22 sont inclus dans la fourchette de ceux précédemment obtenus (tableaux 6.14 à 6.18). Il semblerait que la modification apportée ne soit pas suffisante pour construire une représentation du document plus pertinente. Si l'idée initiale semble intéressante, la mise en œuvre proposée implique des complications trop nombreuses, notamment lorsqu'il s'agit de fixer les paramètres supplémentaires.

La seconde extension consiste à considérer que pour chaque type de primitives, l'information récupérée est correctement représentée dans les hiérarchies mono-critères. Il conviendrait alors de fusionner les hiérarchies mono-critères obtenues. Dans le domaine de la classification des données, il existe un certain nombre de méthodes permettant d'effectuer des consensus de hiérarchies [Gordon 99, sec. 4.5]. Nous avons utilisé l'outil PHYLIP [8] [Felsenstein 89] avec ses paramètres par défaut. L'algorithme de consensus fusionne deux hiérarchies construites avec des signatures différentes, selon la méthode dite du "consensus strict", dans notre cas d'étude. Sont alors conservées dans la hiérarchie résultante les classes communes aux deux hiérarchies. Pour ces expérimentations, les paramètres de notre algorithme de macro-segmentation restent fixés comme indiqué dans le tableau 6.8.

8. Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c, distribuée par l'auteur. Department of Genetics, University of Washington, Seattle.

| documents | $T_{cor}$ | $T_{pre}$ | $T_{del}$ | $T_{ins}$ |
|---|---|---|---|---|
| aim1mb05n1 [$CCV$ et $RythmEdit$] | 0.42 | 0.38 | 0.58 | 0.68 |
| aim1mb05n2 [$CCV$ et $RythmEdit$] | 0.52 | 0.52 | 0.48 | 0.48 |
| aim1mb08 [$H_{col} + L_1$ et $Act_{Sh}$] | 0.26 | 0.26 | 0.74 | 0.74 |
| munich2 [$AutoCrlg$ et $Act_{Sh}$] | 0.51 | 0.37 | 0.49 | 0.88 |
| topa_gainsbourg [$H_{reg} + L_1$ et $H_{cam}$] | 0.41 | 0.33 | 0.59 | 0.82 |

TAB. 6.23: *Résultats obtenus par consensus des hiérarchies mono-critères*

Les performances atteintes (tableau 6.23) sont quelque peu décevantes. Dans deux cas sur cinq, les indicateurs objectifs sont légèrement meilleurs. Nous n'avons pu obtenir d'amélioration sensible de nos segmentations. L'augmentation du nombre d'insertions sur certains documents est due au fait que, compte tenu de la forme de la hiérarchie fusionnée, nous n'avons pu obtenir le nombre de macro-segments souhaités. Notons tout de même qu'une étude plus précise des paramètres de l'algorithme de consensus, des différentes méthodes de fusion (consensus strict, consensus médian, intersection d'arbre), serait nécessaire, ainsi que sa généralisation à un plus grand nombre de hiérarchies.

# Conclusion sur la macro-segmentation

Afin de conclure sur cette partie de nos travaux, nous rappelerons les difficultés auxquelles nous avons été confrontés, les résultats obtenus à l'issue des expérimentations, et les pistes d'études complémentaires qui nous ont paru émerger.

## Des difficultés encore nombreuses

### La gestion des paramètres

Les problèmes de gestion des paramètres évoqués à la sous-section 4.2.5 restent sensibles. Nous revenons, notamment, sur l'utilisation des paramètres $\Delta T$ et $N_{req}$, ainsi que sur le choix des options.

**Utilisation des paramètres $\Delta T$ et $N_{req}$.** Nos conclusions sur la fenêtre temporelle $\Delta T$ sont mitigées. Nous n'avons pu apporter de solution complètement satisfaisante à la gestion du paramètre $\Delta T$, au contraire, nous avons constaté la difficulté d'un réglage adéquat, par la mise à jour des défauts respectifs d'un $\Delta T$ trop long ou trop court (cf. tableaux 6.6 et 6.7). Par défaut, une durée de deux minutes nous a semblé être une valeur acceptable pour ce paramètre.

Par l'utilisation du critère de cohérence $d_m$ et la conversion de la valeur de seuillage $\delta$ en un nombre de séquences requises $N_{req}$, nous estimons avoir proposé une solution plus élégante que celle donnée dans [Yeung 98]. D'un point de vue algorithmique, il n'est plus nécessaire de recalculer le graphe STG lorsque le seuillage varie, et la gestion dynamique de ce paramètre s'en trouve nettement facilitée. Du point de vue de l'utilisateur, la donnée du nombre de séquences requises semble plus intuitive qu'une hypothétique valeur de seuil applicable à une hiérarchie. De plus, dans le cadre de la chaîne documentaire de l'INA, cette donnée s'avère relativement adéquate, notamment pour le traitement de nombreux documents pour lesquels le conducteur est fourni et le nombre des reportages ou des prestations artistiques disponible.

**Choix des options.** En ce qui concerne les différentes options (paramètres, signatures, variantes algorithmiques, etc.), la situation reste aussi contrastée. L'étude de la fonction $W$ semble suggérer le choix d'une fonction non constante, sans que ce choix soit réellement décisif en l'absence de tests complémentaires.

Les résultats obtenus sur la mise à jour de la hiérarchie par différentes méthodes montrent la pertinence des modifications introduites, en particulier la méthode de Ward associée à la cardinalité des classes semble se substituer avantageusement à la méthode du lien maximal.

Une prise en compte plus globale des distances cophénétiques entre plans dans le critère de cohérence $d_m$ reste ouverte, et doit permettre de diminuer le nombre d'oublis, ce que l'utilisation des quantiles d'ordre non nul n'est pas parvenue à effectuer.

D'autre part, l'utilisation d'une légère sur-segmentation permettant la diminution des oublis n'a pu être vérifiée. Enfin, l'introduction du multi-critère ne s'est pas avérée déterminante, tout en introduisant quelques paramètres de pondération supplémentaires.

### Dépendance de l'algorithme aux traitements préalables

Notre algorithme s'est avéré sensible à certaines erreurs de prétraitements, comme la plupart des algorithmes que nous avions étudiés dans la bibliographie. Si l'outil de macro-segmentation paraît capable de corriger en partie des erreurs de sur-segmentation des plans, il est au contraire dépendant de l'extraction des primitives. Les principales difficultés rencontrées sont de nature pratique (intégrer, gérer et paramétrer correctement les outils d'extraction) et théorique (trouver des primitives suffisamment fiables et riches en information). Les erreurs dues à l'inexactitude où à la pauvreté de l'information contenue dans les primitives extraites sont particulièrement pénalisantes lors de la suite des traitements.

### Le saut entre le niveau numérique et l'interprétation sémantique

Nous avions largement anticipé cette difficulté, et c'est sans surprise que nous y avons été confronté. Dans notre corpus d'étude, le document de fiction *aim1mb08*, notamment, nous a permis de mettre en lumière la différence, sans doute en partie irréductible, entre l'interprétation globale et contextuelle d'un document par un documentaliste et l'analyse localisée à laquelle se livrent les algorithmes de macro-segmentation.

L'introduction d'informations *a priori*, comme la structure particulière d'un type de document (par exemple, un journal télévisé suit un canevas bien établi) pourrait être une solution, mais la manière de les formaliser et de les exploiter dans notre algorithme de macro-segmentation reste à définir.

## Bilan des résultats obtenus

### Comparaisons avec d'autres méthodes

Toute comparaison avec les autres méthodes existantes se doit de prendre en compte les remarques énoncées à la sous-section 4.2.6, le fait que notre algorithme de macro-segmentation est à visée généraliste, la nature de notre corpus (diversifié et parmi les plus longs utilisés), et la méthodologie d'évaluation que nous avons voulue rigoureuse. Ainsi, il doit être reprécisé que les segmentations qui nous ont servi de référence ont été construites en fonction d'un usage, d'une pratique professionnelle et non des capacités supposées des algorithmes. C'est notamment le cas pour la fiction *aim1mb08* où une segmentation de référence fondée sur un regroupement des plans similaires visuellement aurait sans doute donné des résultats plus flatteurs.

Si nous restreignons notre propos aux auteurs qui ont fourni des données d'évaluation quantitative sur des corpus conséquents (quelques séquences d'au moins une dizaine de minutes), nous trouvons les performances suivantes :

- Lorsque les outils sont dédiés (segmentation de journaux télévisés), les résultats sont en général nettement meilleurs. Ainsi, dans [Zhang 94], sont annoncés des taux de correction proches de 100% et des taux d'erreur quasi nuls. De même, dans [Ariki 96], les taux de correction sont peu éloignés de 100%, mais ceux-ci sont calculés sur le classement de chaque image et non sur la décision prise par l'algorithme aux ruptures de plan. Notons que dans [Hauptmann 95], des taux d'oubli proches de 27% et des taux de fausses alarmes proches de 8% sont indiqués.

– Dans [Lienhart 99], le corpus est constitué de deux films, et des taux d'oubli compris entre 4% et 8% et des taux de fausses alarmes entre 11% et 18% sont affichés. Dans [Rui 99b], une dizaine d'extraits de dix minutes environ ont été traités, et des taux d'oubli entre 0% et 16% et des taux de fausses alarmes entre 0% et 50% sont atteints. Enfin, dans [Hanjalic 99], il n'est précisé ni la nature du corpus considéré (sans doute des fictions), ni les valeurs obtenues pour son indicateur de qualité $T_{nqual}$. Par contre, les auteurs annoncent un taux de correction de 69%, un taux d'oubli de 31% et un taux de fausses alarmes de 5%.

D'un point de vue purement arithmétique, nos résultats sont en règle générale légèrement inférieurs à ceux communiqués par ces auteurs. Compte tenu des précisions apportées ci-dessus et des contraintes que nous nous sommes données, nos résultats sont cependant largement encourageants. Du point de vue des usages, les premiers retours ont été très positifs sur la moitié du corpus traité.

### Classification hiérarchique et critère de cohérence

Nous avons introduit l'utilisation de la distance cophénétique pour la macro-segmentation. Celle-ci traduit la pertinence de l'information organisée au sein de la hiérarchie. Elle permet aussi le calcul du critère de cohérence $d_m$. Les résultats obtenus pour certaines segmentations, ainsi que l'analyse plus qualitative menée sur les hiérarchies, semblent valider nos choix algorithmiques.

### Proposition d'un cadre méthodologique pour l'évaluation

#### *Proposition d'un corpus de référence*

Nos travaux vont permettre de rendre disponible pour la communauté qui s'y intéresse une première proposition de corpus et une annotation manuelle associée. Nous avons aussi pu initier une réflexion sur les usages liés au développement des outils automatiques de macro-segmentation.

#### *Aspects méthodologiques liés à l'évaluation*

L'utilisation conjointe de critères quantitatifs et d'une analyse qualitative liée à la notion de cohérence de lecture constitue une première suggestion méthodologique pour l'évaluation des résultats de traitements automatiques de documents audiovisuels. Un enrichissement de l'information apportée par une évaluation qualitatitive reste cependant lié à une étude plus approfondie des usages en évolution des différents utilisateurs et clients des fonds de l'INA (à l'Inathèque de France comme au Département des Archives).

### Adéquation à nos objectifs

Nous nous étions posé à la sous-section 6.2.1 une série de questions, auxquelles nous allons essayer de répondre.

1. Le traitement automatique proposé par notre outil de macro-segmentation a-t-il une validité pratique ? La réponse semble positive sur une partie du corpus étudié. Toutefois, le travail doit être prolongé notamment pour ce qui concerne l'extraction de primitives, leur coopération, et l'introduction de contraintes.

2. Le paramétrage de l'algorithme est-il aisé, intuitif? La réponse est plus contrastée. Pour une partie des paramètres, la réponse est non, même si nous avons proposé quelques solutions,

c'est notamment le cas du paramètre $\Delta T$. Pour les autres, il serait nécessaire de conforter nos résultats par des tests complémentaires, mais nos études ont contribué à fixer des options à des valeurs par défaut, et parfois à en proposer des améliorations (calcul du critère $d_m$, méthode de Ward). Nous n'avons su par ailleurs que marginalement utiliser des informations extérieures (à travers le réglage de $N_{req}$), et la question du paramétrisation et du pilotage des algorithmes reste ouverte.

3. Obtenons-nous une évaluation différenciée de l'outil selon la typologie des documents étudiés ? S'agit-il d'un outil générique ? Nous obtenons en effet des résultats différenciés selon les documents, malheureusement notre corpus est insuffisant pour généraliser ces résultats aux genres auxquels ils appartiennent. Enfin, nous obtenons des résultats honorables, bien que perfectibles, sur deux, voire trois, des documents de notre corpus. Compte tenu de la diversité de celui-ci, il semble que l'outil proposé soit en effet généraliste et capable de traiter des types de documents variés. Toutefois, l'algorithme proposé possède aussi ses limites. Celles-ci sont liées aux difficultés du saut à franchir lorsque les liens entre numérique et sémantique sont trop subtils (cas de la fiction), ou à une inadéquation entre le modèle implicitement à l'œuvre et la récurrence de plans similaires intervenant comme une ponctuation du document (cas du journal télévisé).

## Des pistes à suivre

### Un travail à mener sur les usages

Il sera sans aucun doute utile de poursuivre le travail autour de la constitution d'un corpus, ainsi que des usages. Il serait intéressant d'augmenter le corpus afin d'avoir plusieurs documents par genre et de s'affranchir d'éventuelles particularités. De même, il serait sans doute souhaitable de différencier les documents de fiction en catégories plus fines (par exemple, théâtre, film, sitcom, téléfilm, etc.). Ensuite, il conviendrait d'affiner les segmentations de référence, à la lumière de nos résultats, en les confrontant de nouveau aux pratiques des documentalistes. Une réflexion commune sur la catégorisation des erreurs, sur la notion de cohérence de lecture, et sur celle de macro-segments de transition serait sans aucun doute enrichissante. Enfin, ces avancées permettraient de définir des profils de paramétrisation selon des catégories de documents ou des réglages expérimentaux plus aisés dans la suite des travaux de [Rui 98], où les paramètres sont dans la mesure du possible la combinaison d'une constante universelle et d'une variable liée au type du document traité. Une telle stratégie serait particulièrement pertinente dans le contexte de l'indexation menée à l'INA.

### Extraction des primitives

Outre le problème de leur intégration et du réglage des paramètres, il conviendrait d'étudier des primitives plus riches en informations. Pour ce qui concerne la couleur et le mouvement, les travaux développés respectivement dans [Vertan 00] et [Fablet 01] pourraient constituer de bonnes pistes d'étude. Les travaux présentés dans la partie III pourraient offrir des perspectives intéressantes pour la macro-segmentation en proposant comme prétraitements des résultats de classification de segments temporels. Une autre voie est celle de l'utilisation de la bande audio [Saraceno 97], voire de résultats de transcriptions [9].

9. Des travaux sont actuellement en cours à l'INA sur ce thème [Baras 02].

### Construction de la hiérarchie et calcul du critère

Outre une étude des différentes méthodes de fusion des hiérarchies, la recherche de méthodes permettant de prendre en compte successivement différents critères serait sans doute pertinente. De même, une construction de la hiérarchie en plusieurs passages, afin de donner une plus grande cohérence à celle-ci [Llach 99], permettrait notamment de fiabiliser le critère $d_m$. La mise en œuvre d'un seuillage adaptatif ou localisé du critère permettrait peut-être de mieux gérer les passages des documents jugés difficiles.

### Spécialiser l'algorithme, utiliser la para-documentation

Une dernière voie pertinente concerne l'introduction efficace et simple d'informations *a priori*. Ainsi, une typologie des documents, mais aussi l'information contenue dans les conducteurs, ou plus généralement celle présente dans l'ensemble de la para-documentation qui sera rendue disponible en version numérique par les fournisseurs de contenu audiovisuel dans les années à venir[10], pourraient être pris en compte pour guider les algorithmes de macro-segmentation.

10. Voir à ce sujet les projets *Captation* et *Extranorm* programmés à l'INA, ainsi que les travaux en cours, par exemple [LeRoux 02].

## Troisième partie

# Caractérisation de séquences audiovisuelles fondée sur l'analyse du mouvement

# Introduction

La caractérisation du contenu des segments audiovisuels est une des tâches principales de l'indexation avec le repérage temporel de ces segments et la structuration des annotations effectuées.

Nous nous sommes plus particulièrement intéressés à la caractérisation du contenu dynamique des segments temporels. Ceci implique, par conséquent, de considérer une double problématique. Il s'agit, d'une part, de définir des descripteurs du mouvement pour les segments temporels, et, d'autre part, d'identifier une méthode de classification permettant de passer de ces descripteurs numériques au niveau de caractérisation souhaité. Un tel objectif couvre donc un domaine assez large : de l'extraction de primitives du flux audiovisuel à leur interprétation dans le contexte d'une description haut-niveau des segments. Ainsi, nous serons de nouveau confrontés à la question du passage du niveau numérique au niveau sémantique évoquée dans la première partie.

Dans le chapitre 7, nous positionnerons plus précisément notre étude par rapport aux nombreux travaux menés sur l'analyse du mouvement dans des séquences d'images. Nous nous attacherons plus particulièrement à dégager des solutions pour les deux aspects mentionnés plus haut : l'extraction de primitives liées au mouvement et la classification des séquences. Les méthodes développées seront définies et motivées au chapitre 8. Enfin, nous avons souhaité évaluer la chaîne de traitements proposée du point de vue algorithmique et du point de vue des usages. La méthodologie adoptée, la description des expérimentations menées et les commentaires sur les résultats obtenus constituent le chapitre 9.

# Chapitre 7

# Contexte des travaux

## 7.1 Positionnement des objectifs

Nous allons, dans cette section, définir le positionnement de notre étude en fonction du cadre général défini à la section 3.3 et des travaux menés sur l'analyse du mouvement dans des séquences d'images.

### 7.1.1 Quelques considérations générales à propos des travaux sur le mouvement

L'ensemble des études portant sur l'analyse du mouvement dans des séquences d'images forme un domaine vaste et varié. En amont, il peut se fonder sur des approches statistiques, géométriques ou variationnelles, ainsi que sur des travaux concernant le système perceptif humain[1]. En aval, les domaines d'applications sont multiples: indexation vidéo, imagerie médicale, imagerie satellitaire, vidéo-surveillance, robotique ou interface homme-machine. Ce domaine d'étude englobe différents problèmes allant de la détection du mouvement à son interprétation, en passant par des questions de segmentation, de mesure, de suivi temporel ou d'extraction de descripteurs à différents niveaux.

A. Bobick suggère de classer les méthodes liées au mouvement selon que leur objectif est de déterminer le *mouvement*, l'*activité* ou l'*action*[2] [Bobick 97]. Sous le label *mouvement*, cet auteur regroupe l'ensemble des méthodes permettant l'extraction de primitives de bas niveau liées au mouvement. Les méthodes relevant de la notion d'*activité* considèrent généralement le mouvement comme une séquence d'états et synthétisent une information plus globale au travers de statistiques liées à la séquence considérée. Enfin, les méthodes concernées par la notion d'*action* nécessitent des connaissances contextuelles *a priori* de haut-niveau permettant de guider les algorithmes[3]. La classification de A. Bobick est évidemment influencée par ses propres travaux (principalement la reconnaissance de gestes et l'interaction homme-machine), toutefois ce point de vue semble partagé par d'autres, et notamment par H.-H. Nagel[4] [Bobick 97]. La taxonomie de description due à R. Nelson[5] est assez différente de la classification évoquée plus haut, toutefois elle repose aussi

1. Cette préoccupation est présente, par exemple, dans [Adelson 85,Heeger 88,Jasinschi 91,Nelson 92].

2. Soit, en anglais, *Movement*, *Activity*, *Action*.

3. Ces méthodes sont souvent qualifiées, en anglais, de *domain knowledge based*, *closed world* ou *context specific methods*.

4. Sur ce sujet, on pourra se reporter aux réflexions menées au cours des années par cet auteur, de [Nagel 88] à [Nagel 01].

5. La taxonomie proposée par R. Nelson est reprise par O. Chomat selon les catégories suivantes: (i) les actions primitives; (ii) les actions contextuelles simples; (iii) les actions intentionnelles dans un contexte complexe; (iv) les actions complexes et multi-partis; (v) les vecteurs de communication [Chomat 00, p. 16].

sur différents niveaux d'abstraction, et donc sur différents contextes interprétatifs. Si nous nous replaçons dans le contexte de l'indexation vidéo automatique, la problématique du saut sémantique se situerait sans doute entre l'étude des *activités* et des *actions*.

Au-delà de la question du niveau d'interprétation du mouvement, des auteurs ont proposé des états de l'art autour des principales directions de recherche dans ce domaine. Ainsi, R. Fablet énumère dans [Fablet 01] les problématiques suivantes : mesure du mouvement, détection du mouvement, segmentation du mouvement, suivi de primitives ou de régions, interprétation du mouvement (reconnaissance et classification), reconstruction 3D, reconnaissance et modélisation d'activités. Dans [Cedras 95], nous retrouvons ces thématiques présentées selon une grille de lecture quelque peu différente et moins systématique. Quoiqu'il en soit, ces thématiques illustrent à la fois les différents niveaux d'interprétation possibles du mouvement et les applications variées que cela peut concerner.

Un autre point de vue est de considérer les différents types de mouvement étudiés. Nous en trouvons une cartographie assez complète dans [Fablet 01] : mouvement rigide, mouvement articulé, mouvement déformable, mouvement fluide, et textures temporelles. Nous pourrions éventuellement rajouter les mouvements de groupes, dans le cadre de l'étude de mouvements coopératifs d'un ensemble d'entités (voir, par exemple, le cas des sports collectifs).

Enfin, concernant la modélisation de l'activité d'un objet, et notamment de l'activité humaine, un regroupement en trois familles permet de différencier les techniques selon l'approche adoptée [Chomat 00] : méthodes fondées sur un modèle géométrique, méthodes fondées sur l'apparence globale des objets, méthodes fondées sur l'apparence locale des objets.

Rappelons que le mouvement fourni par l'observation de séquences d'images n'est qu'un mouvement apparent, c'est-à-dire la projection 2D d'un mouvement 3D perçu, de plus, à travers la variation temporelle des intensités. Ce mouvement apparent résulte du mouvement relatif entre la caméra et les éléments de la scène, de la profondeur de la scène, des conditions d'illumination de la scène, des caractéristiques intrinsèques de la caméra. Le mouvement apparent est modélisé à l'aide de la *fonction plénoptique* par Adelson et Bergen [Chomat 00, Sec. 3.2], au travers de sept dimensions liées au point de vue, au temps, à la position spatiale, et à la longueur d'onde.

Une autre difficulté classique à laquelle est confrontée l'analyse du mouvement dans une séquence d'images est le problème de l'ouverture (*aperture problem*). L'hypothèse de conservation de l'intensité d'un point en mouvement au cours du temps est exploitée dans la plupart des techniques proposées. Cette hypothèse conduit notamment à la dérivation de l'Équation de Contrainte du Mouvement Apparent (ECMA) [Horn 81]. L'examen de cette équation montre que seule la composante de la vitesse parallèle au gradient spatial d'intensité est directement calculable. Sur ce sujet, nous renvoyons le lecteur à la présentation qui en est faite par exemple dans [Fablet 01, Sec. 1.1.2]. Nous reviendrons brièvement sur ce point, pour l'une des techniques utilisées dans nos travaux, dans le paragraphe 8.2.2.1.

### 7.1.2 Cadrage de nos objectifs

Notre objectif général peut être formulé comme suit : caractériser le contenu de séquences audiovisuelles en fonction de leur contenu dynamique. Ceci implique notamment quelques contraintes sur la globalité de l'information analysée. Il s'agit pour nous de rendre compte d'une activité sur l'ensemble de l'image et sur un intervalle temporel donné. Nous nous positionnons donc dans un contexte de caractérisation d'*activité* au sens de A. Bobick. En particulier, notre cadre d'étude n'englobe pas *a priori* l'étude des mouvements de caméra, la segmentation spatiale au sens du mouvement, l'extraction, la modélisation et le suivi d'objets, l'analyse des trajectoires.

Par conséquent, s'il paraît approprié de partir d'une description de l'apparence locale du mouvement, la classification devra concerner *in fine* une information globale liée à l'activité sur l'ensemble de la séquence d'images considérée.

Il convient, de plus, de préciser ce que nous entendons par *séquence* et l'horizon temporel associé. Dans la partie II, le terme *séquence* était synonyme de *macro-segment*. Il est vrai que nous aurions aimé nous situer directement au niveau des macro-segments (de quelques minutes à plusieurs dizaines de minutes), voire éventuellement au niveau des plans (de quelques secondes à quelques minutes) au sein de ceux-ci. Toutefois, compte tenu de l'état de l'art des méthodes d'analyse du mouvement et de la variabilité du mouvement au sein de ces segments, nous avons souhaité, dans un premier temps, nous situer à un niveau de granularité temporelle nous assurant de l'homogénéité interne des segments considérés. L'approche générale proposée est donc, à partir d'un découpage en plans ou en macro-segments, de sur-segmenter ces unités temporelles, afin d'extraire du flux vidéo des "blocs temporels" d'une quinzaine d'images, supposés homogènes. Cette approche s'inspire de celle proposée dans le domaine du traitement du son, lorsque le flux audio est pré-segmenté en *trames*, avant d'extraire un certain nombre de descripteurs spécifiques [Gauvain 98]. Notre objectif est alors de caractériser ces "blocs temporels". Toutefois, pour des raisons de cohérence avec la plupart des articles cités et traitant de ces sujets, le terme *séquence* renverra donc, dans cette partie, à une succession temporelle d'images et non à la notion de macro-segment.

De plus, nous n'avons pas voulu faire d'hypothèses fortes sur la nature du mouvement (rigide, articulé, déformable, etc.), dans la mesure où tous les types de mouvements peuvent être rencontrés dans les documents audiovisuels traités.

Enfin, même si nous souhaitons nous placer dans un cadre d'étude et une stratégie d'indexation généraux, nous avons appliqué nos travaux à un usage identifié à l'INA et aux corpus disponibles, en nous intéressant plus particulièrement à la caractérisation de séquences dans des documents audiovisuels sportifs. Ces documents, comme la plupart des documents audiovisuels issus du fonds documentaire de l'INA, sont filmés à caméra mobile.

## 7.2 État de l'art

Ces observations préliminaires nous ont guidés dans la présentation de l'état de l'art, pour lequel nous avons suivi le cheminement suivant :

- présentation de quelques méthodes liées à la caractérisation de séquences. Nous verrons que ces méthodes ne sont pas toujours fondées uniquement sur l'information de mouvement, mais les objectifs poursuivis par ces travaux sont similaires aux nôtres ;
- présentation de quelques méthodes de caractérisation du mouvement. Les auteurs se situent alors souvent dans d'autres domaines d'application, mais proposent des méthodes potentiellement utilisables dans notre contexte d'étude ;
- présentation d'algorithmes traitant de l'indexation de documents sportifs. Une revue succincte des différents travaux dans ce domaine d'application spécifique sera faite.

### 7.2.1 Caractérisation du contenu de séquences audiovisuelles

La caractérisation des séquences dans un contexte d'indexation vidéo est abordée soit avec une visée de classification des segments temporels, soit pour la prise en compte de requêtes par similarité. Dans le projet *MoCA* [Fischer 95], une classification des documents en fonction de leur genre

(*actualités*, *course de voiture*, *tennis*, *dessin animé*, *publicité*) est envisagée. Une répartition des documents de fiction en deux classes (*action* ou *romance*) est étudiée dans [Vasconcelos 97c]. Cette idée est reprise dans [Iyengar 98], avec des labels légèrement différents, le choix des classes devient soit *action* ou *personnage*, soit *actualités* ou *sport*. Certaines tâches de classification s'apparentent plus à une détection d'événements, comme la description symbolique de séquences de vidéo-surveillance [Neumann 83], ou la détection des scènes de chasse dans les documentaires animaliers [Qian 99]. Enfin, à mi-chemin entre la caractérisation du contenu des séquences et la détection d'événements, nous pouvons citer les travaux décrits dans [Vasconcelos 98b] qui introduisent les classes *action*, *gros plan*, *foule* et *décor*, et ceux dans [Naphade 98] qui retiennent les classes *explosion* et *chute d'eau*. Une interface graphique de requête sur des séquences est développée dans [Ioka 92], ainsi que dans le projet *VideoQ* [Chang 98]. Des méthodes de requêtes par l'exemple sur des séquences sont proposées dans [Ardizzone 96,Jain 99,Naphade 00b,Fablet 01].

Les primitives extraites des séquences d'images sont généralement de bas niveau et ne comprennent pas que des descripteurs de mouvement. Ainsi, la couleur est utilisée dans [Fischer 95, Naphade 00b] (histogrammes et variance de la distribution des couleurs) ou dans [Vasconcelos 98b] (détection des zones de couleur peau). D'autres descripteurs sont extraits : dans [Fischer 95,Naphade 98] il s'agit de statistiques sur le signal audio, dans [Vasconcelos 97c] de la durée des plans, dans [Vasconcelos 98b,Jain 99,Qian 99] des descripteurs de texture, et dans [Chang 98] des descripteurs de contour ainsi que des informations liées aux caractéristiques temporelles des objets. Lorsque le mouvement est pris en compte, les auteurs utilisent généralement le flot optique, [Fischer 95], ou des primitives qui en sont issues. Le descripteur d'activité est défini comme la variance des composantes du flot optique dans [Vinod 98], ou comme l'énergie du flot optique dans [Iyengar 98]. L'activité est définie par différences entre images successives dans [Vasconcelos 97c] ; c'est également le cas dans [Vasconcelos 98b] après alignement des images. La moyenne des vecteurs de vitesses sur quatre régions et un histogramme normalisé des amplitudes sont extraits du flot optique dans [Ardizzone 96]. Les histogrammes des amplitudes selon les axes verticaux et horizontaux sont calculés à partir du flot optique dans [Jain 99]. Une représentation en trois dimensions des vecteurs de mouvement par blocs est exploitée dans [Ioka 92]. Dans [Fablet 01], les descripteurs utilisés sont fondés sur des cooccurrences de quantités locales de mouvement. Notons que dans [Neumann 83], le niveau symbolique élevé des requêtes en langage naturel considérées ne permet pas d'extraction automatique de l'information nécessaire.

Les méthodes citées diffèrent aussi lorsqu'il s'agit de globaliser l'information utilisée au niveau des séquences. Certains procèdent de manière incrémentale, d'autres réduisent l'information et fournissent des signatures de séquences, d'autres enfin utilisent des connaissances spécifiques au domaine pour organiser l'information[6]. Une stratégie incrémentale est utilisée dans [Fischer 95] où sont extraites, à partir des primitives de bas niveau, des informations telles que la segmentation en plans, la localisation d'images de couleur uniforme, les mouvement de caméra, la segmentation parole/musique/bruit/silence. Les mouvements de caméra sont aussi repérés dans [Ardizzone 96,Chang 98], ainsi que le mouvement dominant et la segmentation en plans dans [Qian 99]. Lorsqu'une signature est calculée au niveau de la séquence, ce peut être par concaténation des données [Naphade 00b], par moyennage [Vasconcelos 97c], par l'utilisation d'histogrammes [Vinod 98], par regroupement en classes [Ioka 92], par analyse en composantes principales (ACP) [Iyengar 98]. Parfois, des primitives calculées autour d'images-clefs deviennent celles de la séquence. Enfin, dans [Fablet 01], les cooccurrences temporelles extraites sont décrites soit par des attributs globaux (certains des descripteurs d'Haralick [Haralick 73]), soit par des modèles probabilistes, en l'occurrence des modèles de Gibbs causaux. Par ailleurs, l'utilisation d'information *a priori* permet

6. Dans ces deux derniers cas, on retrouve les approches *Activité* et *Action* décrites par A. Bobick.

la reconnaissance d'objets (logo) [Fischer 95], ou la labellisation de régions de l'image avec des réseaux de neurones [Qian 99].

En ce qui concerne la caractérisation des séquences à proprement parler, les méthodes mises en œuvre peuvent inclure des modèles et des techniques de classification floue [Fischer 95], un réseau bayésien [Vasconcelos 98b], des chaînes de Markov cachées (HMM) [Iyengar 98]. Les HMM sont également utilisés dans une version hiérarchique dans [Naphade 98], où est aussi envisagée une alternative *via* des mélanges de lois gaussiennes. Des informations spécifiques au domaine permettent la détection de scènes de chasse grâce à des modèles [Qian 99]. Lorsque l'application envisagée est la recherche par l'exemple à l'aide d'une mesure de similarité une distance est généralement proposée sur les primitives [Ioka 92,Ardizzone 96,Vinod 98]. Lorsque les primitives sont de nature différente, les distances correspondantes sont souvent normalisées et les auteurs utilisent une distance pondérée [Chang 98,Jain 99]. Dans [Fablet 01], la recherche d'exemples similaires est formulée dans un cadre bayésien à l'aide du critère du maximum *a posteriori* (MAP).

Citons également les travaux de [Black 98] où est construit un modèle des changements d'apparence dans une séquence vidéo, à l'aide de mélanges de lois. La méthode développée permet de différencier quatre sources de changements d'intensité : l'évolution de la forme des objets, les variations d'illumination, la spécularité, les changements iconiques. Les séquences sont alors caractérisées par la cause des changements d'intensité.

À la lecture de ces travaux, nous observons que, bien souvent, l'information globalisée sur la durée de la séquence est extraite soit par concaténation soit par moyennage, à moins que ce ne soit la signature des images représentatives de la séquence qui soit conservée. Une alternative est proposée dans [Fablet 01], où les signatures fondées sur des cooccurrences temporelles rendent compte, par construction, du contenu dynamique de l'ensemble de la séquence. Ce type de caractérisation du mouvement sur la séquence entière constitue une piste qui remplit les contraintes définies à la sous-section 7.1.2.

### 7.2.2 Autres méthodes pour la caractérisation du contenu dynamique

Les domaines de la vidéo-surveillance et de l'interaction homme-machine, dans la mesure où l'objet est de caractériser des activités humaines, ont aussi suscité des études sur la caractérisation du mouvement. Dans ces domaines d'application, les méthodes sont notamment dédiées à la reconnaissance d'expressions faciales, de gestes et d'activités humaines. Des états de l'art assez complets ont été publiés récemment [Aggarwal 99,Moeslund 01].

Ces domaines d'application privilégient souvent l'intégration d'informations contextuelles pour interpréter l'activité perçue dans les séquences d'images. Celles-ci sont filmées à caméra fixe, ce qui permet la mise en œuvre, avec une difficulté moindre, de méthodes fondées sur de fortes modélisations, ou sur l'apparence globale des objets extraits. Ainsi, pour la reconnaissance d'expressions faciales, de gestes ou d'activités humaines, de nombreux auteurs [Clergue 95,Black 95,Ju 96] se sont fondés sur des modèles plus ou moins précis de leur objet d'étude : un modèle du corps humain ou du visage en trois dimensions constitué de régions et de leurs mouvements relatifs possibles, [Clergue 95], un modèle plan du visage et des modèles paramétriques pour les yeux, la bouche et les sourcils, relatifs au mouvement de la tête [Black 95]. Cette stratégie est étendue aux mouvements du corps humain *via* un modèle 2D dans [Ju 96]. La structure suivie est généralement initialisée à la main, puis l'estimation du flot optique permet la mise à jour du modèle [Black 95,Ju 96]. La mise en correspondance du modèle avec des régions définies par des paramètres caractéristiques peut aussi être assurée au cours du temps par un certain nombre de contraintes spatiales et temporelles reposant sur des *scénarios*. Dans [Yacoob 96], c'est le flot optique qui détermine le suivi de

l'évolution de points particuliers (points de fort gradient d'intensité) et de zones d'intérêt (définies notamment autour des yeux) afin d'analyser des expressions faciales. Dans un contexte légèrement différent, dans [Wilson 98] il est proposé d'estimer l'amplitude ou la direction de certains gestes à l'aide de chaînes de Markov cachées (HMM).

Des méthodes liées aux textures temporelles ont été utilisées par plusieurs auteurs. Dans [Nelson 92], il est suggéré l'utilisation de statistiques du premier et du second ordre sur le flot optique ainsi que de matrices de cooccurrence calculées sur ce dernier. Cette approche, appliquée initialement pour différencier des mouvements rigides simples et des mouvements fluides (feuillages, rivières, etc.), a été reprise et généralisée dans [Fablet 01]. L'efficacité de la méthode statistique développée dans [Fablet 01] a été démontrée pour la reconnaissance et la segmentation du mouvement, ainsi que pour la satisfaction de requêtes par l'exemple, globales ou partielles. Les textures temporelles sont modélisées dans [Szummer 96], par des modèles auto-regressifs spatio-temporels, mais il n'y a pas eu à notre connaissance d'application de cette technique à la reconnaissance d'activités. Dans [Heeger 88], a été introduite une méthode de mesure de mouvement par une famille de filtres de Gabor spatio-temporels, qui a été reprise dans [Chomat 00] pour la reconnaissance d'activités humaines (*descendre* ou *monter* un escalier, *marcher* vers la gauche ou vers la droite, s'*avancer* vers ou s'*éloigner* de la caméra). Les réponses des différents filtres en chaque point de l'image sont stockées dans un histogramme multi-dimensionnel rendant compte de l'activité présente dans la scène observée [Chomat 99a].

L'équipe du laboratoire Multimédia du MIT, animée par A. Bobick, a utilisé dans de nombreux travaux une signature évaluée sur la durée d'une séquence et en chaque point de l'image, et traduisant l'information du dernier instant de présence d'un mouvement [Bobick 01]. Cette signature appelée "image de l'historique du mouvement" (*Motion History Image* - MHI) se fonde sur un seuillage des différences d'images successives dans la séquence. Elle a été appliquée à des domaines variés : classification de mouvements d'aérobic [Davis 97], analyse des gestes d'un chef d'orchestre [Bradski 00], interaction homme-machine notamment dans le projet *Kidsroom*, utilisation de l'imagerie infra-rouge pour des environnements interactifs virtuels [Davis 98].

D'autres primitives, plus ou moins spécifiques ou coûteuses en calculs sont extraites au niveau des séquences, dans le cadre de la reconnaissance d'expressions faciales, de gestes ou d'activités humaines. Dans [Ju 98], la reconnaissance de gestes pour l'annotation automatique de présentations orales s'appuie sur l'analyse des déformations des contours actifs. Dans [Bobick 98], des coefficients de vraisemblance sont calculés à partir de données de bas niveaux et d'une série de HMM, appris chacun sur un geste spécifique élémentaire. La succession temporelle des états constitue un ensemble de primitives permettant de caractériser les séquences d'action. Plus généralement, les données brutes issues des images sont souvent réduites sur une base adaptée de vecteurs propres, souvent par analyse en composantes principales (ACP). C'est notamment le cas dans [Wilson 96], où l'auteur analyse des corrélations internes à certains gestes (dit *bi-phasiques* ou *multi-phasiques*), et dans [Cui 95], pour des travaux sur le langage des signes, où des "fovea vectors" sont issus de la normalisation spatio-temporelle, de la réduction et de la concaténation de données brutes avec l'indication du mouvement global. D'autres auteurs proposent de moyenner temporellement les images [Wilson 95], ou d'extraire, en plus des différences entre images, une quantification vectorielle de données issues de découpages verticaux et horizontaux des images compressées du flux MPEG [Rigoll 96]. Enfin, des ondelettes de Haar concaténées temporellement sont utilisées pour détecter et localiser des piétons en mouvement dans des séquences d'images [Papageorgiou 99].

Pour effectuer la classification du contenu dynamique des séquences, le recours à des modèles est relativement courant. Les auteurs se réfèrent à une grammaire de "gestes atomiques" pour la reconnaissance de geste [Bobick 98], ou à un dictionnaire pour les expressions de visages [Yacoob 96].

Une autre approche est la construction d'un modèle par apprentissage [Kurita 97,Yacoob 99, Bobick 01]. Les HMM et les réseaux de neurones sont des solutions classiques [Rigoll 96]. Les HMM sont utilisés dans [Wilson 95] pour retrouver un même geste sous différents points de vue. Les machines à vecteurs de support (*Support Vector Machines* - SVM) sont une alternative employée par quelques auteurs [Pittore 00,Papageorgiou 99]. Un cadre de reconnaissance bayésien est exploité dans [Chomat 99a]. L'auteur crée des modèles d'apparence pour chaque type d'activité. Chaque point de l'image "vote" pour un des modèles et l'ensemble des votes des points d'une image détermine l'action reconnue.

De cet état de l'art, nous retenons l'utilisation des MHI et des méthodes associées aux textures temporelles qui semblent à même d'obtenir une information du mouvement global sur la séquence. Enfin, les méthodes de classification par apprentissage semblent proposer une intégration de l'information contextuelle sans passer par des modélisations trop coûteuses et trop restrictives.

### 7.2.3 Analyse de documents audiovisuels traitant de thématiques sportives

La mise en image de compétitions sportives, et notamment des sports d'équipes, a donné lieu à de nombreux travaux fondés sur une modélisation des connaissances spécifiques aux domaines. Ces travaux [Intille 94,Gong 95,Yow 95,Miyamori 98] nécessitent généralement une modélisation *a priori* assez complète du terrain, de la balle, voire des joueurs. Un certain nombre de primitives sont extraites des images : objets (ballons, joueurs, lignes de marquage), mouvements de caméra, etc. Celles-ci sont ensuite exploitées au sein du modèle. Un suivi des joueurs lors de matchs de football américain et la possibilité de comparer différentes actions sont appréhendés dans [Intille 94], pour un usage professionnel. Le repérage de certains événements caractéristiques d'un match de football a suscité de nombreux travaux : repérage des *corners*, *tirs au but*, etc. dans [Gong 95], repérage des actions comme des tirs au but marqués ou manqués, ou des résumés d'action sous forme de mosaïques dans [Yow 95]. Dans [Miyamori 98], l'utilisation de HMM permet de visualiser les principales actions d'un match de football sous forme de script. Dans le même ordre d'idée, dans [Chang 96], la détection des événements notables d'un match de football américain a été réalisée avec des techniques proches de celles évoquées à propos de la segmentation temporelle de journaux télévisés (cf. sous-section 4.2.3). L'auteur se fonde sur l'analyse de la bande sonore pour repérer les exclamations de joie du public, et sur un certain nombre de mots-clefs prédéfinis et pré-appris selon une méthode de localisation de mots (*word spotting*). Autour des localisations temporelles de ces événements sonores détectés, les images sont analysées (extraction du marquage au sol, du nombre de joueurs, etc.) et les primitives obtenues sont mises en correspondance avec un modèle de l'événement correspondant au mot-clef trouvé. Notons enfin le travail original décrit dans [Taki 98], qui analyse les mouvements de groupe de joueurs de football grâce à deux primitives : les "régions de dominance" associées aux joueurs puis aux équipes en présence, et un "motif de temps minimal" (*shortest time pattern*) indiquant pour une position donnée de la balle quel joueur (et quelle équipe) en est le plus proche. La première des primitives permet une analyse temporelle des phases de jeu (offensives ou défensives), la seconde permettrait une évaluation des passes effectuées au cours de la partie.

D'autres travaux sur la classification de séquences sportives s'avèrent plus proches de nos préoccupations. Dans [Mohan 98], l'auteur s'intéresse à la détection des événements et de situations sportives particuliers comme un plongeon dans les séquence de natation, un "hit" dans les séquences de baseball, ou plus généralement un ralenti dans les séquences de sport. Pour ce faire, l'auteur utilise une mesure ordinale des images d'une séquence donnée, c'est-à-dire la valeur ordinale des intensités pour les trois canaux de couleurs sur des images réduites. Les descripteurs extraits sont

ensuite concaténés temporellement, et mis en correspondance avec des modèles. Cette méthode permet d'obtenir la localisation temporelle de séquences similaires à la séquence recherchée. La méthode proposée dans [Sahouria 98] vise la classification de séquences selon les types de sport (*basket-ball*, *hockey sur glace*, *volley-ball*). Les vecteurs de mouvement sont extraits du flux MPEG sur les images de type P, et sont ensuite projetés sur une base de vecteurs propres spécifiques au problème considéré. Deux méthodes de classification sont expérimentées : la première utilise des HMM, la seconde une quantification vectorielle (VQ). Une autre méthode pour détecter les ralentis dans des documents de sports est présentée dans [Kobla 99]. Les informations utilisées sont liées au flux MPEG, comme par exemple le type des macroblocs dans les images B ou l'amplitude dominante du champs des vecteurs de mouvement. Ces informations servent notamment à déterminer les images en mouvement et les images immobiles (*shift and still frames*). Les alternances d'une image en mouvement et d'une ou plusieurs images immobiles sont interprétés comme des ralentis. Enfin, dans [Nakano 00] est décrite une méthode permettant de détecter trois actions classiques d'un match de volley-ball (*service*, *réception* et *smash*). La réponse de quatre filtres de Gabor spatio-temporels est calculée sur la zone correspondant au joueur en quinze points prédéfinis. Ces soixante coefficients sont utilisés pour calculer une mesure de corrélation avec des modèles des actions, permettant ainsi la classification des séquences.

Dans le cadre de nos travaux, nous nous situons davantage dans une approche visant à une caractérisation des types de séquences selon les sports, voire la détection des événements notables lors d'une compétition sportive. Nous ne retiendrons pas l'utilisation des vecteurs de mouvement présents dans le flux MPEG. En effet, des expérimentations menées au sein du GRAMM de l'INA ont montré que l'encodage des de vecteurs mouvement pour les documents numérisés auxquels nous avons accès n'était pas assez fiable.

## 7.3 Motivation du choix des algorithmes mis en œuvre

La présentation de ces différents travaux à la lumière de nos contraintes et de nos intérêts va nous permettre de resteindre et de motiver nos choix lors de l'extraction des primitives et de la caractérisation des séquences. La présentation et la mise en œuvre de ces techniques seront détaillées au chapitre 8.

Nous cherchons à développer une stratégie générale pour la caractérisation du contenu dynamique de séquences audiovisuelles en vue de leur indexation, et nous sommes confrontés à une grande variété de mouvements possibles. Ainsi, une modélisation forte du mouvement et des informations contextuelles nous a semblé trop restrictive. Par contre, l'extraction d'informations dynamiques locales en chaque point de l'image paraît intéressante dans la mesure où une description locale permet de prendre en compte des mouvements complexes et de nature variée. Nous avons adopté d'une part le calcul des MHI et d'autre part la notion de textures temporelles avec une caractérisation par des filtres de Gabor spatio-temporels. Ces deux primitives - MHI et filtres de Gabor 3D - présentent en outre l'intérêt de décrire le mouvement de manière naturelle sur plusieurs images. Pour une caractérisation globale des séquences, ces primitives peuvent être exploitées telles quelles, mais une réduction de leur représentation doit aussi être envisagée.

Par ailleurs, les techniques de classification par apprentissage présentent à notre avis un double intérêt. La définition d'ensembles d'apprentissage appropriés semble être une manière souple, générique et efficace de modéliser les différentes apparences d'une même classe de mouvements, de prendre en compte de l'information contextuelle, et de limiter une classification à un monde fermé. La méthode des machines à vecteurs de support a plus particulièrement retenu notre attention. Elle semble accepter en entrée des types très variés de données numériques et être en mesure de traiter

des problèmes complexes.

Nous allons être amenés à utiliser ces deux groupes de primitives dans un contexte plus complexe que celui des travaux antérieurs comme ceux décrits dans [Bobick 01] et [Chomat 00]. Nous avons à appréhender des situations dynamiques plus variées avec de plus une caméra qui peut être mobile.

# Chapitre 8

# Caractérisation du contenu dynamique de séquences courtes

Nous nous sommes efforcés, dans le précédent chapitre, de positionner notre étude en regard d'autres travaux et de motiver le choix des descripteurs du mouvement et d'une technique de classification. Nous allons décrire l'approche retenue pour la caractérisation du contenu dynamique d'une séquence. Nous en donnons tout d'abord une vue d'ensemble, puis les différents modules sont détaillés dans les sections suivantes.

## 8.1 Description globale de la méthode de caractérisation

Notre méthode de caractérisation du contenu dynamique d'une séquence comprend deux modules principaux : l'extraction des primitives de mouvement et la classification de l'activité captée. Un bloc diagramme est donné à la figure 8.1.

L'extraction des deux groupes de primitives retenues, les images de l'historique du mouvement (MHI) et les réponses des filtres de Gabor spatio-temporels, sont détaillées respectivement dans les sous-sections 8.2.1 et 8.2.2.

Concernant les MHI, diverses techniques de globalisation et de structuration spatiale ou temporelle de l'information sont proposées par des matrices de cooccurrence, ou par les descripteurs d'Haralick. Une réduction de l'information par transformée discrète en cosinus (DCT) est aussi envisagée. Par ailleurs, un module de compensation du mouvement dominant estimé permettant le recalage des images est appliqué en entrée des MHI afin d'éliminer le mouvement de la caméra .

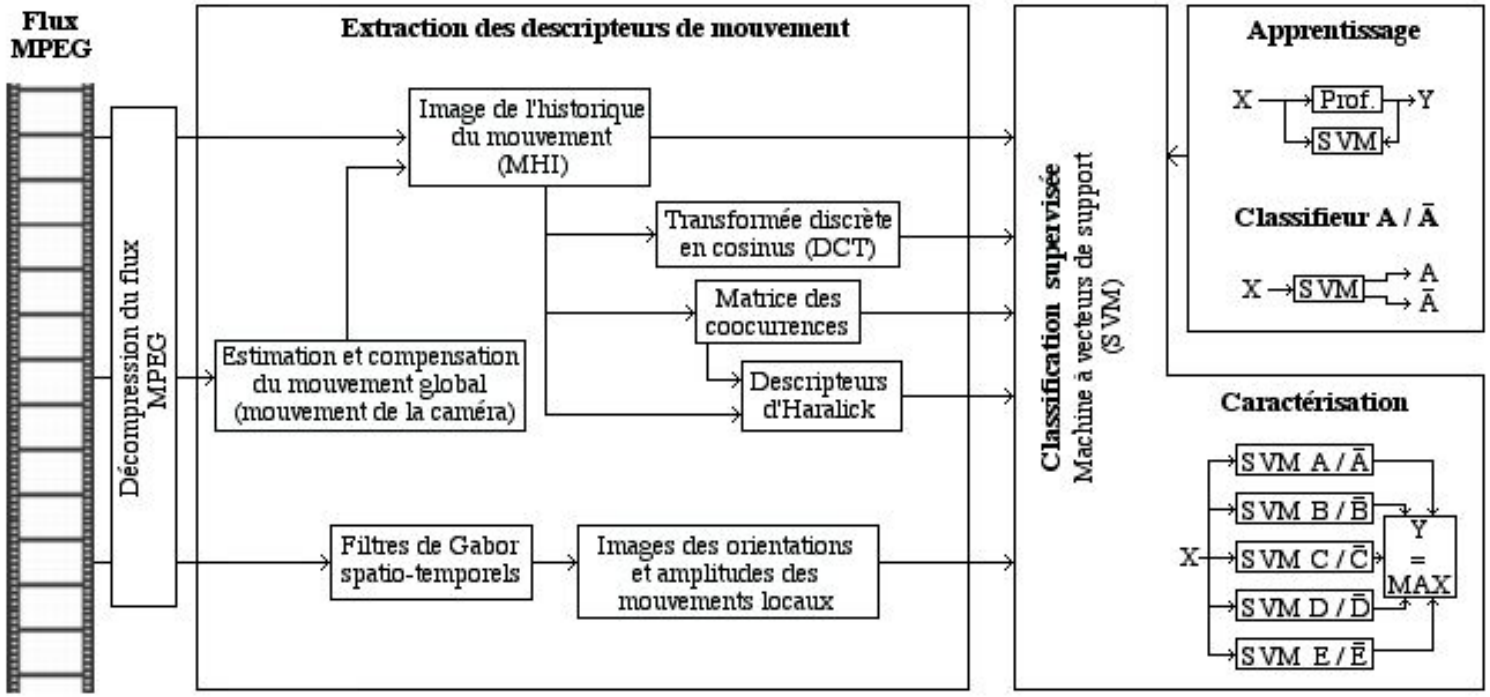
La technique de classification supervisée employée est l'objet de la section 8.3. Soulignons qu'elle se décompose en deux phases : l'apprentissage de classifieurs à deux classes et l'utilisation conjointe de ces classifieurs pour la caractérisation des primitives extraites.

## 8.2 Extraction des primitives du mouvement

### 8.2.1 Utilisation d'images de l'historique du mouvement

#### *8.2.1.1 Définition des images de l'historique du mouvement*

Les images de l'historique du mouvement (*Motion History Image* - MHI) ont été proposées par A. Bobick et son équipe [Bobick 01] sous différentes variantes. Deux signatures sensiblement équivalentes ont été tout d'abord suggérées et ont été notées respectivement BMR (*Binary Motion*




TAB. 8.1: *Bloc-diagramme de la méthode de caractérisation du contenu dynamique d'une séquence comprenant deux modules principaux, le premier concerne l'extraction des descripteurs du mouvement et le second la classification des séquences*

*Region*) [Davis 96] et MEI (*Motion Energy Image*) [Bobick 96]. La signature BMR est formée simplement de la réunion temporelle des cartes binaires des zones mobiles détectées dans les images d'une séquence, tandis que la signature MEI résulte d'une sommation temporelle des cartes binaires des zones en mouvement. Dans le cas d'une caméra fixe, les valeurs non nulles de ces deux signatures forment l'aire balayée par l'objet mobile lors de son déplacement dans la séquence considérée.

La MHI a été définie dans [Davis 96,Davis 97], et est formalisée de manière récursive comme suit :
$\forall (x, y, n) \in \{0, \ldots, W-1\} \times \{0, \ldots, H-1\} \times \{n_i + 1, \ldots, n_f\}$,

$$MHI(x, y, n) = \begin{cases} V_{max} & \text{si } |I_n(x, y) - I_{n-1}(x, y)| \geq \tau \\ \max(0, MHI(x, y, n-1) - 1) & \text{sinon} \end{cases} \quad (8.1)$$

où $V_{max} \in \mathbb{N}$ est une amplitude donnée, $I_n$ est l'intensité au point de coordonnées $(x, y)$ de l'image $\mathcal{I}_n$ , $W$, $H$ et $n$ sont respectivement la largeur, la hauteur et l'indice temporel des images de la séquence $\mathcal{S} = [\mathcal{I}_{n_i}, \ldots, \mathcal{I}_{n_f}]$.

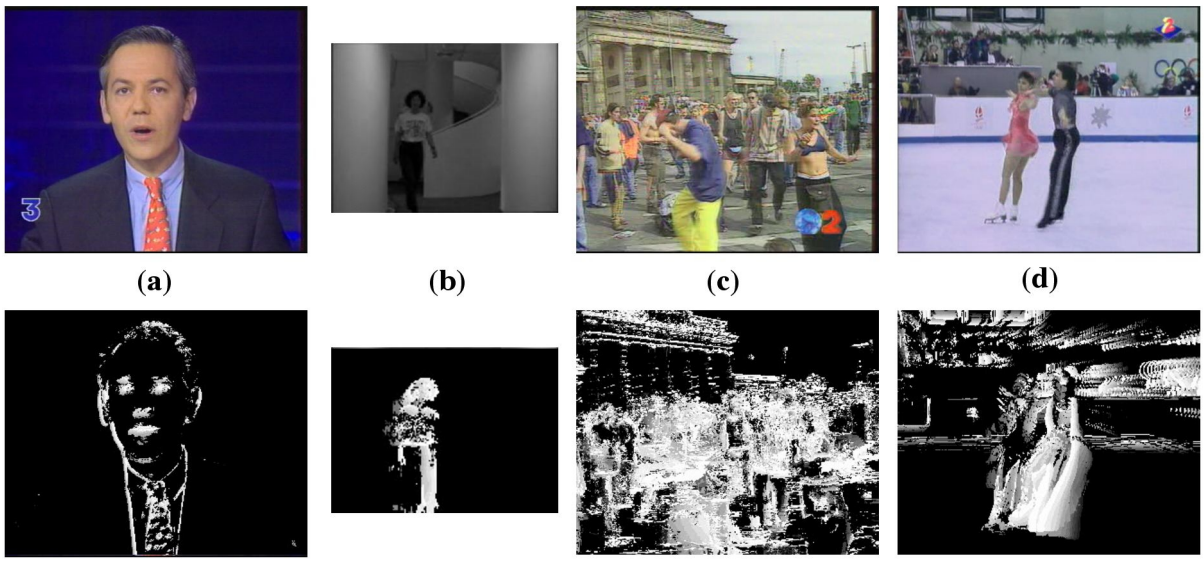



FIG. 8.1: *Représentation de signatures $Map_{MHI}$ : des images représentatives des séquences originales sont regroupées sur la rangée du haut et les MHI correspondantes sur la rangée du bas. Les séquences (a) et (b) sont filmées à caméra fixe. Les paramètres utilisés dans le calcul sont ceux indiqués au paragraphe 8.4.1.*

Des exemples de MHI calculées sur différentes séquences sont montrés à la figure 8.1. Si, dans le cas de la MEI, les points de valeurs maximales sont ceux qui ont le plus souvent bougé dans la séquence considérée, la MHI donne la priorité au dernier mouvement observé en un point. Ainsi, lorsqu'un point est détecté en mouvement à l'instant $n$, toute information sur le mouvement en ce point aux temps antérieurs est effacée. La MHI donne en chaque point l'indication du dernier instant où il y a eu mouvement. Notons enfin que d'après la relation 8.1, toute information de mouvement dont l'éloignement dans le temps est supérieur à $V_{max}$ n'est plus prise en compte.

Initialement, ces signatures ont été utilisées dans le cadre de la modélisation paramétrique du mouvement [Davis 96,Bobick 96], mais rapidement la description du mouvement s'est appuyée

sur des informations extraites des MHI. Ainsi, les MHI exhibant généralement, à caméra fixe, des formes compactes, l'utilisation des moments de Hu a été suggérée [Davis 97]. Dans [Davis 99a], une analyse des gradients de la MHI, calculés à l'aide de masques de Sobel, est menée. En chaque point de l'image, une direction locale du mouvement est ainsi déterminée, et des histogrammes localisés sont construits afin de décrire l'information directionnelle du mouvement dans la séquence d'images. Dans [Bradski 00], l'information d'orientation locale est intégrée par sommation pondérée et permet le calcul de directions moyennes du mouvement sur certaines régions de l'image. Enfin, une variante de la MHI est proposée dans [Bradski 00] afin de permettre une certaine normalisation temporelle de la signature. La tMHI (*timed Motion History Image*) est une signature relativement indépendante du nombre d'images par seconde dans le flux vidéo et de la durée de réalisation d'un mouvement.

Une implantation de ces techniques a été réalisée dans la librairie *CVLib* d'*Intel* [Davis 99b]. Nous avons cependant été amenés à effectuer notre propre mise en œuvre du calcul des MHI.

Nous obtenons ainsi un premier groupe de descripteurs de mouvement pour une séquence $\mathcal{S}$, à savoir le vecteur, noté $Map_{MHI}$, défini par :

$$Map_{MHI}(\mathcal{S}) = \{MHI(0,0,n_f), \ldots, MHI(x,y,n_f), \ldots, MHI(W-1,H-1,n_f)\} \tag{8.2}$$

avec $(x,y) \in \{0,\ldots,W-1\} \times \{0,\ldots,H-1\}$. Les coefficients de la MHI sont ainsi rangés dans l'ordre lexico-graphique. Les valeurs utilisées pour le paramètre $\tau$ seront précisées à la sous-section 8.4.1.

#### *8.2.1.2 Réduction de dimension par transformation discrète en cosinus*

La réduction de la taille du descripteur $Map_{MHI}$ peut avoir deux motivations. Cette signature devant être utilisée comme entrée des classifieurs, il est souhaitable, pour des raisons de vitesse de traitement et d'efficacité, d'obtenir des signatures de plus petite taille, tout en conservant l'information pertinente. Cet argument est néanmoins relativisé par le choix de la méthode de classification. En effet, les machines à vecteurs de support (SVM) sont réputées être peu sensibles à la dimension de l'espace des données d'entrée (voir sous-section 8.3.1). La seconde motivation est que des signatures de taille réduite sont plus aisément manipulables et permettent une économie de stockage. Il nous a donc paru pertinent d'expérimenter une réduction de la dimension des MHI qui, par construction, sont de la taille des images traitées.

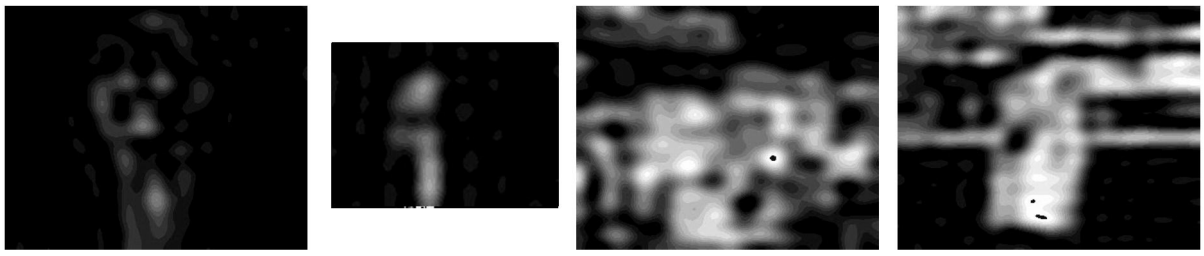


FIG. 8.2: *Représentation de signatures $DCT_{MHI}$ obtenues sur les séquences présentées à la figure 8.1. Les paramètres sont ceux indiqués au paragraphe 8.4.1.*

Deux méthodes classiques, permettant la description du signal sur une base réduite et sa reconstruction avec une dégradation limitée, ont été appliquées : la transformée discrète en cosinus

(DCT) et la transformation de Karhunen-Loeve (KLT). La KLT s'est avérée complexe à utiliser puisqu'elle nécessite un apprentissage supplémentaire. Nous avons donc conservé la DCT, facile à mettre en œuvre et utilisée notamment pour les compressions JPEG et MPEG. Nous avons utilisé la version fournie dans la librairie *IPLib* d'*Intel*[1]. Les coefficients obtenus par transformée discrète en cosinus d'une image $\mathcal{I}$ de taille $W \times H$ sont donnés par la formule suivante :
$\forall (i,j) \in \{0,\dots,W-1\} \times \{0,\dots,H-1\}$,

$$c_{dct}(i,j) = \frac{1}{\sqrt{2W \times H}} \bar{\delta}_0(i)\bar{\delta}_0(j) \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} I(x,y) \cos\left(\frac{(2x+1)i\pi}{2W \times H}\right) \cos\left(\frac{(2y+1)j\pi}{2W \times H}\right) \tag{8.3}$$

avec $\bar{\delta}_0(k) = (1 - \delta_{0,k}) + \delta_{0,k}\frac{1}{\sqrt{2}}$.

En supposant les coefficients $c_{dct}(i,j)$, issus de la transformation de $Map_{MHI}$, lus en *zigzag* et non dans l'ordre lexico-graphique et renumérotés $c_{dct}(k)$, la signature retenue est alors le vecteur, noté $DCT_{MHI}$, des $p$ premiers coefficients obtenus, soit $DCT_{MHI}(\mathcal{S}) = \{c_{dct}(0),\dots,c_{dct}(p-1)\}$. Le choix de $p$ sera abordé à la sous-section 8.4.1. Une visualisation de cette signature est possible en appliquant la transformée inverse sur les $W \times H$ coefficients $\{c_{dct}(0),\dots,c_{dct}(p-1),0,0,0,\dots\}$. Des exemples sont regroupés à la figure 8.2.

#### *8.2.1.3 Matrices de cooccurrence et descripteurs d'Haralick*

Les matrices de cooccurrence ont été utilisées pour décrire des textures spatiales [Aksoy 98] ou temporelles [Nelson 92]. Le calcul des cooccurrences temporelles sur l'ensemble $\{MHI(\mathcal{I}_{n_i}),\dots,MHI(\mathcal{I}_{n_f})\}$, donnant la matrice $Cooc_{MHI}$, nous a paru être une alternative intéressante à la signature $Map_{MHI}$. En effet, cette dernière contient une information localisée spatialement, mais temporellement réduite aux valeurs de $MHI(\mathcal{I}_{n_f})$, tandis que la signature $Cooc_{MHI}$ rend compte de l'évolution temporelle des valeurs des MHI au cours du temps, mais perd par contre la localisation spatiale de l'information.

Plus formellement, nous avons $Cooc_{MHI}(\mathcal{S}) = [v(k,j)]$ avec $(k,j) \in \{0,\dots,V_{max}\}^2$ définie par la formule suivante :

$$\begin{aligned} v(k,j) = \; & \tfrac{1}{(n_f - n_i) \times W \times H} Card\big\{\big((x,y,n-1),(x,y,n)\big) \in (\{0,\dots,W-1\} \times \{0,\dots,H-1\} \\ & \times \{n_i+1,\dots,n_f\})^2 / MHI(x,y,n-1) = k \text{ et } MHI(x,y,n) = j\big\} \end{aligned} \tag{8.4}$$

Compte tenu de la définition des MHI donnée à la relation 8.1, la matrice de cooccurrence contient de nombreuses valeurs nulles. Seuls les coefficients $v(k,V_{max})$, avec $k \in \{0,\dots,V_{max}\}$, et $v(k,k-1)$ avec $k \in \{1,\dots,V_{max}\}$ sont susceptibles d'avoir des valeurs non nulles. La signature $Cooc_{MHI}$ est donc condensée en un vecteur de taille $2V_{max}+1$.

R. Haralick a défini quatorze descripteurs globaux $\{f_1,\dots,f_{14}\}$ calculés à partir des matrices de cooccurrence dans [Haralick 73, p. 619]. Ceux-ci ont été utilisés notamment dans de nombreux travaux en analyse de texture [Gotlieb 90,Vehel 00,Fablet 01]. Nous avons choisi de retenir les onze premiers, calculés sur la matrice de cooccurrence sous sa forme non réduite, et regroupés au sein de la signature $H_{Cooc_{MHI}} = \{f_1,\dots,f_{11}\}$. La définition de ces descripteurs est rappelée ci-dessous :

- $f_1 = \sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}} \big(v(c,l)\big)^2$

1. disponible sur le site http://developer.intel.com/software/products/perflib/ipl/index.htm

- $f_2 = \sum_{n=0}^{V_{max}-1} n^2 \Big\{ \sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}}{}_{|c-l|=n} v(c,l) \Big\}$
- $f_3 = \frac{\sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}} (cl)v(c,l) - \mu_c \mu_l}{\sigma_c \sigma_l}$
- $f_4 = \sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}} (c-\mu)^2 v(c,l)$
- $f_5 = \sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}} \frac{v(c,l)}{1+(c-l)^2}$
- $f_6 = \sum_{c=2}^{2V_{max}} c \cdot v_{c+l}(c)$
- $f_7 = \sum_{c=2}^{2V_{max}} (c-f_8)^2 v_{c+l}(c)$
- $f_8 = -\sum_{c=2}^{2V_{max}} v_{c+l}(c) \log(v_{c+l}(c))$
- $f_9 = -\sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}} v(c,l) \log(v(c,l))$
- $f_{10}$ = variance de $v_{c-l}$
- $f_{11} = -\sum_{c=0}^{V_{max}-1} v_{c-l}(c) \log(v_{c-l}(c))$

avec :

- $v_c = \frac{1}{V_{max}+1} \sum_{l=0}^{V_{max}} v(c,l)$ ;
- $v_l = \frac{1}{V_{max}+1} \sum_{c=0}^{V_{max}} v(c,l)$ ;
- $v_{c+l}(k) = \sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}}{}_{c+l=k} v(c,l)$ ;
- $v_{c-l}(k) = \sum_{c=0}^{V_{max}} \sum_{l=0}^{V_{max}}{}_{|c-l|=k} v(c,l)$,

et $\mu_c$, $\mu_l$, $\sigma_c$, $\sigma_l$ respectivement les moyennes et les écarts-types de $v_c$ et $v_l$, et $\mu$ la moyenne des $v(k,j)$.

Nous avons aussi opté pour l'évaluation de ces mêmes onze descripteurs directement sur la matrice $Map_{MHI}$ pour former la signature $H_{Map_{MHI}} = \{f_1, \ldots, f_{11}\}$. Cette application des descripteurs d'Haralick à la MHI peut être motivée de la manière suivante.

Nous avons évoqué au paragraphe 8.2.1.1 différentes utilisations des MHI, et notamment le calcul du gradient spatial à l'aide de masques de Sobel et le recours aux moments de Hu. Le calcul du gradient spatial n'est effectué que sur des zones relativement homogènes dans [Davis 99a, Sec. 3.2]. De même, l'extraction des sept moments de Hu n'a de réel intérêt que sur des formes compactes [Hu 62]. Or, contrairement aux travaux de A. Bobick *et al.*, le cadre de nos expérimentations (caméra mobile, nombreux objets indépendants en mouvement à des résolutions variables) ne nous assure aucunement la compacité de nos signatures (voir par exemple figure 8.1). Par conséquent, nous avons cherché d'autres descripteurs globaux des MHI. Il se trouve que certains des descripteurs d'Haralick correspondent à des informations identifiables dans des images. Ainsi, $f_1$ représente l'énergie des intensités et $f_9$ l'entropie. D'autres descripteurs peuvent aussi fournir des indications sur la répartition spatiale des intensités selon les diagonales de l'image ou en fonction de la projection des intensités sur les axes horizontaux et verticaux. Nous avons donc décidé d'évaluer les onze premiers descripteurs d'Haralick calculés sur la dernière MHI de la séquence.

#### *8.2.1.4 Compensation du mouvement global*

Les premières expérimentations ont montré que les MHI obtenues pour des scènes complexes avec mouvement de caméra ne traduisaient pas de structures très évidentes. Nous avons donc cherché à atteindre le mouvement des éléments de la scène en estimant, puis en compensant le mouvement dominant dans l'image supposé dû au mouvement de la caméra. L'objectif est que les MHI rendent compte des mouvements propres de la scène ainsi que d'éliminer le mouvement de la caméra dans des situations où ce dernier n'est pas révélateur de la typologie de la séquence visée.

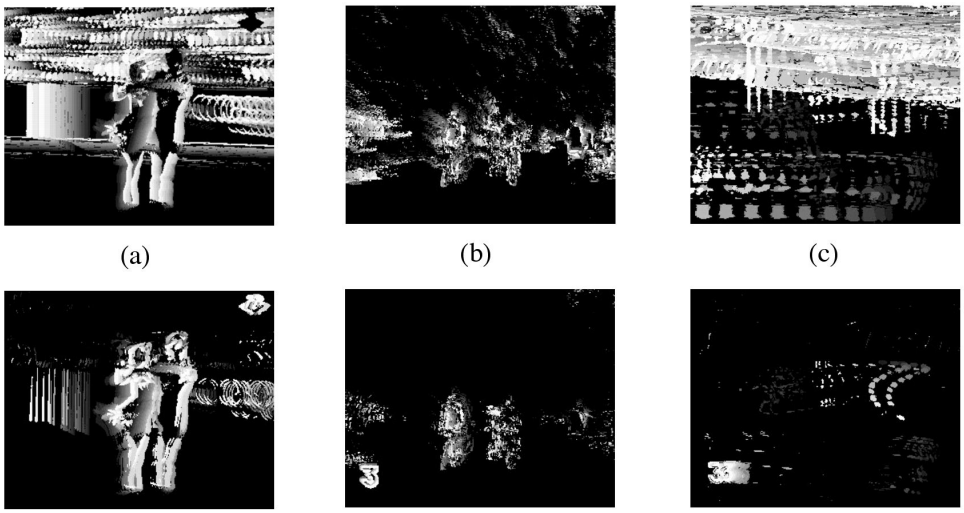



FIG. 8.3: *Effets de la compensation du mouvement dominant sur la signature $Map_{MHI}$. En haut, les MHI incluant les effets des mouvements de caméra, en bas les MHI sur les mêmes séquences dont le mouvement dominant a été compensé à l'aide d'un modèle paramétrique à six paramètres. Les mouvements sont en général complexes dans la mesure où la caméra suit plus ou moins fidèlement le mouvement des sportifs ou l'action retransmise. La compensation semble efficace sur les séquences de patinage (a) et de cyclisme (b). La compensation est plus délicate pour des configurations comme celle de la séquence de football (c).*

Pour ce faire, nous avons eu recours à la méthode d'estimation robuste multirésolution paramétrique décrite dans [Odobez 95] et utilisé la librairie *libMotion-2D* correspondant, développées par le projet VISTA de l'INRIA Rennes. Chaque image $\mathcal{I}_n$ est compensée par rapport à l'image précédente $\mathcal{I}_{n-1}$ à l'aide d'un modèle paramétrique à six paramètres. Les points n'appartenant pas à l'intersection de l'image $\mathcal{I}_{n-1}$ et de l'image $\mathcal{I}_n$ recalée sont fixés artificiellement à une valeur nulle dans la MHI afin d'éviter les effets de bord lors de sa construction.

La compensation du mouvement dominant n'est pas toujours parfaite notamment dans les cas de mouvements complexes de la caméra et d'effets de profondeur significatifs dans la scène, mais la MHI reflète une information plus claire (voir figure 8.3).

### 8.2.2 Utilisation des filtres de Gabor spatio-temporels

Une des limites des MHI est la relative pauvreté de l'information fournie sur le mouvement. En effet, cette signature comprend essentiellement un seuillage des différences d'intensité entre deux images et une datation du dernier changement temporel significatif observé en un point de l'image.

L'intégration temporelle effectuée dans les MHI permet néanmoins de capter spatialement, dans le cas d'un mouvement unique sur un fond fixe, certaines informations sur la nature de ce mouvement à travers l'aire balayée par les éléments en mouvement et le sens de variation des valeurs des MHI associées. Toutefois, dans l'application qui nous intéresse, de nombreux objets en mouvement peuvent être présents sur un fond mobile, ce qui ne permet pas une telle exploitation simple des MHI (voir figure 8.1).

Aussi, nous avons étudié une alternative aux MHI qui puisse fournir une information plus riche sur le mouvement, notamment en termes d'amplitude et de direction. Elle s'appuie sur des filtres de Gabor spatio-temporels, et nous allons la décrire dans les paragraphes suivants.

#### *8.2.2.1 Description des filtres de Gabor spatio-temporels*

Dans [Adelson 85], E. Adelson et J. Bergen ont formulé une appréhension fréquentielle du mouvement apparent dans l'espace spatio-temporel $(x, y, t)$ à trois dimensions que forme une séquence d'images. Un mouvement translationnel correspond ainsi à la focalisation de l'énergie dans un plan particulier de l'espace des fréquences spatio-temporelles, dont l'orientation fournit les caractéristiques de ce mouvement. Les auteurs, s'inspirant du fonctionnement du cortex visuel des primates[2], ont de plus proposé de détecter ces orientations à l'aide d'un ensemble de champs réceptifs locaux.

Deux types d'exploitation de ces champs réceptifs locaux ont été privilégiés. Il s'agit d'une part des méthodes fondées sur des dérivées gaussiennes spatio-temporelles et d'autre part des méthodes énergétiques[3]. Les champs réceptifs peuvent être écrits sous la forme d'un opérateur $F(\vec{p}, \sigma) = A(\vec{p}, \sigma)\phi(\vec{p})$, où $\vec{p}$ indique une position spatio-temporelle et $\sigma$ est le paramètre d'échelle. La fonction $A$ est un noyau gaussien et correspond au support d'estimation locale du mouvement apparent. La fonction $\phi$ est une fonction sinusoïdale.

D. Heeger a utilisé les filtres spatio-temporels de Gabor pour la mesure du flot optique dans des séquences d'images [Heeger 87,Heeger 88]. Dans [Simoncelli 91], il est aussi fait usage de filtres de Gabor dans cette optique, à ceci près que dans [Heeger 88], l'auteur procède à la minimisation de l'écart quadratique des sorties des filtres par rapport aux réponses idéales, tandis que dans [Simoncelli 91] le flot optique est estimé dans un cadre probabiliste. La méthode de D. Heeger a été reprise dans [Spinei 98b] où est effectué un filtrage contrôlé dans les zones où une première estimation du flot optique s'avère imprécise (aux bords des objets en mouvement).

Des primitives ont été définies par combinaison des réponses en énergie des filtres de Gabor pour caractériser six types de mouvement[4] dans [Wildes 00]. Il est possible en théorie de détecter avec un ensemble de filtres de Gabor deux mouvements différents au même endroit, pour peu que les zones en mouvement présentent des fréquences spatiales nettement différenciées [Adelson 86]. Une application à l'estimation des mouvements d'objets transparents a été proposée dans [Spinei 01]. Enfin, des mises en œuvre efficaces ont été proposées par des implantations récursives des filtres

2. Sur les liens avec le système perceptif humain ou animal, citons [Bigun 94] pour la construction d'une famille de filtres de Gabor 3D "imitant" la perception humaine élémentaire du mouvement, et [Jasinschi 91] où certaines activités neurologiques du cortex visuel primaire du chat sont modélisées par des filtrages de Gabor 3D.

3. Ces deux caractérisations sont présentées dans [Simoncelli 91,Chomat 00]. Dans [Chomat 00, Sec. 3.4], l'auteur replace ces méthodes dans le cadre respectivement d'une décomposition spatiale (développement de Taylor) et d'une décomposition spectrale (série de Fourier). Certaines similarités entre ces deux familles ont été notées, par exemple dans [Adelson 86]. Une méthode de filtrage par dérivation spatio-temporelle des intensités a notamment été utilisée dans [Wildes 98] pour la détection de zones d'activité en vidéo-surveillance.

4. Il s'agit des classes suivantes : mouvement stationnaire, mouvement cohérent, mouvement incohérent, mouvement oscillant, scintillation, mouvement non structuré. Une telle stratégie avait déjà été suggérée dans [Adelson 86] à travers son étude de l'axe d'oscillation (*flicker axis*).

[Spinei 98b] ou par le recours à des architectures spécialisées comme les DSP (*Digital Signal Processor*) [Spinei 00].

Avant de passer à la définition mathématique des filtres utilisés, nous allons rappeler quelques problèmes concernant les filtres de Gabor spatio-temporels.

Un reproche parfois fait à l'encontre de ces filtres est qu'ils ne sont pas causaux. Toutefois, il est noté dans [Adelson 86] qu'ils sont mathématiquement efficaces, et les travaux présentés dans [Bigun 94,Jasinschi 91] ont montré qu'ils peuvent constituer une modélisation appropriée du système perceptif pour les traitements de bas niveau. Notons que nous n'avons pas trouvé au cours de nos lectures de proposition de filtres causaux aptes à remplacer efficacement les filtres de Gabor 3D.

La question de la gestion des incertitudes est soulevée par de nombreux auteurs (voir par exemple [Heeger 88, Sec. 5]). L'incertitude sur l'estimation du mouvement provient des deux aspects suivants : (i) la pertinence de la quantification de l'espace spatio-temporel 3D, liée à la localisation et à l'étendue des filtres, et (ii) la nature des textures dans l'image (le mouvement vu comme une "orientation" dans l'espace 3D des fréquences spatio-temporelles sera plus facilement détectable pour une image très texturée). Dans [Jasinschi 91], une étude sur l'influence du paramétrage de la famille de filtres considérée a été menée, et conclut que l'incertitude sur la mesure du mouvement est minimisée si la largeur de bande temporelle est supérieure à la largeur de bande spatiale.

Le problème de l'ouverture et celui de l'aliasage, déjà évoqués brièvement à la sous-section 7.1.1, se posent aussi bien sûr pour ce type de méthodes. Comme il est noté dans [Heeger 88], il semble que pour les méthodes fréquentielles, le problème de l'ouverture soit moins crucial que pour les méthodes différentielles en raison du lissage inhérent au filtrage spatio-temporel effectué [Spinei 98a, p. 8].

Un filtre de Gabor spatio-temporel est défini comme suit [Heeger 88], selon qu'il s'agit d'une modulation en sinus $g_{odd}$ ou d'une modulation en cosinus $g_{even}$ :

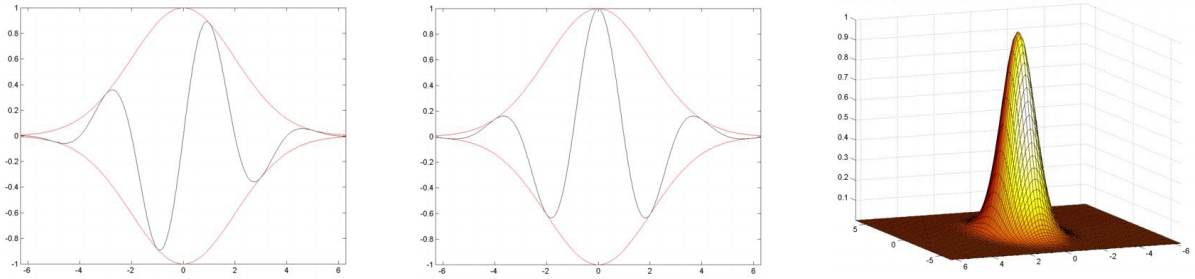
$$\begin{aligned} g_{odd}(x,y,t) &= \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x\sigma_y\sigma_t} \times e^{-\left(\frac{x^2}{2{\sigma_x}^2}+\frac{y^2}{2{\sigma_y}^2}+\frac{t^2}{2{\sigma_t}^2}\right)} \times \sin(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y + 2\pi\omega_{t_0}t) \\ g_{even}(x,y,t) &= \frac{1}{\sqrt{2}\pi^{\frac{3}{2}}\sigma_x\sigma_y\sigma_t} \times e^{-\left(\frac{x^2}{2{\sigma_x}^2}+\frac{y^2}{2{\sigma_y}^2}+\frac{t^2}{2{\sigma_t}^2}\right)} \times \cos(2\pi\omega_{x_0}x + 2\pi\omega_{y_0}y + 2\pi\omega_{t_0}t) \end{aligned} \tag{8.5}$$

où $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$ est la fréquence centrale du filtre (ou orientation du filtre, c'est-à-dire la fréquence spatio-temporelle pour laquelle le filtre répond le plus fortement) et $(\sigma_x, \sigma_y, \sigma_t)$ les écarts-type associés à la gaussienne. Une représentation graphique de ces deux fonctions est fournie à la figure 8.4 en dimension 1.

Lorsque l'orientation et la variance de la gaussienne sont fixées, les filtres $g_{odd}$ et $g_{even}$ ont la même enveloppe et sont en opposition de phase. Afin d'obtenir une réponse en énergie indépendante de la phase, il est classique d'utiliser les filtres précédemment définis *en quadrature*, sous la forme :

$$g_e(x,y,t) = g_{even}(x,y,t)^2 + g_{odd}(x,y,t)^2$$

$g_e$ est souvent appelé *filtre d'énergie* de Gabor [Spinei 98a], son spectre d'amplitude dans le domaine fréquentiel est donné par :



(a) (b) (c)

FIG. 8.4: *Construction d'un filtre d'énergie spatio-temporel de Gabor : (a) filtre à modulation en sinus $g_{odd}$ en dimension 1 ; (b) filtre à modulation en cosinus $g_{even}$ en dimension 1 ; (c) réponse en énergie $g_e$ en dimension 2.*

$$G_e(\omega_x,\omega_y,\omega_t) = \begin{array}{l} \frac{1}{4}e^{-4\pi^2[{\sigma_x}^2(\omega_x-\omega_{x_0})^2+{\sigma_y}^2(\omega_y-\omega_{y_0})^2+{\sigma_t}^2(\omega_t-\omega_{t_0})^2]} \\ +\frac{1}{4}e^{-4\pi^2[{\sigma_x}^2(\omega_x+\omega_{x_0})^2+{\sigma_y}^2(\omega_y+\omega_{y_0})^2+{\sigma_t}^2(\omega_t+\omega_{t_0})^2]} \end{array} \quad (8.6)$$

Le spectre d'amplitude du filtre défini par la relation 8.6 est formé de deux lobes symétriques par rapport à l'origine (voir figure 8.5.a[5]). La vitesse $\vec{v}$ étant traduite par un plan dans l'espace des fréquences spatio-temporelles, la réponse $G_e$ sera d'autant plus forte que ce plan passera près de la fréquence centrale $(\omega_{x_0},\omega_{y_0},\omega_{t_0})$ du filtre. Afin de pouvoir déterminer différentes orientations spatio-temporelles des plans de vitesse, un ensemble de filtres est utilisé. Ils partagent généralement une fréquence spatiale constante $\omega_0$ définie par $\omega_0 = \sqrt{\omega_{x_0}^2+\omega_{y_0}^2}$. Il est commun alors de disposer les filtres "en cylindre" autour de l'axe des fréquences temporelles et de construire ainsi une famille de filtres $\{G_{\theta,\omega_{t_0}}\}$, en faisant varier $\omega_{t_0}$ et $\theta$, où $\theta$ est l'orientation spatiale des filtres définie par $\omega_{x_0} = \omega_0\cos\theta$ et $\omega_{y_0} = \omega_0\sin\theta$. Une illustration d'une telle famille de filtres est proposée à la figure 8.5.b pour quatre valeurs d'orientations spatiales et trois valeurs de fréquences temporelles.

Cet ensemble de filtres ne sera sensible qu'à des vitesses dont les modules sont inclus dans une certaine gamme [Spinei 98a, p. 12]. Afin de disposer d'une famille de filtres sensibles à différentes amplitudes de déplacement, une solution consisterait à faire varier $\omega_0$. Toutefois, il est généralement préféré de conserver une seule famille de filtres à $\omega_0$ constant, et de faire varier la résolution de l'image au sein d'une approche multi-résolution. Une pyramide gaussienne est ainsi construite sur plusieurs niveaux, et chaque niveau est traité par le banc de filtres de Gabor précédemment défini. Ainsi qu'il est indiqué dans [Heeger 88, Sec. 3.2], cette solution équivaut à utiliser des bancs de filtres espacés d'une octave dans le domaine des fréquences spatiales, et centrés sur la même fréquence temporelle. La figure 8.6 contient deux projections et une vue 3D des localisations étagées de ces filtres.

Une telle disposition des filtres d'énergie de Gabor conduit à une caractérisation locale du mouvement apparent par quantification de l'espace spatio-temporel.

5. Sur les figures 8.5 et 8.6, les puissances spectrales des filtres sont représentées dans l'espace des fréquences par une paire d'ellipsoïdes. Sauf mention contraire, les paramètres des filtres ont été fixés arbitrairement afin de bien les séparer et de permettre ainsi une visualisation claire. Une représentation plus proche du paramétrage réel impliquant un fort recouvrement des filtres est montrée à la figure 8.6.d.

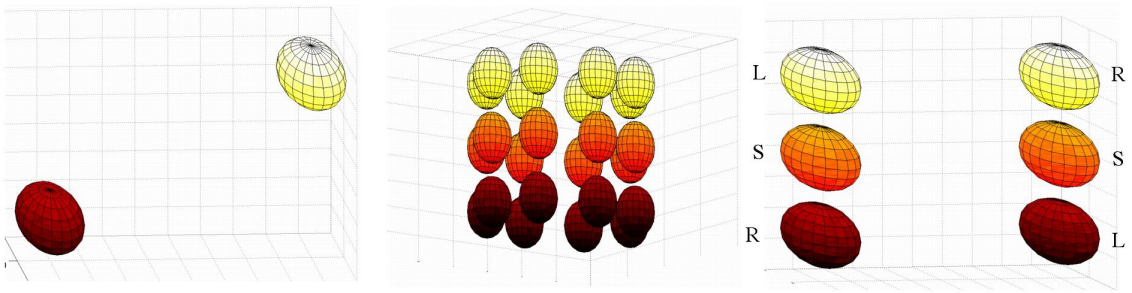



FIG. 8.5: *Représentation de familles de filtres d'énergie spatio-temporels de Gabor dans l'espace des fréquences : (a) les deux lobes d'un filtre $G_e$ ; (b) une famille de filtres paramétrés par $\{-\omega_t, 0, \omega_t\} \times \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$ ; (c) la triade d'orientation $\theta = 0$.*

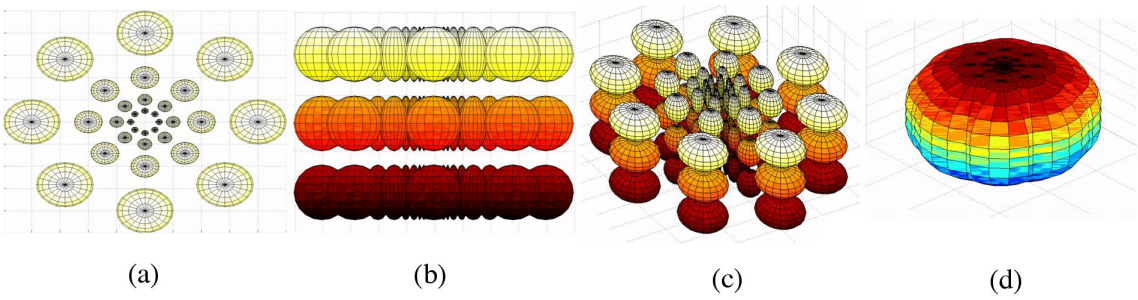



FIG. 8.6: *Représentation de l'ensemble des filtres d'énergie spatio-temporels de Gabor étagés dans l'espace des fréquences : (a) vue 2D selon le plan $(\omega_x, \omega_y)$ ; (b) vue 2D selon le plan $(\omega_x, \omega_t)$ ; (c) vue en trois dimensions ; (d) représentation des mêmes filtres pour les paramètres indiqués au paragraphe 8.4.2.*

Enfin, chacun des filtres d'énergie de Gabor ainsi associés, orientés et étagés a une réponse qui dépend à la fois de la vitesse et du contraste du motif considéré dans l'image. Ainsi, il est nécessaire d'introduire une normalisation pour s'affranchir de la dépendance de la réponse du filtre au contraste spatial dans l'image. La méthode mise en œuvre a été proposée par certains auteurs sous le nom de *triades* de filtres d'énergie de Gabor [Adelson 86,Chomat 00]. Pour un niveau de la pyramide donné et une orientation spatiale donnée, les filtres d'énergie de Gabor sont construits par trois selon des fréquences temporelles symétriques soit $\{-\omega_{t_0}, 0, \omega_{t_0}\}$. La représentation d'une telle triade est donnée figure 8.5.c. Elle contient les deux lobes du filtre d'énergie de Gabor pour $\omega_t = \omega_{t_0}$ noté $G_R$, pour $\omega_t = 0$ noté $G_S$ et pour $\omega_t = -\omega_{t_0}$ noté $G_L$. L'extraction de l'information de mouvement pour une fréquence spatiale et un niveau de pyramide est effectuée par la normalisation suivante :

$$g_T(x, y, t) = \frac{g_R(x, y, t) - g_L(x, y, t)}{g_S(x, y, t)}$$

Pour notre part, nous n'avons pas pris en compte le sens du mouvement associé à une orientation

donnée, et nous avons considéré la variante suivante[6] :

$$g_T(x,y,t) = \frac{|g_R(x,y,t) - g_L(x,y,t)|}{g_S(x,y,t)}$$

Notons qu'une alternative est utilisée dans [Spinei 98b,Wildes 00]. La réponse d'un filtre d'énergie de Gabor pour une orientation donnée est normalisée par la somme des réponses des filtres correspondant à cette orientation.

En conclusion, si nous choisissons $N_\theta$ orientations spatiales, $N_{\omega_t}$ fréquences temporelles et $N_{pyr}$ niveaux de la pyramide gaussienne, nous obtenons une famille de $N_{pyr} \times N_\theta \times (2N_{\omega_t} + 1)$ filtres d'énergie de Gabor[7], ou encore $N_{pyr} \times N_\theta \times N_{\omega_t}$ triades de filtres d'énergie de Gabor.

Les différents paramètres utilisés lors des expérimentations seront détaillés à la sous-section 8.4.2.

#### *8.2.2.2 Quantification des orientations et des amplitudes*

Nous allons décrire comment nous exploitons et transformons l'information fournie par les filtres spatio-temporels de Gabor en vue de la caractérisation du contenu dynamique des séquences. Pour chaque triade correspondant à l'orientation $\theta_i$, $i \in \{0, \ldots, N_\theta - 1\}$, la réponse en chaque point des différents niveaux $L_j$, $j \in \{0, \ldots, N_{pyr} - 1\}$ de la pyramide (où $L_0$ correspond à l'image à la résolution initiale) est binarisée à l'aide de seuils $\tau_j$, $j \in \{0, \ldots, N_{pyr} - 1\}$. Pour chaque orientation $\theta_i$, $i \in \{0, \ldots, N_\theta - 1\}$, une carte de la taille de l'image initiale est ainsi construite. Nous attribuons à chaque point de cette carte une valeur dans l'ensemble $\{0, \ldots, 2^{N_{pyr}-1}\}$. La valeur 0 correspond à une absence de réponse après seuillage sur l'ensemble des niveaux de la pyramide gaussienne, la valeur $2^j$ correspond à une réponse supérieure au seuil obtenue pour la première fois, les niveaux étant parcourus du plus grossier au plus fin et les fils et descendants d'un pixel donné héritant du label de leur père ou ancêtre. Nous obtenons ainsi une cartographie de l'amplitude des mouvements locaux quantifiée sur $N_{pyr} + 1$ niveaux. Un exemple de cette quantification du mouvement est présenté à la figure 8.7.

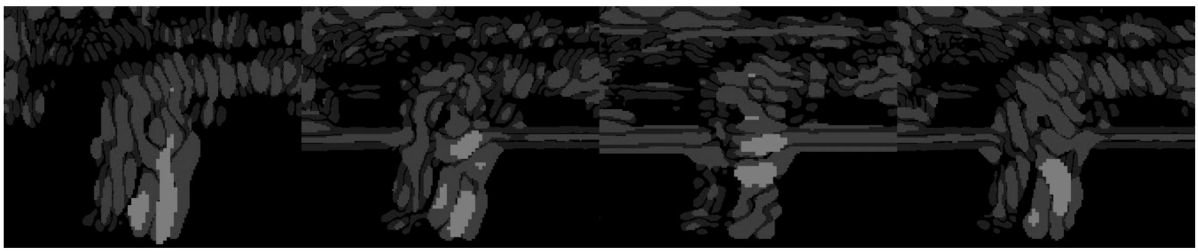



FIG. 8.7: *Représentation d'une carte des amplitudes quantifiées de mouvement pour les orientations* $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$ *sur la séquence (d) présentée à la figure 8.1.*

Les images des amplitudes quantifiées sont ainsi calculées pour l'ensemble des $\theta_i$, $i \in \{0, \ldots, N_\theta - 1\}$. Une nouvelle opération est alors effectuée tenant compte des amplitudes, et des orientations.

6. L'utilisation ultérieure que nous ferons de l'information ainsi extraite reste cependant identique dans le cas où l'on souhaiterait garder une réponse signée.

7. En effet, à chaque $\omega_t$ correspondent deux filtres de type $R$ et $L$ ; le filtre de stationnaire de type $S$ est commun à toutes les triades pour une orientation $\theta$ donnée.

En chaque point, l'orientation retenue est celle pour laquelle l'amplitude évaluée est maximale. L'image résultante est une carte des orientations et des amplitudes quantifiées respectivement sur $N_\theta$ et $N_{pyr}+1$ niveaux. Cette signature est notée $Map_{STG}$. Une valeur nulle indique l'absence de mouvement, et chaque combinaison des orientations et des amplitudes $\{\theta, 2^j\}$ correspond à une des $N_\theta \times N_{pyr}$ valeurs de quantification[8]. La figure 8.8 contient plusieurs exemples de cartes $Map_{STG}$ calculées sur les séquences présentées à la figure 8.1. Par convention, une séquence sera représentée par la $Map_{STG}$ correspondant à l'instant médian de la séquence (voir sous-section 8.2.3).

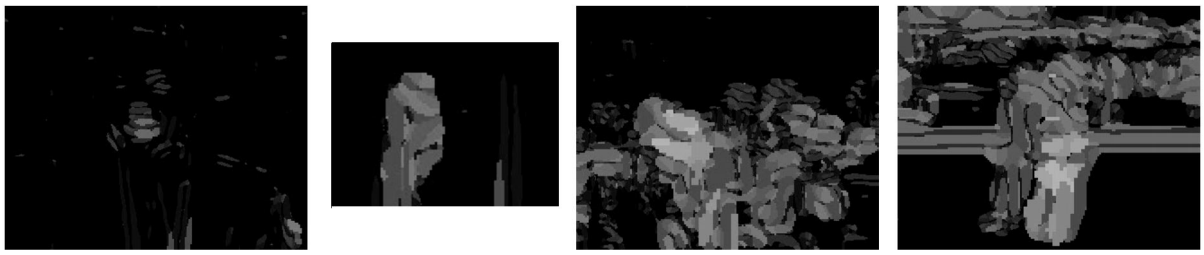


FIG. 8.8: *Représentation de signatures $Map_{STG}$ calculées sur les séquences présentées à la figure 8.1.*

Notons que si nous avions gardé une réponse signée pour chaque triade de filtres d'énergie de Gabor, la même méthode pouvait être appliquée, et la quantification du mouvement aurait été réalisée sur $1+2\times N_\theta \times N_{pyr}$ valeurs.

Le paramétrage de l'extraction des primitives retenues sera abordé et détaillé à la sous-section 8.4.2

### 8.2.3 Quelques commentaires sur les primitives extraites

Entre les différentes primitives introduites à la sous-section précédente, une des différences principales est la taille des vecteurs de primitives considérées qui formeront les entrées du classifieur. Même si nous avons opté pour un algorithme de classification réputé être modérément sensible à cet aspect, ce point n'est pas pour autant indifférent puisqu'il influe sur les temps de calcul et les espaces de stockage nécessaires. Le tableau 8.2 contient, pour chacun des descripteurs, sa taille théorique et un ordre de grandeur numérique de cette taille pour le paramétrage retenue à la sous-section 8.4. L'échelle de variation de ces valeurs est donc importante : jusqu'à $10^4$ environ.

| Signatures | $Map_{MHI}$ | $DCT_{MHI}$ | $Cooc_{MHI}$ | $H_{Cooc_{MHI}}$ | $H_{Map_{Mhi}}$ | $Map_{STG}$ |
|---|---|---|---|---|---|---|
| Taille | $W \times H$ | $p$ | $2V_{max}+1$ | 11 | 11 | $W \times H$ |
| Ordre de grandeur | $65 \cdot 10^3$ | 325 | 31 | 11 | 11 | $65 \cdot 10^3$ |

TAB. 8.2: *Expression théorique et ordre de grandeur des tailles des différentes signatures extraites*

Par ailleurs, la nature de ces primitives est variable, même si toutes rendent compte d'une information temporelle. Nous avons introduit des descripteurs localisés spatialement ($Map_{MHI}$, $Map_{STG}$) qui fournissent une valeur en chaque point de l'image, et des descripteurs globalisant

8. Au total $1+N_\theta \times N_{pyr}$ valeurs de quantification sont utilisées pour rendre compte des amplitudes et des orientations locales du mouvement.

l'information sur l'ensemble de l'image ($DCT_{MHI}$, matrice de cooccurrence, descripteurs d'Haralick). De même, certains descripteurs sont relatifs à un instant donné de la séquence ($Map_{MHI}$, $Map_{STG}$, $DCT_{MHI}$ et $H_{Map_{MHI}}$), tandis que d'autres reflètent une information cumulée temporellement ($Cooc_{MHI}$ et $H_{Cooc_{MHI}}$).

En outre, les descripteurs issus des images de l'historique du mouvement et ceux provenant des triades de filtres de Gabor spatio-temporels sont intrinsèquement différents. Dans le cas des MHI, les différences temporelles sont calculées sur des paires d'images successives, puis l'information est cumulée dans le temps. Cette opération est menée en chaque point indépendamment de son voisinage spatial. De plus, à un instant donné, seule l'information antérieure est prise en compte. Pour ce qui concerne les triades de filtres de Gabor, l'application d'un noyau de convolution spatio-temporel implique que le calcul effectué en un point intègre des informations du voisinage spatial et temporel de ce point. La signature $Map_{MHI}$ donne une datation du dernier mouvement perçu en un point indépendamment de son voisinage, tandis que la signature $Map_{STG}$ contient des informations d'amplitude et d'orientation locales, et intègre l'information sur un voisinage spatio-temporel de chaque point. Comme nous pouvons le voir sur les figures 8.1 et 8.8, cette richesse est obtenue au prix d'un certain "moyennage" de l'information.

Dans les expérimentations, nous avons considéré des séquences constituées de 15 images dans la mesure où le traitement de séquences plus longues n'aurait pas été pertinent pour ce qui est de $Map_{MHI}$, compte tenu du $V_{max}$ choisi (voir sous-section 8.4.1). Il faut alors souligner que l'horizon temporel des filtres de Gabor étant de 7 images, seule une partie des séquences intervient dans le calcul de la signature $Map_{STG}$.

Nous avions souhaité des signatures capables de décrire des séquences sur un horizon temporel de l'ordre de la seconde (voir sous-section 7.1.2). En fait, seuls les matrices de cooccurrence temporelle et les descripteurs d'Haralick qui en sont issus peuvent satisfaire cette contrainte.

Par souci de cohérence, nous avons choisi de garder une longueur de séquence de 15 images, et d'évaluer la signature $Map_{STG}$ en leur milieu. Enfin, pour des raisons pratiques, nous n'avons pas pris un compte un recalage des images après compensation du mouvement dominant pour les signatures fondées issues des filtres spatio-temporels de Gabor. Compte tenu de l'étendue temporelle de ces filtres, il aurait fallu recaler à chaque fois un ensemble de sept images successives. Lorsque le mouvement de caméra est important (ce qui est fréquent en sport), la zone d'intersection des images recalées peut être très réduite, ce qui rend le filtrage difficile à mettre en pratique.

## 8.3 Classification des séquences par les machines à vecteurs de support

Nous avons fait le choix, pour la méthode de classification, des machines à vecteurs de support (*Support Vector Machines* - SVM) qui nous ont paru être un outil efficace. Comme affirmé dans [Scholkopf 98], les SVM présentent notamment deux avantages majeurs :

1. les machines à vecteurs de support sont issues de la théorie de l'apprentissage statistique qui définit un cadre théorique rigoureux ;

2. les SVM affichent pour des applications pratiques de très bonnes performances.

La première affirmation semble tout à fait justifiée, et nous avons essayé de donner quelques éléments de compréhension de ce cadre théorique dans l'annexe C. La seconde paraît moins évidente. Comme le reconnait B. Schölkopf lui-même, les SVM ont donné des résultats comparables à l'état de

l'art du domaine, mais n'ont pas fait franchir de saut quantitatif notable aux outils de classification. Quelques travaux ont évalué, sur des applications spécifiques, les performances des SVM par rapport aux réseaux de neurones [Scholkopf 95], aux perceptrons [Pontil 98b], à différents algorithmes de recherche (arbres de décision, réseaux bayésiens, etc.) [Dumais 98b], à des classifieurs fondés sur la méthode du plus proche voisin avec ou sans information *a priori* [Roobaert 99,Guo 00], à la méthode de détection de visage de Sung & Poggio [Osuna 97a] et à diverses heuristiques [Blanz 96]. Ces travaux ont montré que les SVM fournissent des résultats en général équivalents ou légèrement supérieurs aux autres méthodes. Un autre avantage des SVM parfois mentionné est de ne pas nécessiter la définition de l'architecture du classifieur avant son utilisation (comme c'est le cas par exemple pour les réseaux de neurones), ce qui réduit le nombre de paramètres à régler [Smola 96]. Par ailleurs, elles semblent offrir un bon pouvoir de généralisation.

Après un rappel succinct du principe des SVM, nous décrirons les mises en œuvre effectuées pour l'apprentissage des classifieurs et pour l'utilisation conjointe de bancs de classifieurs à des fins de caractérisation du contenu dynamique de séquences audiovisuelles.

### 8.3.1 Principe des machines à vecteurs de support

Les SVM sont une méthode de classification par apprentissage fondée sur la théorie de l'apprentissage statistique [Vapnik 95,Osuna 97b,Burges 98,Cristianini 00]. Leur principe est la recherche d'une surface de décision optimale par la minimisation sous contraintes d'une fonction définissant un hyperplan séparateur optimal ("Optimal Separating Hyperplane" - OSH). La capacité théorique de généralisation des SVM repose sur le choix de la surface de décision associée à la plus grande marge possible compte tenu des contraintes. Une caractéristique intéressante est la possibilité de transformer aisément un problème non linéaire en un problème linéaire par extension de la dimension de l'espace des variables, tout en ne nécessitant que la définition d'une fonction particulière, appelée *noyau*, et le calcul de produits scalaires. Pour ce qui nous concerne, nous nous sommes contentés d'utiliser les SVM comme un outil "boîte noire", en ayant recours à des logiciels disponibles sur l'internet. Nous avons néanmoins rassemblé dans l'annexe C les informations qui nous ont paru nécessaires à l'utilisation de cet outil et au réglage de ses paramètres.

L'étude des SVM est un domaine de recherche encore relativement émergent qui a connu ces dernières années un certain succès au travers d'applications variées. Parmi les problématiques traitées par cette méthode, nous pouvons citer, notamment, la catégorisation de textes [Dumais 98a], la détection d'événements [Papageorgiou 97,Papageorgiou 99,Pittore 00], la classification de gestes [Pittore 00], la détection de visages [Osuna 97a,Osuna 97b,Papageorgiou 97], la recherche par similarité sur des textures [Guo 00], la reconnaissance d'objets 3D [Blanz 96,Pontil 98b,Roobaert 99] ou 2D sous différents angles de vue [Karlsen 00], la classification de chiffres écrits à la main [Scholkopf 95], la vérification d'identité [Gutschoven 00], l'identification de particules [Barabino 99], etc.

Il nous faut toutefois préciser que nous nous trouvons dans un cas de figure légèrement atypique dans l'utilisation que nous faisons des SVM. Dans les travaux cités plus haut, la dimension des vecteurs d'entrée est très variable (entre 3 et 1024), le nombre d'exemples disponibles pour l'apprentissage est généralement élevé (entre 13 et plusieurs centaines par classe). Comme nous l'avons vu à la sous-section 8.2.3, les dimensions de nos vecteurs d'entrée sont assez diverses et plutôt plus élevées, mais les SVM sont considérés comme n'étant que peu sensibles à la taille du vecteur d'entrée (voir [Karlsen 00]). Par contre, les tailles de nos bases d'apprentissage sont comparativement faibles (entre 3 et 20 exemples par classe, voir tableau 9.1). Nous n'avons pas trouvé, dans les différents articles sur le sujet, d'indications claires concernant le nombre d'exemples nécessaires à l'apprentissage, mais il n'est pas exclu, compte tenu des chiffres cités plus haut, que nos bases d'apprentissage

soient un peu trop réduites pour les descripteurs utilisés. Dans [Osuna 97a], il est recommandé, après une étude empirique, que la taille de l'ensemble d'apprentissage soit de 20% supérieur au nombre de vecteurs de support permettant la classification souhaitée. Ce nombre n'étant pas connu *a priori*, cette indication ne peut être aisément mise en pratique.

Pour nos expérimentations, nous avons utilisé la librairie *LibSVM* développée à l'Université Nationale de Taiwan [Chang 01]. Cette librairie[9] permet l'utilisation des noyaux les plus classiques. La gestion de ses paramètres sera abordée à la sous-section 8.4.3.

### 8.3.2 Stratégies de classification par des machines à vecteurs de support

La librairie *LibSVM* permet également de traiter des problèmes de régression et de classification multi-classes. Toutefois, considérant que l'utilisation de classifieurs multi-classes nécessitait une compréhension plus approfondie et que ce type de solution était encore fort récent, nous nous sommes contentés d'utiliser des classifieurs à deux classes. Il existe différentes stratégies envisageables pour traiter des problèmes multi-classes avec des classifieurs à deux classes. Les deux principales sont la stratégie du "un contre tous" (*One versus All*) et celle du tournoi de tennis (*Tennis Tournement*) [Pittore 00, Sec. 6.3]. La méthode "un contre tous" nécessite un classifieur par classe, soit $N_c$ classifieurs s'il y a $N_c$ classes. Chaque classe est alors apprise contre le "reste du monde" ; la difficulté est de définir un ensemble d'apprentissage suffisamment représentatif du "reste du monde". Les exemples de test sont ensuite présentés à l'ensemble des classifieurs: $N_c$ tests sont nécessaires afin de prendre une décision. Dans la méthode du "tournoi de tennis", des "matchs" sont organisés entre classes. Il est alors nécessaire d'entraîner $N_c!$ classifieurs en mode "un contre un". La prise d'une décision nécessite de soumettre l'exemple de test à $N_c - 1$ classifieurs.

Comme nous allons le voir plus en détails dans les paragraphes suivants, nous avons opté pour la stratégie du "un contre tous" afin de limiter le nombre de classifieurs nécessaires.

### 8.3.3 Construction d'un classifieur et apprentissage

Considérons un problème de classification à $N_c$ classes pour lequel une base de $n_e$ exemples par classe serait disponible. Les exemples sont répartis entre $n_l$ exemples d'apprentissage et $n_t$ exemples de test pour chaque classe. Cette répartition est effectuée au hasard et est appliquée similairement à la construction des $N_c$ classifieurs. Compte tenu de la relative[10] faiblesse numérique des exemples disponibles dans nos expérimentations, la plupart des exemples sont utilisés à des fins d'apprentissage dans un rapport d'environ $\frac{n_t}{n_l} = \frac{1}{2}$. Les exemples d'apprentissage sont censés être représentatifs de la variabilité éventuelle des éléments d'une classe, et, après apprentissage, les vecteurs de support tiendront lieu de modélisation de la classe apprise.

Nous noterons $N_e = n_e \times N_c$ le nombre total de séquences disponibles dans la base, $N_l = n_l \times N_c$ le nombre total d'échantillons d'apprentissage et $N_t = n_t \times N_t$ le nombre total d'échantillons de la base de test. Chaque classifieur $Cl_i$, $i \in \{1, \ldots, N_c\}$ est entraîné afin de reconnaître la classe associée $C_i$ au sein de l'ensemble des classes définies sur les $\frac{2}{3}$ des $N_e$ exemples de la base. Nous procédons, dans une certaine mesure, à une simplification puisqu'au "reste du monde" est substitué le "reste des classes", ce qui nous permet de contourner le problème de la définition d'un ensemble d'exemples représentatifs du "reste du monde". De plus, ce choix nous place dans une configuration de type "monde fermé" : l'univers est restreint aux classes en présence.

9. Disponible à l'adresse `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

10. Nous avons noté à la sous-section 8.3.1 que nous avions peu d'exemples en comparaison de l'utilisation habituelle des SVM, toutefois par rapport à d'autres travaux sur la classification de séquences, leur nombre est tout à fait honorable.

Lorsque nous présentons une séquence $\mathcal{S}$ de la base à un classifieur $Cl_i$, $i \in \{1, \ldots, N_c\}$ correctement entraîné, la réponse valuée $V_{Cl_i}(\mathcal{S})$ doit être positive si $\mathcal{S}$ appartient à la classe $C_i$ et négative sinon.

Afin de pouvoir évaluer la construction des classifieurs obtenue, nous avons exhibé, pour chaque classifieur $Cl_i$, $i \in \{1, \ldots, N_c\}$, les informations ci-après :

- les matrices de confusion des classes calculées séparément sur les ensembles d'apprentissage et de test ;
- le taux $T^l_{pos}(C_j)$ de classification correcte pour chaque classe $C_j$, $j \in \{1, \ldots, N_c\}$ calculé sur l'ensemble d'apprentissage ;
- le taux $T^t_{pos}(C_j)$ de classification correcte pour chaque classe $C_j$, $j \in \{1, \ldots, N_c\}$ calculé sur l'ensemble de test ;
- le taux $T^l_{pos}$ de classification correcte pour l'ensemble des classes calculé sur l'ensemble d'apprentissage, $T^l_{pos} = \frac{1}{N_c} \sum_{j=1}^{N_c} T^l_{pos}(C_j)$ ;
- le taux $T^t_{pos}$ de classification correcte pour l'ensemble des classes calculé sur l'ensemble de test, $T^t_{pos} = \frac{1}{N_c} \sum_{j=1}^{N_c} T^t_{pos}(C_j)$ ;
- le taux $T^{l+t}_{pos}$ de classification correcte sur l'ensemble des échantillons disponibles, $T^{l+t}_{pos} = \frac{N_l T^l_{pos} + N_t T^t_{pos}}{N_e}$ ;
- la valeur de sortie du classifieur $V_{Cl_i}(\mathcal{S}_k)$, $k \in \{1, \ldots, N_e\}$ pour l'ensemble des exemples ;
- la valeur moyenne de la sortie du classifieur $V^l_{Cl_i}(C_j)$ pour chaque classe $C_j$, $j \in \{1, \ldots, N_c\}$ sur les exemples d'apprentissage ;
- la valeur moyenne de la sortie du classifieur $V^t_{Cl_i}(C_j)$ pour chaque classe $C_j$, $j \in \{1, \ldots, N_c\}$ sur les exemples de test ;
- le nombre de vecteurs de support $N_{SV}$ et leur identité (pour cela nous avons dû modifier légèrement la librairie *LibSVM*). Nous avons aussi récupéré le nombre de vecteurs de support "problématiques" $N^c_{SV}$, c'est-à-dire ceux dont le multiplicateur de Lagrange associé est maximal (voir les explications données en annexe à la sous-section C.4.2).

### 8.3.4 Caractérisation des séquences

Une fois les $N_c$ classifieurs construits, chaque échantillon de test $\mathcal{S}_k$, $k \in \{1, \ldots, N_t\}$ est présenté à l'ensemble des classifieurs $Cl_i$, $i \in \{1, \ldots, N_c\}$. La décision d'associer la séquence $\mathcal{S}_k$ à l'une des classes $C_j$, $j \in \{1, \ldots, N_c\}$ est prise comme suit :

$$C(\mathcal{S}_k) = argmax\{V_{Cl_i}(\mathcal{S}_k),\ i \in \{1, \ldots, N_c\}\},\ \text{pour } k \in \{1, \ldots, N_t\} \tag{8.7}$$

Les informations extraites à cette étape sont :

- la matrice de confusion des classes calculée sur l'ensemble de test ;
- le taux $T_{pos}(C_j)$ de classification correcte pour chaque classe $C_j$, $j \in \{1, \ldots, N_c\}$ calculé sur l'ensemble de test ;

- le taux $T_{pos}$ de classification correcte pour l'ensemble des classes calculé sur l'ensemble de test ;
- la valeur de sortie de chacun des classifieurs pour chacun des exemples de test.

Lors de la présentation des expérimentations menées (cf. chapitre 9), seules les matrices de confusion et les taux de classification correcte seront systématiquement donnés. Les autres informations ne seront fournies que dans la mesure où elles apportent un éclairage pertinent sur les résultats obtenus.

## 8.4 Gestion des paramètres pour la caractérisation des séquences

Nous allons dans cette section passer en revue les différents paramètres liés à l'extraction des primitives sur les séquences et à l'utilisation des machines à vecteurs de support. Nous avons essayé pour la plupart d'entre eux de fixer des valeurs pertinentes sur l'ensemble des expérimentations. L'exploration de toutes les variantes possibles aurait été bien entendu trop coûteuse en temps et aurait rendu difficile la lecture des résultats.

Certains paramètres ont été fixés empiriquement à partir d'expérimentations menées sur un ensemble réduit mais représentatif des séquences étudiées. Pour quelques autres, nous avons repris des valeurs communément utilisées ; pour d'autres encore, nous n'avons pu proposer de solution entièrement satisfaisante.

### 8.4.1 Paramètres liés à l'utilisation de l'image de l'historique du mouvement

Les deux paramètres nécessaires à l'extraction des MHI sont $V_{max}$ et $\tau$. Pour l'extraction de la signature $DCT_{MHI}$, il faudra de plus fixer une valeur pour $p$.

Dans [Davis 97], il est indiqué pour $V_{max}$ des valeurs comprises entre 11 et 19. Les quelques expérimentations que nous avons faites nous ont amenés à prendre une valeur $V_{max} = 15$. Pour des $V_{max}$ plus grands, l'information captée dans les MHI devient vite illisible ou "surchargée", dans la mesure où trop de mouvements sont "mémorisés" et superposés au sein de la MHI. La valeur de $\tau$ a été fixée sans trop de difficultés après quelques expérimentations. Il est apparu que, pour les séquences à caméra fixe et à caméra mobile, les valeurs $\tau = 30$ et $\tau = 50$ étaient respectivement optimales. Un seuillage plus fort pour les séquences à caméra mobile permet, en effet, une légère modération du "bruit" lié au mouvement du décor.

Pour la signature $DCT_{MHI}$, nous avons empiriquement défini la valeur de $p$ selon la formule $p = \frac{q(q+1)}{2}$, avec $q = \frac{1}{10}\min(W, H)$. Cette valeur de $q$ fixée après expérimentations permet une réduction importante de l'information sans la dégrader de manière rédhibitoire.

### 8.4.2 Paramètres liés à l'utilisation des filtres de Gabor spatio-temporels

La construction des familles de filtres spatio-temporels de Gabor nécessite de fixer plusieurs paramètres : la largeur de l'enveloppe gaussienne $(\sigma_x, \sigma_y, \sigma_t)$ et les fréquences centrales $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$. Pour tous ces paramètres, nous avons repris les valeurs les plus communément utilisées. Comme dans [Heeger 88], nous avons retenu les valeurs $\sigma_x = \sigma_y = 4$ et $\sigma_t = 1$. La taille du noyau de convolution [11] est alors de 23 pixels pour les dimensions spatiales et de 7 images pour la dimension

11. Afin d'éviter les effets de bord aux frontières des images ou aux bornes des séquences considérées, nous avons normalisé la réponse des filtres de Gabor en un point par la taille du support "actif" lors de la convolution effectuée en ce point.

temporelle[12]. Notons toutefois que cette configuation ne nous assure pas d'après [Jasinschi 91] une minimisation de l'incertitude (voir paragraphe 8.2.2.1).

En ce qui concerne les fréquences centrales, nous avons aussi repris les valeurs choisies dans [Heeger 88]. La fréquence spatiale a été fixée à $\omega_0 = \frac{1}{4}$ de cycles par point d'image, et les quatre orientations spatiales suivantes ont été retenues : $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$. Les fréquences temporelles, outre $\omega_t = 0$, ont été fixées à $\omega_t = \pm\frac{1}{4}$ de cycles par image pour les filtres de type $L$ et $R$. Comme il est noté dans [Spinei 98a, p. 20], D. Heeger, pour paramétrer sa famille de filtres, s'est appuyé sur des données biologiques afin de tenter de reproduire les performances du système visuel humain, mais il n'apporte aucune justification théorique. A. Spinéi propose dans ses travaux une approche fondée sur un critère de distance permettant de justifier le choix des valeurs de $\sigma_x$, $\sigma_y$, et $\sigma_t$. Ses conclusions lui permettent d'aboutir à une valeur optimale de $\sigma_t$ et à un domaine de validité pour $\sigma_x$ et $\sigma_y$ qui s'avèrent finalement compatibles avec le paramétrage initialement retenu par D. Heeger [Spinei 98a, Sec. 3.2.4]. Toutefois, le choix de ces valeurs reste délicat. Si, nous avons opté pour le paramétrage proposé par [Heeger 88], il convient de souligner que dans les travaux cités, ces filtres sont utilisés pour estimer le flot optique, ce qui n'est pas notre cas. Notons que O. Chomat utilise un paramétrage différent (mais aussi arbitraire) avec $\sigma_t = \sigma_x = \sigma_y = 1.49$ [Chomat 99b, Sec. 5].

Concernant la construction de la pyramide gaussienne, nous avons introduit quatre niveaux de résolution. Pour la configuration correspondant des filtres de Gabor étagés, les domaines de vitesse détectés aux différents niveaux de la pyramide sont estimés à $[0, 1.25]$ points par image pour le niveau $L_0$, puis à $[1.25, 2.5]$, $[2.5, 5]$ et $[5, 10]$ pour les trois niveaux suivants $L_1$, $L_2$ et $L_3$ respectivement [Spinei 98a, Sec. 2.3.2].

Le choix des paramètres $(\tau_0, \tau_1, \tau_2, \tau_3)$ a été malaisé. Nous avons expérimenté plus d'une vingtaine de jeux de paramètres sans parvenir à une solution pleinement satisfaisante. La difficulté d'un choix empirique est liée à l'interdépendance des seuillages aux différents niveaux, et à la nécessité d'équilibrer, sur un ensemble de séquences plus ou moins représentatives, l'information issue des différents niveaux de la pyramide. Nous avons finalement opté pour le paramétrage suivant : $\tau_0 = 0.005$, $\tau_1 = 0.01$, $\tau_2 = 0.025$, $\tau_3 = 0.05$.

### 8.4.3 Paramètres liés à l'utilisation des machines à vecteurs de support

Le principal paramètre lié aux SVM est le paramètre d'apprentissage $C$, qui contrôle, dans la fonction coût, l'importance[13] accordée aux erreurs de classification (voir les explications apportées, en annexe, à la section C.4.2). Les autres options sont liées à des choix de définition du SVM (SVM linéaire, utilisation d'un noyau, type de fonction noyau utilisé et ses paramètres), ou à des paramètres spécifiques au fonctionnement de la librairie *LibSVM*.

Les paramètres ou options relatifs à l'utilisation de la librairie *LibSVM* (notamment la taille du cache mémoire, le critère d'arrêt des itérations, l'utilisation d'une décomposition du problème en sous-problèmes) ont été pris aux valeurs par défaut indiquées dans la documentation technique accompagnant cette librairie. Par ailleurs, nous nous sommes situés dans le cadre de classifieurs à deux classes, et nous n'avons pas utilisé de pondération différenciée des erreurs en fonction des classes, lors de l'apprentissage.

Pour les choix de configuration du SVM, nous nous sommes fiés aux travaux d'évaluation des SVM sur différents problèmes (voir section 8.3), desquels nous avons conclu que, globalement, les

12. Notons, comme conséquence, que les horizons temporels des signatures $Map_{MHI}$ et $Map_{STG}$ sont différents. La $Map_{MHI}$ a un horizon temporel de 15 images vers le passé, et la $Map_{STG}$ a un horizon temporel de 3 images vers le passé et de 3 vers le futur (voir sous-section 8.2.3).

13. D'après [Pontil 98a, Sec. 3.2.3], plus $C$ sera faible plus l'*OSH* maximisera la marge, et plus $C$ sera grand plus l'*OSH* sera positionné afin de minimiser le nombre d'erreurs sur l'ensemble d'apprentissage.

SVM associés à des noyaux de fonctions à base radiale (*Radial Basis Function* - RBF) donnaient, en général, des résultats légèrement supérieurs. Une fois ce choix effectué, il restait à fixer une valeur pour le paramètre $\gamma$ apparaissant dans la formule de la RBF. Nous avons, une fois encore, opté pour la configuration par défaut. Dans l'utilisation de *LibSVM*, le paramètre $\gamma$ est donc calculé au sein de la librairie en fonction des données d'apprentissage.

Enfin, le paramètre $C$ de pondération des erreurs a été plus problématique à choisir. D'une part, nous n'avons pas trouvé d'information précise à ce sujet, utilisable pour nos expérimentations. D'autre part, la documentation technique fournie avec la librairie *LibSVM* ne donne aucune indication numérique pour ce paramètre. La valeur par défaut proposée est $C = 1$. Les quelques expériences limitées que nous avons menées ont indiqué que, pour des valeurs de $C$ inférieures à 1, l'apprentissage devenait impossible, et que, pour des valeurs de $C$ supérieures à 100, aucune modification des résultats n'était plus perceptible lors de la caractérisation des séquences. Cette observation est cohérente avec les remarques trouvées dans [Pontil 98a] sur l'influence du paramètre $C$ dans la définition de l'*OSH*. Les auteurs signalent notamment l'existence de plages de valeurs de $C$ pour lesquelles l'ensemble des vecteurs de support reste constant, et la faible influence, sous certaines conditions, des petites variations de $C$ sur la construction de l'*OSH*. La gestion du paramètre $C$ au sein du domaine $[1, 100]$ est restée problématique dans nos expérimentations. Ce point sera abordé de nouveau au paragraphe 9.2.1.2.

# Chapitre 9

# Expérimentations

Nous avons testé les techniques proposées sur cinq bases d'expérimentations différentes. Plusieurs essais ont été nécessaires au préalable pour fixer les options et la paramétrisation des algorithmes développés. Nous avons été amenés à étudier près de 400 versions de classifieurs. Les sections suivantes présentent respectivement la méthodologie adoptée, les résultats expérimentaux obtenus, ainsi que quelques expérimentations complémentaires.

## 9.1 Objectifs et méthodologie d'évaluation des algorithmes

Afin d'évaluer les différentes signatures à la fois d'un point de vue algorithmique et du point de vue des usages, nous avons organisé cinq expérimentations séparées.

| Tests | $N_c$ | $N_e$ | $\frac{N_l}{N_c}$ | $W \times H$ | Thématique | Commentaires |
|---|---|---|---|---|---|---|
| Test 1 | 5 | 85 | 12 | $304 \times 224$ | type de mouvements | séquences de mouvements homogènes et bien différenciés, utilisation d'échantillons redondants |
| Test 2 | 6 | 30 | 3 | $176 \times 128$ | activité humaine | différents types de déplacements à caméra fixe |
| Test 3 | 3 | 54 | 12 | $176 \times 126$ | mouvement de caméra | différents mouvements de caméra sur un motif texturé |
| Test 4 | 5 | 100 | 14 | $304 \times 224$ | type de mouvements | mouvements de nature différente dont l'apparence présente des variations (éclairage, orientation, amplitude, valeur de plan, texture, angle de vue, etc.) |
| Test 5 | 5 | 150 | 20 | $304 \times 224$ | type de sports | séquences de différents sports à caméra mobile présentant diverses variations (direction, angle de vue, valeur de plan, amplitude, etc.) |

TAB. 9.1: *Présentation des cinq bases d'expérimentations retenues (où $N_c$ est le nombre de classes, $N_e$ et $N_l$ sont respectivement le nombre total d'extraits disponibles et le nombre d'exemples utilisés pour l'apprentissage)*

Les trois premières ont pour objectif de valider les algorithmes mis en œuvre, les deux dernières doivent permettre une évaluation de la portée applicative de nos travaux dans le contexte d'usage considéré. Le tableau 9.1 contient une fiche technique décrivant synthétiquement les cinq bases de séquences utilisées. Les documents, dont nous avons extrait les séquences formant les différentes bases d'expérimentations, sont décrits dans le tableau A.1 de l'annexe A.

***Expérimentation 1***

Cette première base d'expérimentation a été construite afin d'évaluer l'efficacité des différentes primitives sur un exemple très simple. Elle est constituée de cinq classes. Sur les 17 extraits dispo-

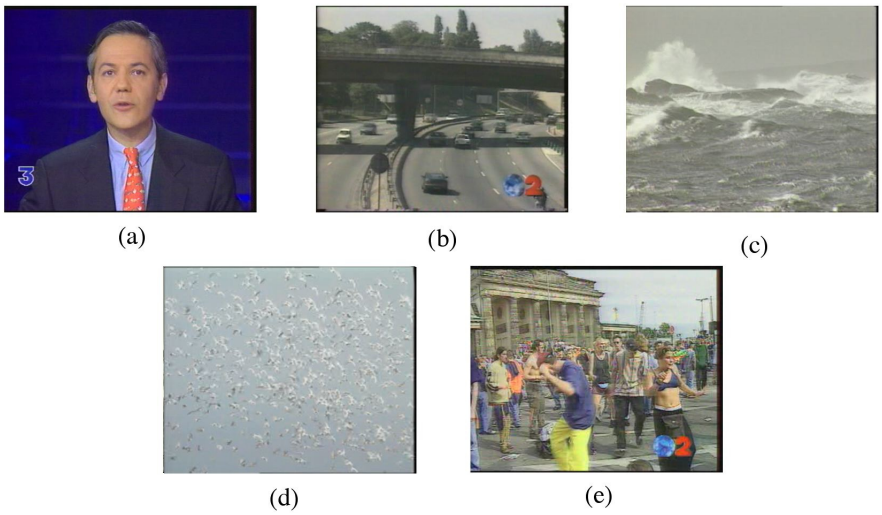



FIG. 9.1: *Exemples de séquences de la base d'expérimentation 1 : (a)* Présentateur, *(b)* Voiture, *(c)* Océan, *(d)* Oiseau, *(e)* Danseur.

nibles dans chaque classe, 12 sont utilisés lors de la phase d'apprentissage. Dans cette expérimentation, nous avons souhaité évaluer les algorithmes sur des classes très homogènes. Nous avons donc sélectionné cinq extraits de vidéo d'une durée d'environ cent images, correspondant à différents types de mouvement. Chacun des cinq extraits a été découpé en 17 séquences de 15 images. Chaque séquence présente un décalage de cinq images par rapport à la précédente. Un exemple possède une zone de chevauchement de cinq ou dix images avec au moins deux autres exemples de la même classe. Ceci nous assure une forte cohérence interne de chaque classe. Les classes sont labellisées *Voiture*, *Océan*, *Oiseau*, *Danseur*, *Présentateur*, et correspondent respectivement à des exemples de mouvement rigide, fluide, chaotique, articulé, ou à une faible activité. La classe *Voiture* est constituée d'images de plusieurs voitures roulant dans différentes directions sur un échangeur. La classe *Océan* contient les mouvements d'une mer agitée. Les extraits de la classe *Oiseau* montrent l'envol de nombreux oiseaux. Un ensemble d'individus dansent sur un rythme soutenu sur les images de la classe *Danseur*. Les séquences de la classe *Présentateur* sont issues d'un plan de plateau lors d'un journal télévisé. Ces séquences sont illustrées par des images-clefs à la figure 9.1.

Toutes les séquences sont filmées à caméra fixe (en dépit d'un léger défaut de stabilisation sur les séquences de la classe *Danseur*). Même si la construction de cette base d'expérimentation implique des classes relativement simples, cohérentes et bien séparées, observons que les scènes sont pourtant déjà complexes, en terme d'analyse d'image, à cause de la présence d'objets multiples et de phénomènes d'occultations.

Dans les autres expérimentations, aucun recouvrement entre échantillons n'a été autorisé.

***Expérimentation 2***

Pour cette expérimentation, l'objectif est de s'intéresser à un problème de reconnaissance d'activités humaines, en l'occurrence différents types de déplacements humains dans un hall, filmés à caméra fixe. La base d'expérimentation est constituée de six classes : *Aller à gauche*, *Aller à droite*, *S'approcher de la caméra*, *S'éloigner de la caméra*, *Monter l'escalier*, *Descendre l'escalier*. Chaque

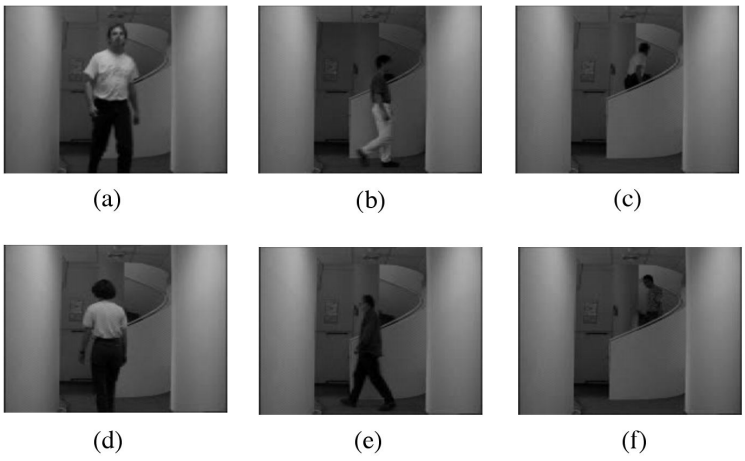



FIG. 9.2: *Exemples de séquences de la base d'expérimentation 2 : (a)* S'approcher de la caméra, *(b)* Aller à droite, *(c)* Monter l'escalier, *(d)* S'éloigner de la caméra, *(e)* Aller à gauche, *(f)* Descendre l'escalier.

type de déplacement est effectué une fois par cinq personnes différentes. Nous avons utilisé des séquences plus longues (24 images), afin d'appréhender la plus grande partie possible de chaque action. Nous avons fixé, en conséquence, le paramètre $V_{max}$ à $V_{max} = 24$. Compte tenu de l'horizon temporel considéré, seuls cinq échantillons sont disponibles par classe de déplacement, et nous en avons gardé trois par classe pour l'apprentissage. Des images représentatives des déplacements sont fournies à la figure 9.2.

Notons qu'un même déplacement peut être effectué avec des durées variables par les différents intervenants, et rappelons que nos algorithmes n'incluent aucune normalisation temporelle des séquences d'activité. Cette base d'expérimentation a été obtenue auprès du laboratoire GRAVIR de l'IMAG, et a été notamment exploitée dans les travaux décrits dans [Chomat 00,Bruno 01,Fablet 01].

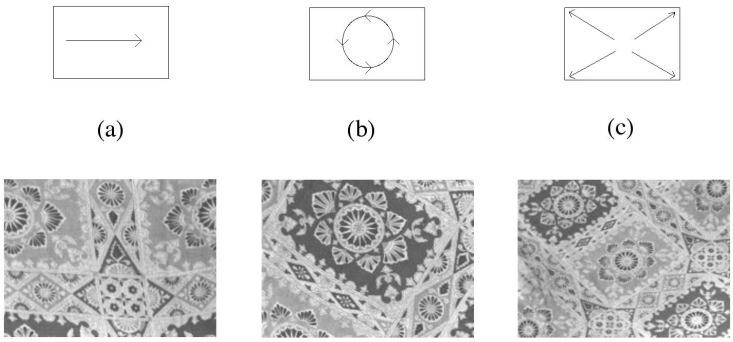



FIG. 9.3: *Exemples de séquences de la base d'expérimentation 3 : (a)* Translation, *(b)* Rotation, *(c)* Divergence.

***Expérimentation 3***

Cette expérimentation porte sur les mouvements de caméra. La troisième base d'exemples contient donc trois types de mouvements de caméra réalisés sur un motif texturé : *Translation* (ou *travelling latéral*), *Divergence* (ou *travelling en profondeur*) et *Rotation* selon l'axe de visée (qui ne correspond à aucun déplacement de caméra identifié du langage audiovisuel). Chaque classe contient 18 extraits dont 12 servent lors de la phase d'apprentissage. Ces séquences, dont des images représentatives sont présentées à la figure 9.3, ont été acquises à l'IRISA par une caméra portée par un robot cartésien.

Cette base nous permet d'expérimenter nos algorithmes sur des mouvements liés à la caméra. Rappelons que les signatures que nous avons retenues ont été proposées initialement dans un cadre d'analyse à caméra fixe (cf. section 7.3). Une difficulté éventuelle liée à cette base d'exemples est que les mouvements de caméra d'un type donné peuvent être effectués dans des directions différentes : la classe *Translation* comprend des *travellings* verticaux, horizontaux ou selon la diagonale, la classe *Divergence* est constituée de *travellings* avant ou arrière, et dans la classe *Rotation* la caméra peut tourner dans les sens trigonométrique ou anti-trigonométrique selon son axe de visée.

***Expérimentation 4***

La centaine d'extraits qui constituent la quatrième base d'exemples sont toujours filmés à caméra fixe (ou presque). Les classes considérées correspondent à différents types d'activités peu éloignés de ceux définis dans la première base : *Eau*, *Oiseau*, *Studio*, *Ballet*, *Automobile*. Toutefois, les segments utilisés sont cette fois disjoints, les types de mouvement étudiés sont plus généraux et la variabilité du mouvement au sein d'une classe est sensible. Les exemples ont été extraits de 25 séquences elles-mêmes issues de 11 documents. Ainsi, les mouvements représentés au sein des classes *Eau*, *Studio*, *Ballet*, *Automobile* sont beaucoup plus hétérogènes que ceux relevant des classes *Océan*, *Présentateur*, *Danseur*, *Voiture*. La figure 9.4 illustre, au travers d'images représentatives des séquences traitées, la variabilité des contenus dynamiques, des conditions de prise de vue, etc. Nous disposons de 14 échantillons par classe pour l'apprentissage.

Nous espérons, grâce à cette expérimentation, pouvoir évaluer les capacités d'apprentissage et de généralisation de nos classifieurs. En outre, cette base constitue un exemple intéressant de reconnaissance du mouvement. Si cette classification présente peu d'intérêt pour l'indexation proprement dite (du point de vue de l'usage final), elle constitue un intérêt du point de vue de la nature de l'information de mouvement extraite[1].

***Expérimentation 5***

Avec cette dernière expérimentation, nous abordons un exemple de mise en œuvre de nos algorithmes lié à des usages identifiés d'indexation audiovisuelle. Les exemples, au nombre de 150, sont extraits de 23 séquences provenant de 9 documents différents. Les extraits sont filmés avec une caméra mobile, et les échantillons sont par conséquent très hétérogènes. Les classes ont été définies selon les sports présents dans les retransmissions sportives issues des corpus disponibles. Les cinq classes considérées sont : *Aviron*, *Cyclisme*, *Patinage*, *Football*, *Formule 1*. Vingt extraits

1. Nous faisons référence aux remarques énoncées à la section 2.2. Dans le cadre de l'indexation des documents audiovisuels mise en pratique par les documentalistes de l'INA, une discrimination des séquences selon les classes *Ballet* et *Automobile*, par exemple, n'a que peu d'intérêt dans la mesure où elle ne correspond pas à une stratégie d'indexation définie. Par contre, la description proposée du contenu dynamique des séquences peut s'avérer utile comme résultat intermédiaire, pouvant être fourni en entrée d'autres outils d'analyse automatique (voir, par exemple, les perspectives d'utilisation incrémentale des résultats évoquées en conclusion).

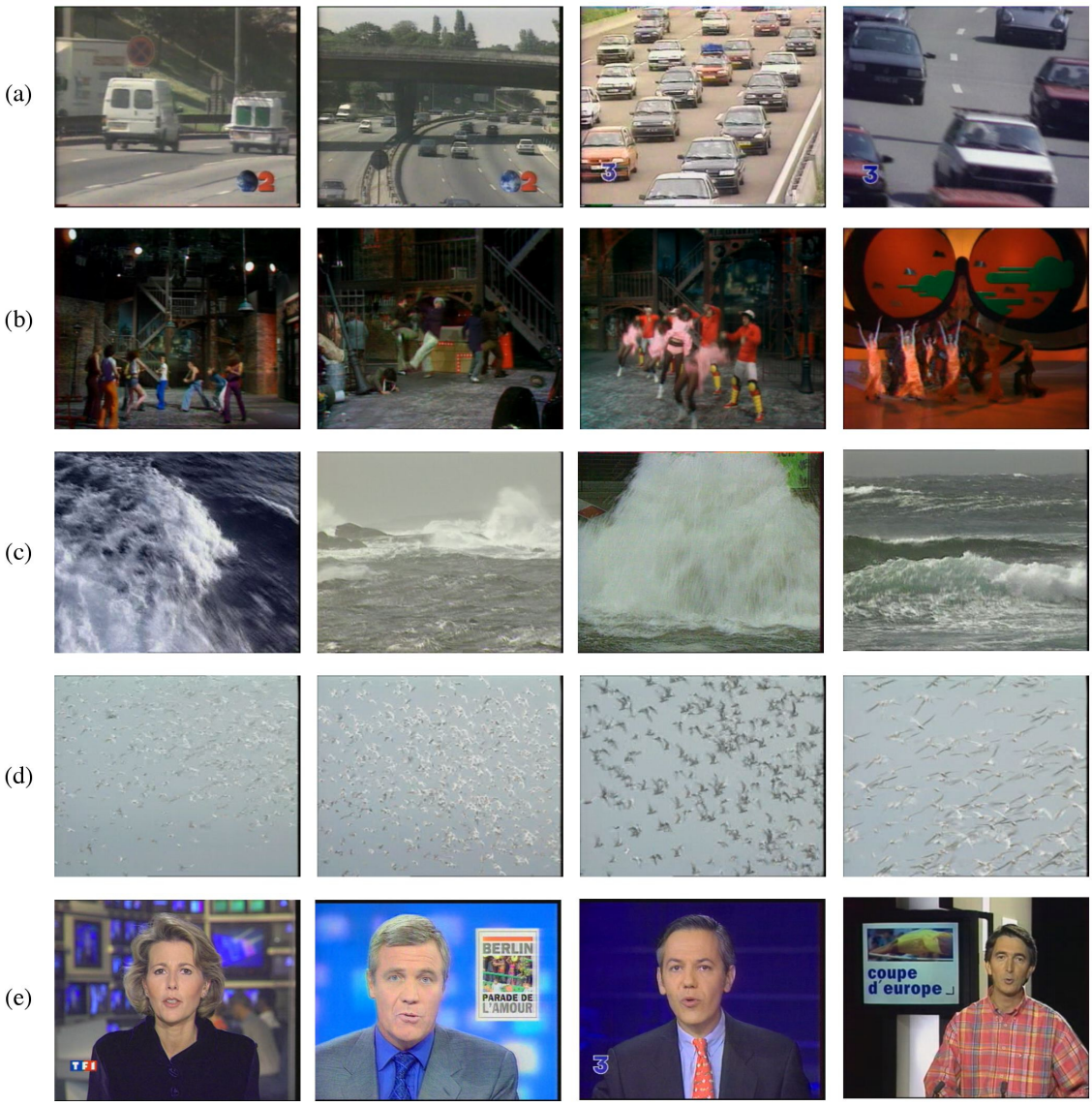



FIG. 9.4: *Exemples de séquences de la base d'expérimentation 4 : (a)* Automobile, *(b)* Ballet, *(c)* Eau, *(d)* Oiseau, *(e)* Studio.


FIG. 9.5: *Exemples de séquences de la base d'expérimentation 5 : (a)* Football, *(b)* Formule 1, *(c)* Cyclisme, *(d)* Aviron, *(e)* Patinage.

sont disponibles par classe pour l'apprentissage. Comme pour la précédente base d'exemples, les mouvements représentés au sein des classes sont très variables. Cette diversité est renforcée par les mouvements de caméra. La figure 9.5 contient des images représentatives des séquences considérées.

**En résumé,** nous proposons deux lectures des expérimentations. La première distingue les expérimentations 1 à 3 dont l'objectif est de valider nos algorithmes sur des problèmes classiques de caractérisation du contenu dynamique dans des séquences d'images (caractérisation simple de différents types de mouvement, d'activités humaines, de mouvements de caméra), et les expérimentations 4 et 5 plus proches à notre sens de la réalité des contenus dynamiques des documents audiovisuels (prise en compte de la variabilité des mouvements des objets, des mouvements de caméra, d'un usage défini).

La seconde approche concerne les expérimentations 1, 4 et 5 qui s'appuient sur des séquences issues de corpus INA, et incluent une gradation des difficultés : classification sur des classes homogènes, introduction de variabilités sensibles au sein des classes, prise en compte des mouvements de caméra.

## 9.2 Résultats expérimentaux

Une fois fixé le cadre d'évaluation, nous allons présenter les résultats obtenus sur la classification des séquences pour les descripteurs retenus. Les résultats sont introduits d'un point de vue global avant d'être détaillés pour chaque base d'expérimentations. Nous nous intéresserons aussi à la validité des apprentissages effectués pour les différents classifieurs et à l'influence du paramètre $C$, paramètre lié aux SVM et aux erreurs d'apprentissage, dont la gestion s'est avérée problématique. Nous présenterons également quelques résultats complémentaires offrant des perspectives intéressantes.

### 9.2.1 Commentaires généraux sur les résultats expérimentaux

Avant de passer aux résultats détaillés, nous allons aborder :

- les résultats obtenus avec l'ensemble des descripteurs de mouvement à travers les taux globaux de classification correcte $T_{pos}$ pour les différentes expérimentations ;
- l'influence du paramètre $C$, lié à l'apprentissage des SVM ;
- les apprentissages effectués pour les différents classifieurs, en évoquant le nombre de vecteurs de support $N_{SV}$ et le nombre de vecteurs de support "problématiques" $N_{SV}^{c}$ ;

#### *9.2.1.1 Étude des taux de classification correcte*

Les principaux résultats obtenus avec les descripteurs introduits sur les cinq bases d'expérimentation sont résumés à la figure 9.6 pour deux valeurs du paramètres $C$.

Selon les descripteurs, ces résultats sont très variables. Ainsi, les classifications réalisées à partir de la signature $Map_{MHI}$ sont plutôt satisfaisantes ($T_{pos} \in [0.75, 1.00]$), par contre celles fondées sur la signature $Cooc_{Map_{MHI}}$ sont très décevantes ($T_{pos} \in [0.00, 0.61]$). La comparaison des résultats indique l'ordre de performance décroissant suivant : $Map_{MHI}$, $Map_{STG}$, $DCT_{MHI}$, $H_{Map_{MHI}}$, $H_{Cooc_{MHI}}$ et $Cooc_{MHI}$. Il paraît surprenant que la signature $H_{Cooc_{MHI}}$ donne de meilleurs

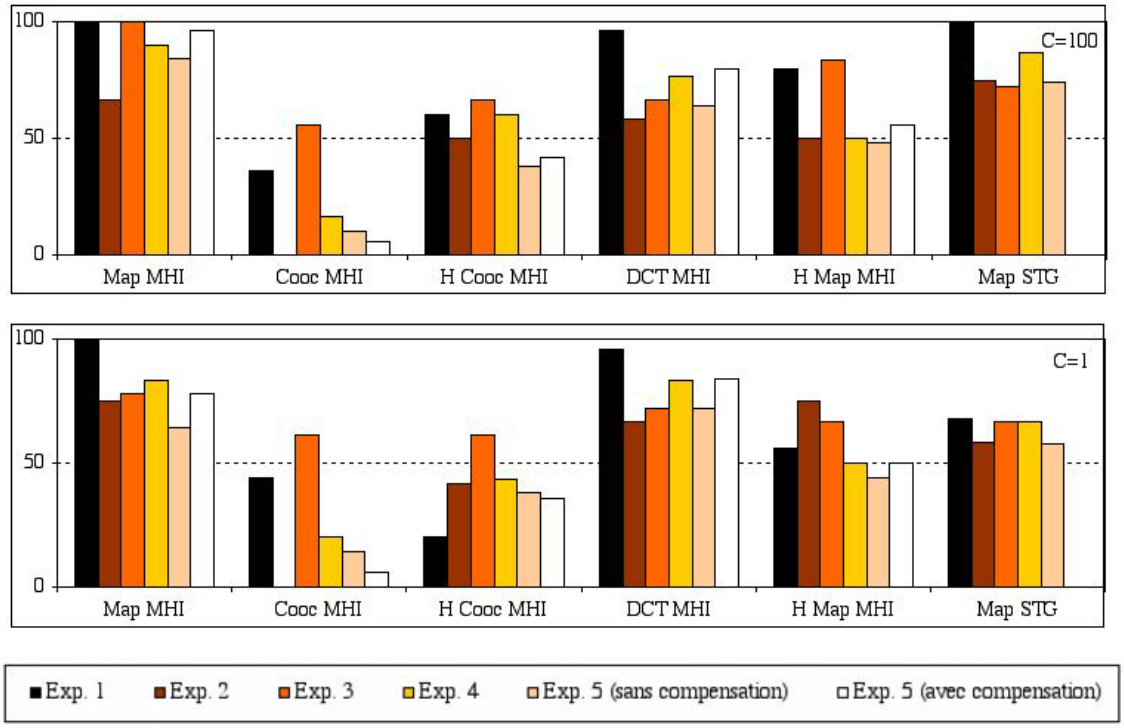



FIG. 9.6: *Taux de classification correcte $T_{pos}$ (%) sur l'ensemble des classes considérées pour les cinq bases d'expérimentations avec chacun des descripteurs, $C = 100$ (en haut) et $C = 1$ (en bas).*

résultats que la signature $Cooc_{MHI}$, dont elle est issue. Nous tenterons d'élucider ce point lors de l'étude détaillée des différents résultats.

Les résultats obtenus sur les différentes bases d'expérimentation semblent cohérents, et ceux des expérimentations 1, 4 et 5 sont en accord avec le contenu des bases d'expérimentation. En effet, à une exception près, la difficulté croissante de caractérisation du contenu dynamique (information redondante dans la première base, introduction de variations intra-classe sensibles dans la quatrième, utilisation de séquences à caméra mobile dans la cinquième) explique la hiérarchisation des résultats obtenus. Conformément à nos espoirs, dans l'expérimentation 5, les résultats sont dans leur grande majorité supérieurs, avec une compensation du mouvement de la caméra.

Les résultats obtenus avec la signature $Map_{STG}$ sont similaires ou légèrement inférieurs, à une exception près, à ceux relatifs à la signature $Map_{MHI}$. Le recours à une caractérisation du mouvement fondée sur les filtres spatio-temporels de Gabor n'a donc pas entièrement répondu à notre attente au vu des résultats sur les cinq bases d'expérimentation. En particulier, l'enrichissement de l'information de mouvement, par une prise en compte plus fine des amplitudes et des orientations, n'a pas permis une amélioration notable des résultats.

Les deux valeurs du paramètre $C$ retenues pour les expérimentations rassemblées à la figure 9.6 sont les bornes de l'intervalle "utile" défini lors de la mise en œuvre des SVM (cf. sous-section 8.4.3). Que nous fassions une lecture des résultats par signature ou par base d'expérimentation, l'influence de ce paramètre semble difficile à interpréter. Nous revenons sur cette question dans le paragraphe suivant.

#### *9.2.1.2 Influence du paramètre $C$*

Afin de compléter les résultats de la figure 9.6, de mieux comprendre l'influence du paramètre $C$ lié à l'apprentissage des SVM, et éventuellement d'en proposer une valeur par défaut, nous avons mené, pour les descripteurs fondés sur les MHI, et sur la base d'expérimentation 1, qui présente un cas de classification simple, des tests complémentaires présentés dans les tableaux 9.2 et 9.3.

| $T^l_{pos}$ (%) | $Map_{MHI}$ | $Cooc_{MHI}$ | $H_{Cooc_{MHI}}$ | $DCT_{MHI}$ | $H_{Map_{MHI}}$ |
|---|---|---|---|---|---|
| $C = 100$ | $\{100, 100, 100, 100, 100\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 100\}$ | $\{80, 80, 80, 80, 82\}$ | $\{80, 80, 80, 97, 100\}$ |
| $C = 75$ | $\{100, 100, 100, 100, 100\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 95, 100\}$ |
| $C = 50$ | $\{100, 100, 100, 100, 100\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 95, 100\}$ |
| $C = 10$ | $\{100, 100, 100, 100, 100\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 83, 100\}$ |
| $C = 5$ | $\{90, 100, 100, 100, 100\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 98\}$ |
| $C = 1$ | $\{80, 80, 80, 80, 100\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ | $\{80, 80, 80, 80, 80\}$ |

TAB. 9.2: *Influence du paramètre $C$ sur le taux de classification correcte $T^l_{pos}$ pour chacun des classifieurs sur l'ensemble d'apprentissage de l'expérimentation 1*

Lors de ces tests, les classifieurs ont été appris pour différentes valeurs de $C$ dans l'intervalle $[0, 100]$ fixé antérieurement. Le tableau 9.2 contient les résultats de la classification des échantillons d'apprentissage par chacun des cinq classifieurs utilisés. Ces tests s'avèrent cohérents avec les travaux menés dans [Pontil 98a, Sec. 3] à propos de l'influence du paramètre $C$ sur la définition de l'*OSH*, et donc sur la surface de décision associée à chacun des classifieurs. En effet, nous observons à la fois une certaine stabilité des résultats et un apprentissage plus précis lorsque $C$ augmente. Nous retrouvons donc la notion qu'une contrainte forte sur les erreurs d'apprentissage permet de définir une surface de décision plus précise sur l'ensemble d'apprentissage.

D'après [Pontil 98a], cette précision est obtenue au détriment d'une marge plus faible associée à la surface de décision et, en théorie, d'une capacité de généralisation plus faible (voir sous-section 8.4.3). Afin d'en évaluer les conséquences, nous présentons dans le tableau 9.3, les résultats obtenus pour la caractérisation des séquences de l'ensemble de test, lorsque les classifieurs sont utilisés ensemble (voir sous-section 8.3.4).

| $T_{pos}$ (%) | $Map_{MHI}$ | $Cooc_{MHI}$ | $H_{Cooc_{MHI}}$ | $DCT_{MHI}$ | $H_{Map_{MHI}}$ |
|---|---|---|---|---|---|
| $C = 100$ | 100 | 36 | 60 | 96 | 80 |
| $C = 75$ | 100 | 36 | 60 | 96 | 80 |
| $C = 50$ | 100 | 40 | 60 | 96 | 76 |
| $C = 10$ | 100 | 40 | 60 | 96 | 64 |
| $C = 5$ | 100 | 32 | 40 | 96 | 56 |
| $C = 1$ | 100 | 44 | 20 | 96 | 56 |

TAB. 9.3: *Influence du paramètre $C$ sur le taux de classification correcte $T_{pos}$ sur l'ensemble de tests de l'expérimentation 1*

La comparaison des tableaux 9.2 et 9.3 met en évidence les deux points suivants. D'une part, d'après la donnée de $T^l_{pos}$ les différents classifieurs semblent correctement appris[2]. Toutefois, des résultats plutôt stables d'un descripteur à l'autre lors de l'apprentissage ($T^l_{pos} \in [80, 100]$) correspondent à des performances contrastées ($T_{pos} \in [32, 100]$) sur la base de test. D'autre part, hormis pour la signature $Cooc_{MHI}$, le taux de classification correcte $T_{pos}$ augmente avec la valeur de $C$, ce qui semble contradictoire avec la théorie. Nous pouvons avancer une double explication pour ce

2. Nous serons amenés à revenir sur ce point plus en détail dans le paragraphe suivant.

phénomène. Lors de la caractérisation des séquences, les classifieurs sont utilisés conjointement, ce qui atténue les erreurs produites individuellement par certains classifieurs. En outre, le lien entre l'amplitude de la marge de l'*OSH* associé à un classifieur et sa capacité de généralisation n'est que théorique[3]. Dans la pratique, d'autres éléments influencent les résultats obtenus : la constitution de la base d'exemples, le choix des exemples d'apprentissage, l'espace de représentation des données, la répartition des points des différentes classes dans cet espace. Ceci pourrait expliquer aussi les résultats contradictoires observés dans certains cas à la figure 9.6.

Ainsi, pour construire des classifieurs optimaux, nous devrions, en toute rigueur, nous donner des bases d'exemples beaucoup plus larges, étudier les différents choix possibles pour les échantillons d'apprentissage afin de conserver les plus représentatifs, faire varier le paramètre $C$ pour obtenir le meilleur compromis entre précision et généralisation, et ce pour chacune des classes de nos cinq expérimentations. Compte tenu des problèmes pratiques soulevés, nous avons choisi nos bases d'apprentissage au hasard, comme indiqué à la sous-section 8.3.3. Nous avons fixé la paramètre $C$ d'après les résultats de la figure 9.6 confirmés par le tableau 9.3, dont il ressort que les performances sont globalement meilleures pour $C = 100$. Ce paramétrage a été retenu pour l'ensemble des expérimentations menées.

#### *9.2.1.3 Analyse de la phase d'apprentissage des classifieurs*

Une source d'information supplémentaire sur la construction des classifieurs et le déroulement de l'apprentissage est l'étude des vecteurs d'entrée sélectionnés pendant l'apprentissage pour être des vecteurs de support, et particulièrement de ceux qui révèlent la complexité de l'apprentissage. Nous avons étudié cette information par le biais des indicateurs suivants : le nombre de vecteurs de support $N_{SV}$ et le nombre de vecteurs de support "problématiques" $N_{SV}^c$ obtenus lors de l'apprentissage de chaque classifieur.

Rappelons qu'un vecteur de support est un vecteur d'entrée dont le multiplicateur de Lagrange associé est strictement positif ($\alpha_i > 0$). Dans le cas d'un problème de classification non séparable, les vecteurs de support "problématiques" sont ceux pour lesquels $\alpha_i = C$, c'est-à-dire les vecteurs mal classés à l'issue de l'apprentissage ou ceux qui sont à une distance de la surface de décision inférieure à la marge associée à l'*OSH* (cf. sous-section C.4.2).

L'analyse de l'ensemble des classifieurs est contenue dans le tableau 9.4. Avant d'essayer d'en tirer des enseignements, donnons quelques indications utiles :

- l'ensemble des vecteurs de support contient la totalité de l'information présente dans la base d'apprentissage. En particulier, si les autres vecteurs d'entrée étaient retirés de la base d'apprentissage, la surface de décision obtenue serait identique. Par conséquent, si $\frac{N_{SV}}{N_l}$ est faible, cela signifie que l'information contenue dans la base d'apprentissage est très redondante. Dans le cas contraire, la définition de la surface de décision est fondée sur un grand nombre de points. Parmi les causes qui nous paraissent propices à l'obtention d'un grand nombre de vecteurs de support, citons la dimension de l'espace de représentation des vecteurs d'entrée[4]

3. En effet, si on se réfère à la relation C.1 de l'annexe C, le paramètre $C$ influe sur le compromis entre le risque empirique et l'intervalle de confiance. Le principe du SRM permet de borner le risque d'erreur avec une certaine probabilité, et non de s'assurer d'une capacité de généralisation donnée.

4. Dans [Pontil 98a], l'auteur rappelle notamment que pour un espace de représentation de dimension $n$, l'OSH est entièrement défini par $n - 1$ vecteurs de support non "problématiques". Dans le cas de l'utilisation de noyaux non linéaires, et plus précisément des RBF dont la dimension VC est infinie, l'auteur indique qu'un nombre fini de vecteurs de support non "problématiques" ne serait pas suffisant pour déterminer entièrement l'*OSH*. Nous n'avons pas trouvé d'indications claires concernant l'influence de la dimension de l'espace de représentation originel sur la sélection des vecteurs de support lorsqu'un noyau de dimension VC infinie est utilisé. Toutefois, il semble que des

| Expérimentation | | Exp. 1 | Exp. 2 | Exp. 3 | Exp. 4 | Exp. 5 |
|---|---|---|---|---|---|---|
| $N_l$ | | 60 | 18 | 36 | 70 | 100 |
| $Map_{MHI}$ | $N_{SV}$ | $\{37, 23, 32, 46, 49\}$ | $\{13, 10, 13, 12, 15, 10\}$ | $\{24, 32, 17\}$ | $\{58, 45, 57, 49, 31\}$ | $\{45, 66, 82, 45, 60\}$ |
| | $N^c_{SV}$ | $\{0, 0, 0, 0, 0\}$ | $\{0, 0, 0, 0, 0, 0\}$ | $\{0, 0, 0\}$ | $\{0, 0, 0, 0, 0\}$ | $\{0, 0, 0, 0, 0\}$ |
| $Cooc_{MHI}$ | $N_{SV}$ | $\{40, 34, 36, 38, 28\}$ | $\{10, 10, 9, 11, 10, 12\}$ | $\{32, 26, 27\}$ | $\{40, 40, 59, 43, 38\}$ | $\{56, 58, 61, 57, 55\}$ |
| | $N^c_{SV}$ | $\{15, 18, 13, 14, 19\}$ | $\{3, 3, 4, 3, 4, 3\}$ | $\{18, 21, 23\}$ | $\{21, 22, 18, 20, 24\}$ | $\{25, 25, 21, 28, 31\}$ |
| $H_{Cooc_{MHI}}$ | $N_{SV}$ | $\{27, 22, 24, 25, 26\}$ | $\{8, 7, 4, 7, 7, 6\}$ | $\{24, 25, 12\}$ | $\{33, 12, 32, 32, 30\}$ | $\{40, 59, 50, 42, 54\}$ |
| | $N^c_{SV}$ | $\{22, 20, 24, 23, 23\}$ | $\{5, 5, 4, 5, 5, 6\}$ | $\{24, 23, 10\}$ | $\{23, 12, 25, 24, 27\}$ | $\{40, 36, 36, 39, 35\}$ |
| $DCT_{MHI}$ | $N_{SV}$ | $\{28, 25, 29, 28, 31\}$ | $\{9, 8, 9, 10, 11, 9\}$ | $\{24, 27, 24\}$ | $\{38, 35, 37, 33, 30\}$ | $\{38, 49, 57, 44, 44\}$ |
| | $N^c_{SV}$ | $\{20, 23, 20, 20, 20\}$ | $\{4, 5, 4, 3, 3, 5\}$ | $\{24, 20, 24\}$ | $\{21, 23, 22, 24, 23\}$ | $\{38, 31, 30, 37, 37\}$ |
| $H_{Map_{MHI}}$ | $N_{SV}$ | $\{26, 5, 16, 25, 25\}$ | $\{8, 6, 6, 8, 10, 8\}$ | $\{20, 25, 3\}$ | $\{30, 24, 33, 30, 23\}$ | $\{22, 43, 66, 35, 42\}$ |
| | $N^c_{SV}$ | $\{23, 2, 13, 22, 22\}$ | $\{5, 6, 6, 4, 5, 5\}$ | $\{17, 23, 0\}$ | $\{27, 24, 25, 25, 20\}$ | $\{19, 37, 26, 32, 38\}$ |
| $Map_{STG}$ | $N_{SV}$ | $\{34, 16, 29, 45, 40\}$ | $\{14, 6, 6, 12, 12, 8\}$ | $\{24, 34, 14\}$ | $\{46, 44, 51, 54, 26\}$ | $\{39, 66, 64, 30, 40\}$ |
| | $N^c_{SV}$ | $\{0, 0, 2, 0, 7\}$ | $\{0, 1, 0, 2, 4, 1\}$ | $\{19, 20, 0\}$ | $\{0, 0, 1, 8, 1\}$ | $\{0, 0, 0, 0, 0\}$ |

TAB. 9.4: *Analyse des vecteurs de support des classifieurs sur l'ensemble des expérimentations. $N_l$ est le nombre d'échantillons dans l'ensemble d'apprentissage, $N_{SV}$ le nombre de vecteurs de support, et $N^c_{SV}$ le nombre de vecteurs de support "problématiques".*

et la complexité du problème à traiter [5] ;

- les vecteurs de support "problématiques" sont des vecteurs mal classés ou proches de la surface de décision. Si $\frac{N^c_{SV}}{N_{SV}}$ est élevé, c'est le signe que le problème de classification est complexe et/ou qu'aucune solution satisfaisante n'a été trouvée lors de l'apprentissage. *A contrario*, si les données sont séparables, les classifieurs seront parfaitement appris sur l'ensemble d'apprentissage et nous aurons $N^c_{SV} = 0$.

Dans le tableau 9.4, le rapport $\frac{N_{SV}}{N_l}$ est globalement élevé (souvent supérieur ou égal à $\frac{1}{2}$, parfois supérieur à $\frac{3}{4}$), ce qui semble montrer que l'information est relativement peu redondante dans les bases d'apprentissage. Ceci pourrait indiquer que des bases d'exemples plus larges auraient été nécessaires.

Par ailleurs, sur toutes les bases d'expérimentation, les données d'apprentissage issues de la signature $Map_{MHI}$ semblent être séparables ($N^c_{SV} = 0$). Pour cette signature, nous pouvons donc attribuer les erreurs de caractérisation des séquences de tests à un défaut de généralisation des classifieurs. Toutefois, si $N^c_{SV}$ est nul, $N_{SV}$ reste relativement élevé. Chaque point de l'image est associé à une dimension de l'espace de représentation initial des données ; nous pouvons donc faire l'hypothèse que les variabilités seront importantes au sein d'une même classe sur de nombreuses dimensions. La complexité du problème de classification traité est sans doute une explication des valeurs $N_{SV}$ obtenues. La situation est similaire pour près de la moitié des classifieurs construits à partir de la signature $Map_{STG}$.

Pour les autres signatures, les rapports $\frac{N_{SV}}{N_l}$ et $\frac{N^c_{SV}}{N_{SV}}$ semblent globalement cohérents avec les résultats obtenus lors de la caractérisation des séquences (figure 9.6). En effet, lorsque les résultats de la caractérisation $T_{pos}$ sont faibles, en général, les deux rapports étudiés tendent à croître. Cependant, il convient de noter que les variations des taux entre signatures sont parfois peu sensibles, même si les résultats de la caractérisation sont très différents et que les rapports $\frac{N^c_{SV}}{N_{SV}}$ sont très élevés, voire proches de 1. Les écarts entre l'apprentissage et les résultats de la caractérisation nous

données représentées dans un espace de petite dimension occuperont un sous-espace plus réduit de l'espace d'arrivée et corresponderont *a priori* à un problème de classification plus simple dans l'espace de projection.

5. La complexité du problème traité dépend notamment de l'information synthétisée par les différents descripteurs. Ceux-ci étant représentés dans des espaces de grande dimension, leur visualisation, et par conséquent la compréhension de la complexité du problème de classification induit, restent problématiques.

renvoient aux explications données plus haut sur la généralisation. Le nombre important de vecteurs de support "problématiques" constitue une indication intéressante. Il semble que pour les signatures autres que $Map_{MHI}$, la détermination de la surface de décision (l'$OSH$) ait été problématique et ait laissé un grand nombre de points mal classés ou proches de celle-ci. Les données d'apprentissage ayant été difficilement apprises par les classifieurs, les résultats sur la caractérisation de séquences par l'utilisation conjointe de ceux-ci sont donc très dépendants de la configuration globale et respective des ensembles d'apprentissage et de test. À titre d'exemple, les bons résultats obtenus par la signature $DCT_{MHI}$ sur la base d'expérimentation 1 doivent être considérés avec davantage de circonspection à la lumière des informations disponibles sur l'apprentissage réalisé ($\frac{N_{SV}^c}{N_{SV}} \in [\frac{2}{3}, 1[$).

Avant de passer à l'étude détaillée des résultats, soulignons que les taux de classification correcte $T_{pos}$ obtenus paraissent plutôt satisfaisants. Si nous avons insisté sur certaines des difficultés rencontrées et sur quelques problèmes apparus lors des expérimentations, notamment dans l'utilisation des SVM, il reste qu'avec $C = 100$, le taux de bonne classification $T_{pos}$ est au-dessus de 60% pour près des $\frac{3}{5}$ des expérimentations, et au-dessus de 75% pour près de la moitié de celles-ci (cf. figure 9.6).

### 9.2.2 Commentaires détaillés des résultats expérimentaux

#### *9.2.2.1 Résultats de l'expérimentation 1*

Les matrices de confusion et les taux de classification correcte sont donnés, dans l'annexe D, aux tableaux D.1 à D.6 pour l'ensemble des primitives.

Les résultats obtenus avec la signature $Map_{MHI}$ (tableau D.1) sont excellents, l'étude de chacun des classifieurs et de leur apprentissage le confirme. Toutes les séquences de tests ont été correctement caractérisées.

Les résultats pour la signature $Cooc_{MHI}$ sont décevants (tableau D.2) : seules les classes *Oiseau* et *Danseur* sont correctement reconnues. L'étude des classifieurs montre que l'apprentissage n'a pas été concluant et confirme les observations faites à partir du tableau 9.4. En particulier, la réponse moyenne $V_{Cl_i}^l(C_i)$ du classifieur $Cl_i$ sur l'ensemble des exemples d'apprentissage de la classe $C_i$ est négative, alors qu'elle devrait être positive à l'issue de l'apprentissage. Ceci nous amène à nous interroger sur la validité de l'information contenue dans la signature $Cooc_{MHI}$, et plus précisément sur la réduction d'information à laquelle nous avons procédé. Ainsi, l'évolution temporelle des MHI réduite à un vecteur de taille 29 sans localisation spatiale n'est pas une bonne description du mouvement apparent.

Il est surprenant, *a priori*, que la signature $H_{Cooc_{MHI}}$ issue de la précédente donne de meilleurs résultats (tableau D.3). Les réponses $V_{Cl_i}^l(C_j)$ de chaque classifieur $Cl_i$ sur les exemples de la base d'apprentissage montrent à nouveau que le problème a été mal appris. Toutefois, à la différence de la signature précédente, la réponse moyenne $V_{Cl_i}^l(C_i)$ sur la classe $C_i$ est plus élevée que les réponses $V_{Cl_i}^l(C_j)_{j\neq i}$ du classifieur $Cl_i$ sur le "reste des classes", ce qui explique les résultats corrects obtenus lorsque les classifieurs sont utilisés ensemble (voir la relation 8.7). Nous ne pouvons faire que des conjectures pour expliquer que le problème de classification semble relativement plus simple dans le cas présent qu'avec $Cooc_{MHI}$. Les plus probables intuitivement sont que l'information extraite par les descripteurs d'Haralick est sans doute plus discriminante que celle contenue dans la matrice de cooccurrence, ou que l'espace de représentation est plus favorable dans le cas des descripteurs d'Haralick. Nous n'avons cependant pas de moyens de vérifier ces hypothèses.

Les résultats obtenus avec la signature $DCT_{MHI}$ semblent tout à fait satisfaisants (tableau D.4) : toutes les séquences ont été correctement caractérisées sauf une séquence de la classe *Voiture*. Toutefois, l'étude de l'apprentissage des différents classifieurs montre des résultats d'apprentissage

identiques aux deux cas précédents (réponse globalement négative du classifieur sur les exemples d'apprentissages de la classe à laquelle il est dédié), ce qui était déjà apparu dans l'étude du tableau 9.4. Les données n'ont donc pas permis un apprentissage correct des classifieurs ; seule la plus grande proximité des échantillons d'une classe $C_i$ à la surface de décision associée au classifieur $Cl_i$ par rapport aux exemples des autres classes nous assure les résultats satisfaisants obtenus lors de la caractérisation des séquences.

Le cas de figure est d'ailleurs identique pour trois des classifieurs construits à partir de la signature $H_{Map_{MHI}}$. L'utilisation conjointe des classifieurs produit cependant davantage de confusions entre classes que dans le cas précédent (cf. tableau D.5).

Les classifieurs construits à partir de la la signature $Map_{STG}$ opèrent une classification sans erreur sur l'ensemble de tests. Chaque classifieur, y compris ceux présentant un nombre de vecteurs de support "problématiques" $N^c_{SV}$ non nuls, ont été correctement appris, d'après l'étude de leurs réponses moyennes $V^l_{Cl_i}(C_j)$ sur les échantillons d'apprentissage. Cette première comparaison entre les descripteurs $Map_{MHI}$ et $Map_{STG}$ indique que la contrepartie à l'enrichissement de l'information de mouvement en amplitude et en orientation dans la signature $Map_{STG}$ est une plus grande variabilité des données. Nous pouvons faire l'hypothèse que le problème de classification devient plus complexe, et que le nombre de séquences d'apprentissage n'est pas suffisant. Cette hypothèse expliquerait les valeurs non nulles de $N^c_{SV}$ pour deux des classifieurs fondés sur $Map_{STG}$.

Compte tenu de la perception que nous avons de ces différents types de mouvements, nous nous étions attendus à une confusion entre les classes *Océan* et *Oiseau*, celles-ci pouvant être assimilées à des textures en mouvement. Aucune explication de ce type ne paraît pouvoir être apportée ici.

En résumé, nous retenons des résultats obtenus sur cette première expérimentation les idées suivantes :

- une première validation des descripteurs indique l'ordre d'efficacité décroissant[6] : $Map_{MHI}$, $Map_{STG}$, $DCT_{MHI}$, $H_{Map_{MHI}}$, $H_{Cooc_{MHI}}$ et $Cooc_{MHI}$. En particulier, dans le cas de la signature $Cooc_{MHI}$, l'information spatiale perdue ne semble pas compensée par l'information temporelle plus riche introduite dans celle-ci ;

- sur les trois descripteurs dont les taux de classification correcte sont supérieurs à 90%, seul le descripteur $Map_{MHI}$ permet un apprentissage satisfaisant de tous les classifieurs ;

- les difficultés d'apprentissage observées pour les autres classifieurs fondés sur les MHI montrent que les réductions mises en œuvre sur les données tendent à rendre l'information plus confuse (car moins riche) en entrée des SVM, le problème à traiter est alors complexe, en dépit de la définition artificiellement simple des classes. Une explication similaire peut être apportée pour le descripteur $Map_{STG}$, l'éventuelle complexité étant due à l'enrichissement de l'information et non à sa réduction. Nous avons confirmation, par la même occasion, de la capacité des SVM à accepter en entrée des données associées à des espaces de grande dimension ;

- lorsque le problème de classification s'avère complexe, il semble que les échantillons d'apprentissage soient trop peu nombreux. Nous aurions aimé remédier à cela dans les expérimentations suivantes. Malheureusement, les expérimentations 2 et 3 sont menées sur des bases d'exemples extérieures à nos corpus que nous n'avons pu augmenter. Pour les expérimentations 4 et 5, nous avons sensiblement augmenté le nombre de séquences disponibles pour chaque classe (voit tableau 9.1). Toutefois, la complexité de la caractérisation augmentant nettement elle aussi, il aurait fallu idéalement disposer de bases d'expérimentation de l'ordre du millier d'échantillons. Cela ne nous a pas été possible pour des raisons d'ordre pratique ;

6. avec des réserves évoquées à la sous-section 9.2.1 quant à la validité des apprentissages.

– les difficultés observées lors de l'apprentissage de certains classifieurs sont atténuées par l'utilisation coopérative qui est faite de ceux-ci. Ainsi, il suffit pour effectuer la classification des séquences que le classifieur $Cl_i$ réponde plus fortement aux exemples de la classe $C_i$ que les autres classifieurs. Lorsque ce n'est pas le cas, nous pouvons imaginer que les vecteurs d'entrées des différentes classes apparaîtraient très "mélangés" dans leur espace de représentation respectif, ce qui nous renvoie, dans ce problème simple de classification, à la validité de nos méthodes de réduction de l'information.

Ces résultats sont confirmés, en général, dans les autres expérimentations (cf. figure 9.6). Toutefois, compte tenu du caractère expérimental de notre démarche, nous avons tenu à présenter, par la suite, les résultats obtenus à partir de tous les descripteurs mis en œuvre, y compris ceux qui ne semblent pas validés par cette première expérimentation.

#### *9.2.2.2 Résultats de l'expérimentation 2*

Sur cette base de séquences, nous souhaitons évaluer nos algorithmes sur un exemple classique de caractérisation du contenu dynamique: la reconnaissance d'activité humaine. Les matrices de confusion et les taux de classification correcte sont présentés dans les tableaux D.7 à D.12 de l'annexe D pour l'ensemble des descripteurs.

Nous observons une chute importante des taux de classification correcte $T_{pos}$ pour l'ensemble des descripteurs. Deux explications peuvent être avancées pour rendre compte de cette dégradation des résultats.

La première est le nombre sans aucun doute insuffisant d'échantillons disponibles pour l'apprentissage. Nous avons déjà abordé ce point ; avec trois exemples par classe dans la base d'apprentissage, nous atteignons sans doute un seuil critique. La seconde est liée à l'absence de normalisation dans nos algorithmes de la durée des actions effectuées par les différents sujets, ce point étant moins sensible pour les autres bases de séquences. En effet, les déplacements sont effectués en des temps différents en fonction des actions (en l'occurrence, il est plus long de monter l'escalier que de traverser la pièce), et selon les individus. Nous avions tenté de pallier cela en prenant $V_{max} = 24$, soit le mimimum des temps d'actions observés afin de considérer l'information sur l'horizon temporel le plus large possible, mais ces séquences sont constituées d'extraits de déplacements qui ne sont ni normalisés ni systématiquement alignés temporellement. Ceci doit sensiblement augmenter la variabilité des séquences au sein des classes de déplacement.

Notons que l'ordre de performance des signatures est sensiblement le même que dans la précédente expérimentation, que les résultats sur la signature $Cooc_{MHI}$ sont mauvais, et que les performances des descripteurs d'Haralick qui en sont issus sont notablement meilleurs.

Nous pouvons envisager comme erreurs acceptables les confusions suivantes :

1. les confusions liées au sens du mouvement : *S'approcher*/*S'éloigner*, *Descendre*/*Monter*, et *À gauche*/*À droite* ;

2. les confusions liés à des composantes directionnelles communes : *À droite*, *S'approcher*, *Monter* ont une composante de mouvement apparent vers la droite, et *À gauche*, *S'éloigner*, *Descendre* une composante vers la gauche (voir [Chomat 00]).

D'après les matrices de confusion des descripteurs fondés sur les MHI (tableaux D.7 à D.11), 58% des erreurs observées sont de type 1 et 27% de type 2. Ces résultats sont cohérents avec nos remarques concernant la relative pauvreté de l'information de directionnalité apportée par les MHI

(cf. sous-section 8.2.2). *A contrario*, les bons résultats relatifs de la signature $Map_{STG}$ peuvent être le signe de la prise en compte de cette directionnalité des mouvements au sein de la signature.

De cette deuxième expérimentation, nous retiendrons le défaut de nos algorithmes de ne pas incorporer une normalisation et un alignement temporel des séquences [7]. Notons, de plus, la confirmation des limites des MHI dans la description du mouvement apparent, et l'éventuel apport de l'utilisation des filtres de Gabor spatio-temporels.

#### *9.2.2.3 Résultats de l'expérimentation 3*

Cette base est dédiée à la caractérisation des mouvements de caméra. Trois types de déplacements de caméra au-dessus d'un motif texturé sont considérés. Les matrices de confusion et les taux des classification correcte sont présentés dans les tableaux D.13 à D.18 de l'annexe D pour l'ensemble des descripteurs.

Les résultats obtenus sont plutôt satisfaisants : les taux de classification correcte $T_{pos}$ sont parmi les meilleurs obtenus pour l'ensemble des expérimentations. L'ordre de performance des signatures reste sensiblement similaire, à l'exception des signatures $DCT_{MHI}$ et $Map_{STG}$ dont les résultats sont relativement moins bons pour cette tâche. Il est possible que les filtrages spatiaux nécessaires lors de l'extraction de ces descripteurs ne permettent pas de rendre compte de manière satisfaisante du mouvement de la texture du motif.

Lorsque nous analysons les confusions produites, les séquences de la classe *Rotation* sont fréquemment associées à la classe *Divergence*, et, dans une moindre mesure, celles de la classe *Divergence* à la classe *Translation*.

Ces bons résultats n'ont pas été confirmés par les expérimentations menées sur les séquences de la base 5. Dans ce dernier cas, les mouvements de caméra sont nettement plus complexes et deviennent moins perceptibles du fait de la présence d'objets en mouvement. C'est pourquoi nous avons été amenés à intégrer un module de compensation du mouvement dominant afin d'atténuer le bruit provoqué par le mouvement apparent des décors (voir paragraphe 8.2.1.4).

#### *9.2.2.4 Résultats de l'expérimentation 4*

Les classes de séquences dans cette expérimentation reprennent les types de mouvement pris en compte dans l'expérimentation 1. Toutefois, chaque classe est représentée par plus d'échantillons, et inclut des apparences de mouvement plus diversifiées. Les matrices de confusion et les taux des classifications correctes sont regroupés dans les tableaux D.19 à D.24 de l'annexe D pour l'ensemble des descripteurs.

Si les résultats obtenus sont inférieurs à ceux de l'expérimentation 1, peu de nouveaux éléments viennent nourrir nos réflexions. Les descripteurs $Map_{MHI}$, $H_{Cooc_{MHI}}$ et $Map_{STG}$ résistent plutôt bien à l'introduction d'une plus grande variabilité dans les séquences, tandis que les autres subissent des baisses plus sensibles de leurs performances.

La relative stabilité des résultats peut sans doute être mise au crédit des SVM. Cette technique de classification s'avère robuste à l'introduction de variabilités sensibles au sein des classes et les SVM semblent conserver une partie importante de leur capacité de généralisation dans un cas complexe.

Comme lors de l'expérimentation 1, les classifieurs, sauf ceux fondés sur les signatures $Map_{MHI}$ et $Map_{STG}$, semblent, dans la plupart des cas, difficilement appris. Nous retrouvons notamment des

7. Dans le mesure où cela n'est problématique que pour l'expérimentation 2, nous n'avons pas envisagé, dans le cadre des travaux présentés dans cette étude, de modifier nos algorithmes.

réponses négatives des classifieurs sur les exemples d'apprentissage issus de la classe à laquelle ils sont associés. Nous espérions corriger ce problème en augmentant le nombre d'exemples disponibles, mais cela n'a pas été suffisant compte tenu de la plus grande complexité du problème de classification à traiter. Nous retrouvons ces problèmes d'apprentissage dans l'étude du tableau 9.4, où l'on note que l'accroissement de $N_l$ est, en général, concomitant avec ceux de $N_{SV}$ et $N^c_{SV}$.

Nous retrouvons aussi que de bons résultats sont obtenus avec la signature $Map_{MHI}$. Dans ce cas, les erreurs de caractérisation des séquences sont attribuables à des difficultés de généralisation, tous les classifieurs ayant été parfaitement entraînés lors de l'apprentissage. L'ordre de performance des différents descripteurs reste globalement inchangé, si ce n'est le bon comportement de la signature $H_{Cooc_{MHI}}$ dont les résultats sont identiques à ceux obtenus sur la base d'expérimentation 1.

Lorsqu'on observe les confusions entre classes, il apparaît que les séquences issues des classes *Oiseau* et *Studio* sont celles sur lesquelles il y a le moins d'erreurs de classification. Ceci peut s'expliquer par la constitution de ces deux classes, dont l'apparence du mouvement semble relativement plus stable que dans les autres cas (voir figure 9.4). Lorsqu'une séquence est mal caractérisée, elle est souvent attribuée à la classe *Studio*. En particulier, il semble qu'il y ait une confusion répétée entre les classes *Studio* et *Ballet*. Il est vrai, lorsque nous visualisons les MHI des séquences de ballets, que peu d'information de mouvement y est présente lorsque les danseurs sont filmés en plan large. Notons que les confusions sont, malgré tout, assez largement distribuées.

Nous retiendons donc que les résultats restent satisfaisants après introduction d'une variabilité sensible dans les classes de séquences, que la capacité de généralisation des SVM est notable, et qu'en particulier les signatures $Map_{MHI}$ et $Map_{STG}$ offrent de bonnes performances.

#### *9.2.2.5 Résultats de l'expérimentation 5*

Cette dernière expérimentation porte sur des séquences de sports et nous permet de valider nos algorithmes sur une application liée à un problème d'indexation de documents sportifs. Cela permettra également d'évaluer la robustesse des descripteurs introduits dans le cas d'importants déplacements de caméra. Les matrices de confusion et les taux des classifications correctes sont rassemblés dans les tableaux D.25 à D.30 de l'annexe D pour l'ensemble des descripteurs.

Avec ou sans compensation du mouvement dominant, cette expérimentation confirme les bonnes performances des signatures $Map_{MHI}$ et $Map_{STG}$, l'ordre de performance déjà constaté pour les différentes signatures, et les problèmes d'apprentissage pour les classifieurs fondés sur les signatures autres que $Map_{MHI}$ et $Map_{STG}$. Les taux de classification correcte $T_{pos}$ sont inférieurs à ceux de l'expérimentation 4, et meilleurs avec compensation du mouvement dominant que sans compensation de ce dernier. Les résultats confirment aussi les remarques énoncées à propos de l'expérimentation 4 : l'augmentation du nombre d'exemples d'apprentissage a permis de mieux appréhender la complexité croissante du problème de classification, mais n'a pas été suffisante pour permettre un entraînement correct de tous les classifieurs. De même, la robustesse des SVM à la variabilité des classes est réaffirmée.

Lorsque le mouvement dominant n'est pas compensé, les taux de classification correcte sont de 84%, 74% et 64%, respectivement pour les signatures $Map_{MHI}$, $Map_{STG}$ et $DCT_{MHI}$. Ces résultats nous permettent de valider une partie de nos choix algorithmiques sur une base de séquences issues des corpus de l'INA et liée à un usage d'indexation établi (identifier différents types de sport dans un document télévisé).

L'étude des confusions entre classes montre que les plus difficilement reconnues sont les classes *Cyclisme* et *Formule 1*, respectivement confondues avec *Aviron* et *Football*. Cela est sans doute dû

à une plus grande variation des conditions de prises de vue dans les séquences représentatives de ces deux sports.

La compensation du mouvement dominant conduit à de meilleurs résultats sauf pour la signature $Cooc_{MHI}$. La compensation du mouvement dominant est particulièrement pertinente pour les classes *Formule 1* et *Football*, qui sont deux fois mieux reconnues. Il ne semble pourtant pas que les mouvements de caméra dans les séquences relevant de ces deux sports soient moins complexes ou plus faciles à compenser. L'amélioration des résultats indique aussi que, lorsque les informations du mouvement des objets et de la caméra sont mêlées au sein des signatures, le mouvement de caméra est davantage un facteur de bruit que discriminant.

Nous pouvons donc faire valoir les résultats satisfaisants obtenus avec certains descripteurs sur cet exemple réel d'indexation, et valider l'utilité de la compensation des mouvements de caméra. Le recours aux MHI donne des résultats intéressants, mais leur utilisation hors du cadre "fond fixe-caméra fixe" reste problématique et leur capacité à intégrer l'information du mouvement de caméra de façon discriminante semble faible. Une signature intégrant séparément les informations de mouvement des objets et de la caméra serait sans doute plus pertinente au vu de cette expérimentation.

## 9.3 Résultats complémentaires : vers la détection d'événements dans la vidéo

En complément des expérimentations analysées à la section 9.2, nous avons étudié, sur des séquences plus longues, l'évolution temporelle des descripteurs d'Haralick calculés sur la signature $Map_{MHI}$ (sans compensation du mouvement de la caméra). Parmi les onze descripteurs retenus (voir paragraphe 8.2.1.3), deux nous ont semblé présenter des profils d'évolution intéressants. Il s'agit de l'énergie, $f_1$, et de l'entropie des MHI, $f_9$. Nous commenterons les résultats les plus significatifs, obtenus avec le descripteur d'entropie sur trois séquences, d'une durée d'environ une minute. Ces séquences[8] correspondent respectivement à un saut en longueur (figure 9.7), un saut à la perche (figure 9.8) et un départ de course (figure 9.9) - en l'occurrence, un faux départ.

Sur ces trois exemples, nous notons que des variations fortes de l'entropie sont dues à des ruptures de plans. Dans le cadre d'un traitement automatique, il serait envisageable de les isoler à l'aide d'un algorithme de détection automatique des changements de plans. Si nous faisons abstraction de ces variations spécifiques, l'étude du profil des courbes ouvre quelques perspectives prometteuses, même si une mise en œuvre effective reste à valider :

1. Les pics principaux de l'entropie correspondent aux moments d'intérêt dans les séquences étudiées : course et saut sur la figure 9.7, saut sur la figure 9.8, départ de la course sur la figure 9.9. Cette impression doit être relativisée par la forte entropie liée aux gros plans, toutefois un traitement spécifique de ces segments reste envisageable dans le cadre de l'analyse automatique de la vidéo. Une utilisation simple pourrait être un seuillage du critère, qui dans les exemples présentés, permettrait aux documentalistes de se positionner rapidement sur des segments d'intérêt ;

2. La courbe d'évolution de ces descripteurs lors des ralentis présente une forte instabilité locale et/ou une baisse durable de la valeur du descripteur pour $f_9$. Cette instabilité est certainement due à la succession d'images en mouvement et d'images fixes dans les ralentis notée dans [Kobla 99]. Cette caractéristique devrait permettre, après modélisation ou après apprentissage

8. Celles-ci sont extraites du document *munich2*, le saut en longueur correspond aux images 10850 à 12474, le saut à la perche aux images 39890 à 41370, et le départ de course aux images 41955 à 43215.

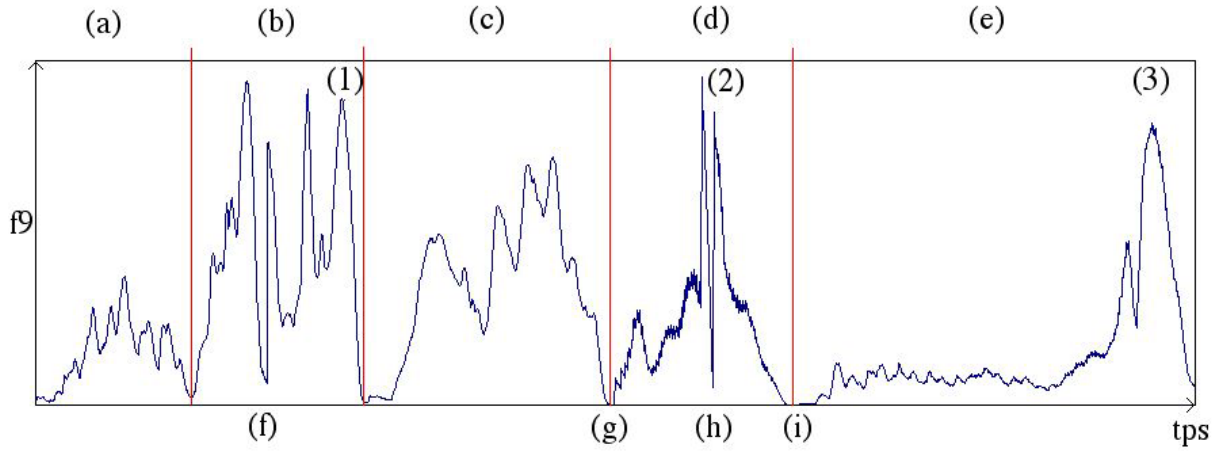



FIG. 9.7: *Évolution du descripteur $f_9$ (entropie) calculé sur la signature $Map_{MHI}$ pour une séquence de saut en longueur : (a) échauffement, (b) course d'élan, (c) sortie de l'athlète et* zoom *de la caméra , (d) premier ralenti (caméra latérale), (e) deuxième ralenti (caméra frontale), (f) rupture de plan, (g) transition progressive, (h) double rupture de plan, (i) transition progressive, (1-3) sauts.*

dans le cadre que nous avons proposé, de retrouver les séquences de ralentis. Dans certains cas, lorsque le ralenti est présenté sous le même angle de vue que l'épreuve, nous retrouvons un motif temporel similaire mais atténué, décalé dans le temps (par exemple, les zones (d) des figures 9.7 et 9.8), ce qui constitue un sujet d'étude potentiel ;

3. Un dernier point à noter, qu'il serait envisageable de prendre en compte dans la stratégie de classification que nous avons proposée, est que certains profils de courbes correspondent à des événements notables (faux-départ, saut, etc.) ou à des prototypes d'épreuves si l'on souhaite considérer un horizon temporel plus large. La détection de tels événements reste conditionnée par une étude extensive, sur des exemples plus diversifiés, de la reproductibilité des profils de courbes observés ici.

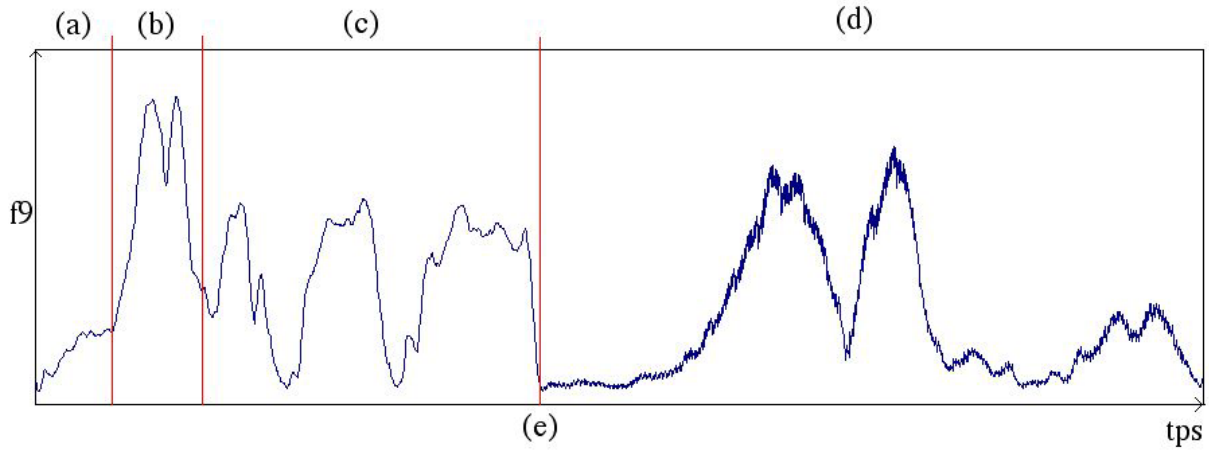



FIG. 9.8: *Évolution du descripteur $f_9$ (entropie) calculé sur la signature $Map_{MHI}$ pour une séquence de saut à la perche : (a) course d'élan, (b) saut, (c) sortie de l'athlète et* zoom *de la caméra, (d) ralenti, (e) transition progressive.*

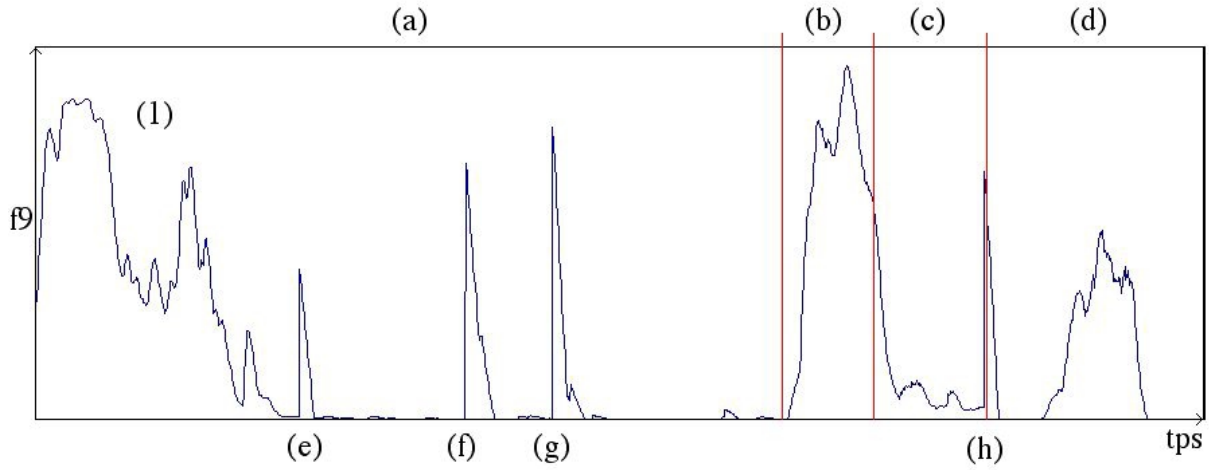



FIG. 9.9: *Évolution du descripteur $f_9$ (entropie) calculé sur la signature $Map_{MHI}$ pour un faux-départ pour une course : (a) échauffements des athlètes, (b) départ, (c) arrêt de la course, (d) ralenti, (e-h) rupture de plans, (1)* zoom *de la caméra et gros plan sur les athlètes.*

# Conclusion sur la caractérisation de séquences courtes

L'étude décrite dans cette partie était clairement exploratoire et expérimentale. Nous avons essayé d'étendre la définition et l'emploi de descripteurs conçus et utilisés généralement pour d'autres contextes (les images de l'historique du mouvement, les filtres spatio-temporels de Gabor), et nous avons mis en œuvre des méthodes de classification récentes (machines à vecteurs de support), dans des contextes relativement nouveaux pour lesquels nous avons été confrontés à des problèmes de paramétrisation. L'évaluation méthodique menée a permis de valider certains résultats, d'identifier de nombreuses difficultés, et d'ouvrir des perspectives de recherche.

## Bilan des résultats obtenus

**Comparaison avec d'autres méthodes.** Il convient d'être prudent sur la validité des comparaisons possibles : les cadres d'expérimentation, les domaines applicatifs, et les méthodes d'évaluation varient d'un auteur à l'autre. Certains auteurs caractérisent des séquences, d'autres les comparent à un modèle, d'autres enfin donnent des résultats de requêtes par similarité. Toutefois, comme ce sont les seuls éléments de comparaison que nous ayons, il nous paraît intéressant de les signaler à titre indicatif:

- dans le domaine de la reconnaissance de gestes, d'activités, et d'expressions faciales, les performances varient entre 60% et 100% de bonne classification. Pour la reconnaissance de gestes, elles sont de l'ordre de 85% et 90% [Cui 95,Rigoll 96,Kurita 97]. Les expressions de visage sont reconnues à près de 100% dans [Yacoob 96]. Dans [Yacoob 99], le taux de reconnaissance des différents types de marches et de leur direction est de 80% environ, et celui des mouvements de lèvres évolue entre 60% et 70% ;

- les résultats sur la détection des événements varient entre 75% et 95% de reconnaissance correcte dans [Naphade 98]. Ceux annoncés dans [Iyengar 98], pour des problèmes à deux classes, sont aux environs de 90%. Les résultats sur les requêtes par similarité sont assez disparates : les taux de précision et de rappel varient respectivement entre 12% et 56%, et 20% et 93% dans [Jain 99], ceux annoncés dans [Qian 99] où l'information contextuelle est fortement modélisée sont meilleurs, entre 68% et 100% pour le taux de précision et entre 44% et 100% pour le taux de rappel ;

- dans les travaux traitant de documents sportifs, la reconnaissance de gestes sportifs est de 100% chez [Nakano 00] sur un tout petit corpus, la classification correcte des plans varient entre 40% et 100% pour [Gong 95], la détection d'événements est exacte à 80% dans

[Chang 96]. Dans le cadre de requêtes par similarité, un taux de précision maximal de 60% est annoncé dans [Mohan 98].

Enfin, rapportons les résultats que nous avons pu trouver dans les deux travaux les plus proches des nôtres. Dans [Chomat 00], pour la base d'expérimentation 2, aucun taux de bonne classification et aucune matrice de confusion ne sont indiqués, mais dans [Chomat 99b], il semble que ce taux soit autour de 50%[9]. Dans [Chomat 00], des résultats sont par contre donnés sur la caractérisation de séquences d'activités humaines dans un bureau. Les taux de reconnaissance s'étagent, selon la paramétrisation, entre 20% et 100%, et certains jeux de paramètres permettent d'obtenir des taux presque exclusivement au-dessus de 75%. Dans [Fablet 01], dont nous nous sommes inspirés pour la constitution des bases d'apprentissage, les résultats de reconnaissance annoncés varient entre 60% et 100% selon les méthodes utilisées et les tests effectués.

Ainsi, si nous considérons les descripteurs les plus efficaces, les résultats que nous avons obtenus parviennent à égaler les performances similaires à celles trouvées par ailleurs pour des problèmes plus ou moins proches de reconnaissances dans des vidéos.

**Généralisation de l'utilisation des images de l'historique du mouvement.** Nous avons développé une extension des MHI plus générale. Associée aux SVM, la signature $Map_{MHI}$ nous a permis d'obtenir des résultats satisfaisants sur l'ensemble des bases d'expérimentation. À ce titre, la généralisation des MHI à des séquences filmées à caméra mobile et leur utilisation pour décrire des mouvements de nature variée et des scènes complexes où plusieurs objets sont en mouvement est un des résultats prometteurs des travaux que nous avons menés. Toutefois, nous n'avons pu proposer de descripteurs extraits de ces MHI généralisés pouvant se substituer à l'utilisation des moments de Hu ou à l'extraction du gradient spatial, proposée par A. Bobick et son équipe [Bobick 01]. Un constat est aussi que la compensation du mouvement dominant en amont du calcul des MHI semble préférable à l'intégration des mouvements de caméra dans les MHI. Enfin, les MHI restent des descripteurs du mouvement localisés sur un horizon temporel d'une quinzaine d'images ; ceci implique de définir des stratégies de globalisation de l'information extraite lorsque nous souhaiterons utiliser ces descripteurs sur des segments temporels plus longs.

**Proposition d'une alternative aux images de l'historique du mouvement.** La signature $Map_{STG}$, qui s'appuie sur une "cartographie" du mouvement plus riche que les MHI, a donné des résultats prometteurs en égalant les performances du descripteur $Map_{MHI}$. Il y a toutefois une certaine déception à constater que l'enrichissement de l'information de mouvement fournie par le recours aux filtres de Gabor spatio-temporels n'a pas entraîné une amélioration sensible des résultats. Cela appelle deux commentaires. Le corollaire de l'enrichissement de l'information de mouvement est une plus grande variabilité des signatures. Le nombre relativement faible d'échantillons d'apprentissage a peut-être davantage mis en exergue dans nos expérimentations la complexité accrue plutôt que la richesse informationnelle. De plus, comme nous l'avons déjà signalé, la paramétrisation des familles de filtres de Gabor utilisée reste à optimiser pour cet usage.

**Utilité et robustesse des machines à vecteurs de support.** Les SVM sont une technique de classification qui connaît un certain succès ces dernières années. Nous avons pu mesurer leur intérêt pour la classification des séquences et, éventuellement, pour la détection d'événements, ainsi que leur robustesse et leur efficacité démontrées dans le cas des signatures $Map_{MHI}$ et $Map_{STG}$.

9. Cette indication issue de [Chomat 99b, Fig. 8] est à considérer avec précaution, les indicateurs considérés sont les probabilités de reconnaissance d'une action au cours du temps.

Si la question de l'amélioration des signatures extraites et de la définition d'une méthodologie de construction des ensembles d'apprentissage reste ouverte, l'utilisation des SVM nous paraît, par contre, validée pour ce type d'usage. Nous jugeons aussi prometteuse la mise en œuvre de plus en plus fréquente dans les outils rendus publics des différentes variantes des SVM (SVM pondérés, classification multi-classes).

**Un cadre général pour la classification de séquences vidéos.** En dépit des difficultés rencontrées et des questions méritant des compléments d'expérimentation, nous sommes parvenus à proposer un cadre général de classification des séquences vidéos. Il permet d'envisager une stratégie assez générale de caractérisation du contenu des séquences ou de détection d'événements qui ne se limite ni aux types de sports abordés dans cette étude, ni aux seuls documents sportifs.

**Apports des travaux menés et perspectives.** Nous avons listé les principaux résultats obtenus, et nous détaillerons ci-après les difficultés rencontrées, et quelques pistes de recherche envisageables. Il nous paraît important de mettre en exergue les deux voies d'investigation pour lesquelles nous espérons avoir pu ouvrir des perspectives.

La stratégie que nous avons développée nous a permis, au moins avec les signatures $Map_{MHI}$ et $Map_{STG}$, d'obtenir des résultats satisfaisants concernant la caractérisation du mouvement présent dans des séquences vidéos. Les descripteurs proposés peuvent avoir d'autres applications, comme la recherche par similarité dans des bases de séquences. Les segments temporels pourraient aussi être caractérisés par ce biais en vue d'autres traitements. On pourrait ainsi associer à un plan une signature, constituée des réponses d'une famille de classifieurs, qui indiquerait la similitude probable entre le contenu dynamique du plan et différents types de mouvement. L'information associée aux classifieurs dépendrait de l'usage visé par l'utilisateur (classe de mouvements généraux, activités spécifiques, événements notables, etc.). Ces descripteurs pourraient ainsi être exploités dans les méthodes de macro-segmentation décrites dans la partie II.

Même si des difficultés perdurent (définition des ensembles d'apprentissage, localisation temporelle des séquences lors d'un traitement extensif des documents), les résultats obtenus ouvrent la voie à l'intégration de nos outils dans des environnements d'indexation semi-automatique. La labellisation de segments temporels d'un magazine sportif en fonction de différents types de sports serait ainsi possible. Nous pouvons même envisager de ne rendre actifs que les classifieurs correspondant aux sports annoncés dans la para-documentation (conducteurs, etc.) disponible à l'INA lors de l'indexation du flux. Notre stratégie d'indexation étant envisageable dans un cadre plus général, elle pourrait permettre l'accès à certains types de séquences (un sport donné, un but, un départ de course, etc.) en fonction d'un usage qu'il conviendrait de définir comme nous l'avons fait dans la partie II pour la macro-segmentation. Cette assistance à l'indexation de vidéos et à la navigation dans des bases de vidéos est bien sûr possible dans d'autres contextes que celui des magazines sportifs.

## Difficultés rencontrées

Nous avons été confrontés à de nombreuses difficultés aussi bien lors de l'extraction des descripteurs que lors de l'utilisation des SVM pour la classification. Un des problèmes principaux a notamment été la difficulté récurrente de paramétrisation des algorithmes.

**La gestion des paramètres.** La paramétrisation des MHI a été la moins problématique. Pour le paramètre $V_{max}$, nos expérimentations préliminaires ont rapidement conduit au choix d'une valeur

en accord avec les recommandations de A. Bobick *et al.* [Bobick 01]. Il en va de même pour le paramètre $\tau$, en dépit du caractère simplificateur lié au seuillage.

La paramétrisation des filtres de Gabor est apparue plus délicate. Une partie des paramètres concerne, notamment, la largeur de la gaussienne $(\sigma_x, \sigma_y, \sigma_t)$ et le choix des fréquences centrales $(\omega_{x_0}, \omega_{y_0}, \omega_{t_0})$. Lors de nos expérimentations, certains constats visuels (voir figure 8.8) ont rejoint les arguments développés par O. Chomat [Chomat 00], qui propose une famille de filtres plus sélectifs spatialement et moins temporellement. Il est possible que, pour une utilisation des réponses de ces filtres visant à constituer une carte des orientations et des amplitudes locales du mouvement, un autre paramétrage des filtres aurait été plus pertinent. Des expérimentations plus complètes à ce sujet seraient sans doute nécessaires. Le problème est similaire avec les paramètres de seuillage $\{\tau_0, \ldots, \tau_3\}$. Les valeurs ont été fixées expérimentalement, ce qui ne paraît pas entièrement satisfaisant. Comme nous n'avons pu définir un cadre d'évaluation rigoureux, l'interaction entre les différents niveaux et orientations des filtres a rendu l'évaluation visuelle malaisée.

Enfin, le choix du paramètre $C$ intervenant dans la classification par les SVM nécessiterait sans doute aussi des expérimentations supplémentaires, et certains des problèmes observés restent en suspens.

**L'extraction des descripteurs** Les deux signatures, $Map_{MHI}$ et $Map_{STG}$, qui se sont avérées les plus performantes, ont le désavantage d'être de grande taille (cf. tableau 8.2). Si les SVM sont peu sensibles à cet aspect de taille, ce dernier a néanmoins un impact sur la question du stockage et du temps de calcul. Un autre argument pour la réduction de la signature $Map_{MHI}$ était de globaliser spatialement l'information localisée en chaque point du MHI. Force est de constater que nos tentatives de réduire l'information contenue dans les MHI n'ont pas été couronnées de succès. Les descripteurs de taille réduite issus des MHI n'ont jamais pu égaler les performances produites par la signature $Map_{MHI}$.

Nous avions aussi pour objectif d'étudier le comportement des MHI dans le contexte plus difficile des séquences d'images à caméra mobile. Les résultats obtenus sur la base d'expérimentation 3 et le taux de classification correcte de 84% sur la base d'expérimentation 5 pour $Map_{MHI}$ semblent montrer que, malgré la superposition des informations de mouvements d'objets et de caméra, une classification satisfaisante peut être atteinte. Néanmoins, de meilleurs résultats ont été obtenus après compensation du mouvement dominant. Par ailleurs, rappelons que la mise en œuvre de la compensation du mouvement dominant en amont des filtres de Gabor ne nous a pas paru aisément réalisable (cf. sous-section 8.2.3).

En outre, la normalisation et l'alignement temporel des séquences n'ont pas été abordés dans cette étude. Il est vrai que nous n'avons été confrontés à ce problème que pour la base d'expérimentation 2, mais une utilisation opérationnelle à grande échelle des outils développés pourrait le requérir.

Enfin, un outil de visualisation des séquences dans l'espace de représentation des différentes signatures nous a manqué. Il nous aurait permis de mener une analyse plus fine de la validité des signatures et de la pertinence des classes retenues. Nous avons dû inférer l'éventuelle disparité des descripteurs à l'intérieur des classes et la séparation des différentes classes entre elles à partir des résultats finaux de l'apprentissage ou de la classification.

**Des problèmes d'apprentissage** Nous souhaitions utiliser les SVM comme une simple technique de classification "boîte noire". L'usage spécifique que nous en avons fait et les difficultés que nous avons rencontrées, largement évoquées dans le descriptif de nos expérimentations, nous ont amenés à nous interroger plus qu'envisagé initialement sur le fonctionnement des SVM et la théorie

qui les sous-tend. Sur ce point, de nombreuses questions restent à éclaircir davantage que nous avons pu le faire : influence réelle du paramètre $C$ et de la taille des données en entrée, définition qualitative et quantitative des bases d'apprentissage, etc.

Les techniques par apprentissage nous paraissaient séduisantes pour passer du niveau numérique à un certain niveau sémantique sans avoir à recourir à une modélisation fine, spécifique, ou coûteuse de l'information contenue dans les images. Les SVM, en particulier, semblaient avoir l'avantage d'une certaine simplicité de mise en œuvre associée à une grande efficacité (voir section 8.3). Ces deux avantages doivent néanmoins être relativisés à la lumière de nos expérimentations. Le réglage du paramètre $C$ et la définition d'un ensemble d'apprentissage adéquat restent assez ouverts : Comment fixer le compromis entre la précision du classifieur sur l'ensemble d'apprentissage et sa capacité de généralisation ? Combien faut-il d'exemples pour entraîner correctement les classifieurs ? Comment choisir les échantillons d'apprentissage pour modéliser, en toute généralité, une classe donnée ?

## Des pistes à suivre

Le premier travail à mener à la suite de nos expérimentations est sans aucun doute la mise en œuvre de tests et d'outils complémentaires afin d'analyser plus finement l'influence des différents paramètres, le contenu informatif des signatures proposées, la pertinence des corpus et des ensembles d'apprentissage, etc.

Concernant la réduction de la représentation des descripteurs extraits, plusieurs pistes pourraient être explorées. Le recours à un histogramme multi-dimensionnel [Chomat 00] ou à une analyse en composante principale (ACP) [Iyengar 98], notamment, pourraient être envisageables. Une autre difficulté est d'exploiter à bon escient les mouvements dus à la caméra et ceux dus aux objets. Nous pourrions envisager de concaténer les MHI calculées après compensation du mouvement dominant aux paramètres de celui-ci, en entrée des SVM [Cui 95]. De manière générale, il pourrait être intéressant, compte tenu de la capacité des SVM à accepter des vecteurs d'entrée de grande dimension, de concaténer des descripteurs autres que ceux de mouvement, nous rapprochant ainsi des méthodes évoquées à la sous-section 7.2.1.

Sur la question de la normalisation temporelle des informations contenues dans des séquences de durée variable, citons l'approche proposée dans [Bradski 00] où une variante des MHI, notée tMHI (pour *timed Motion History Image*) permet de rendre les descripteurs relativement indépendants du nombre d'images par seconde dans le flux vidéo et de la durée de réalisation d'un mouvement. Par ailleurs, les articles traitant de la reconnaissance des gestes proposent souvent des méthodes de réduction et de normalisation temporelles des données en entrée, méthodes qui pourrait être étudiées et adaptées à notre cadre d'étude.

Les SVM offrent des variantes d'utilisation sur lesquelles nous ne nous sommes pas penchés au cours de nos travaux. Ainsi, certains auteurs ont proposé un apprentissage pondéré des problèmes de classification à deux classes [Chang 01]. Il s'agit de pénaliser différemment les erreurs sur les échantillons de la classe apprise et sur ceux des autres classes. Le paramètre $C$ est alors différent selon la classe des échantillons dans les équations régissant les SVM. Les problèmes d'apprentissage rencontrés consistaient en une mauvaise réponse du classifieur à un exemple de la classe associée : celle-ci n'aurait en quelque sorte pas été suffisamment apprise. L'utilisation de SVM pondérés pourrait peut-être améliorer ce point. Malheureusement, cette option n'était pas offerte dans la librairie *LibSVM* dans la version 2.1 que nous avons utilisée, elle l'est dorénavant dans la version 2.31 de *LibSVM*. Par ailleurs, certains auteurs utilisent des stratégies d'optimisation des ensembles d'apprentissage. Citons, notamment, la technique du *bootstrapping* utilisée par E. Osuna [Osuna 97a].

Il s'agit d'inclure les fausses alarmes à l'ensemble d'apprentissage afin d'enrichir la représentation du "reste des classes". Nous aurions alors des ensembles d'apprentissage spécifiques à chaque classifieur. Compte tenu du nombre limité d'exemples dont nous disposions, nous n'avons pu mettre en œuvre une telle stratégie. De plus, les outils de classification par SVM disponibles intègrent de plus en plus la gestion des apprentissages multi-classes ; il conviendrait d'évaluer ces outils, la classification à réaliser ne nécessitant plus que la construction d'un seul classifieur.

Enfin, il faudrait bien sûr augmenter considérablement le nombre d'exemples d'apprentissage disponibles, condition nécessaire pour espérer un entraînement satisfaisant des classifieurs, et une étude approfondie de leur comportement.

# Conclusion générale

Nous avons abordé les problématiques de la segmentation temporelle des documents audiovisuels en séquences, et de la caractérisation du contenu dynamique de segments vidéos. Les missions de catalogage, d'archivage, et d'indexation de l'INA ont constitué le cadre applicatif de nos travaux, l'objectif étant de proposer des algorithmes pouvant enrichir les outils d'analyse, automatiques ou semi-automatiques, utilisés par les documentalistes pour mener à bien l'indexation des documents audiovisuels. Nous étions tenus de porter une attention particulière aux usages prescrits par ce contexte applicatif, à l'évaluation des méthodes développées, et à la gestion des paramètres d'algorithmes destinés au traitement de masse de documents audiovisuels.

Nous proposons, dans les sections suivantes, de conclure sur les apports des recherches menées, ainsi qu'une synthèse des difficultés rencontrées et des perspectives ouvertes par nos travaux.

## Synthèse des travaux effectués

Nous avons considéré divers aspects de l'analyse du signal vidéo, de l'extraction de descripteurs, aux questions de classification des contenus dynamiques. Nos recherches sur l'analyse de l'information contenue dans le flux vidéo se sont portées sur l'extraction et l'utilisation conjointe de descripteurs, notamment de couleur, pour des images fixes représentatives des segments, ainsi que sur la définition de descripteurs du mouvement pour de courtes séquences d'images. Concernant la classification des séquences vidéos, nous nous sommes plus particulièrement intéressés à des méthodes de classification hiérarchique, ainsi qu'à des méthodes de classification par apprentissage. Ces travaux ont donné lieu à de nombreux développements informatiques, et nous ont permis d'étudier des domaines variés comme la reconnaissance de forme, les filtres de Gabor spatio-temporels, et les machines à vecteurs de support.

La méthode de macro-segmentation définie repose sur celle initialement proposée par Yeung et *al.* [Yeung 98]. Tenant compte des limites de cette dernière, nous avons proposé, et en partie validé, des améliorations comme l'introduction de fonctions non constantes pour la contrainte temporelle, l'utilisation de la méthode de Ward pour la construction de la hiérarchie des classes, et la définition d'un critère fondé sur la distance cophénétique pour déterminer les ruptures entre séquences. Le calcul de ce critère permet notamment de remplacer la détermination d'un niveau de seuillage par la donnée du nombre de séquences recherchées, ce qui correspond mieux aux besoins de l'utilisateur final.

Nous avons défini une stratégie pour la caractérisation du contenu des documents audiovisuels. Nous l'avons appliquée à la caractérisation du contenu dynamique de séquences sportives, mais le cadre proposé est générique et pourrait s'adapter à d'autres primitives, et à d'autres applications. Deux descripteurs du mouvement dans des séquences d'images ont été développés. Le premier est l'image de l'historique du mouvement, introduite par Bobick et *al.* [Bobick 01], dont nous avons généralisé l'usage. Le deuxième exploite des filtres de Gabor spatio-temporels, et fournit

une cartographie du mouvement qui intègre l'information locale d'amplitude et d'orientation du mouvement apparent. Nos expérimentations de classification du mouvement nous ont aussi amenés à valider l'utilisation des machines à vecteurs de support dans notre cadre applicatif.

Notre démarche, guidée par les usages d'indexation de l'INA, a été également l'occasion de mener de nombreuses réflexions sur la paramétrisation et l'évaluation des algorithmes. Ainsi, nous avons effectué des expérimentations portant sur les paramètres principaux des méthodes utilisées, et notamment sur l'influence des paramètres $\Delta T$, lié à la contrainte temporelle sur les similarités entre plans pour la macro-segmentation, et $C$, associé aux erreurs lors de l'apprentissage des classifieurs pour la classification. Nous avons constitué, pour l'évaluation des outils de macro-segmentation, une première proposition de corpus de référence annoté manuellement, ainsi qu'une méthodologie, à la fois quantitative et qualitative, d'évaluation. Nous avons mis en œuvre, pour la caractérisation du contenu dynamique, une stratégie d'évaluation progressive intégrant des corpus de tests issus de la communauté scientifique, et des bases de séquences issues des corpus INA.

Lors des évaluations des différents algorithmes, nous avons obtenu des performances proches de celles annoncées pour des problèmes voisins. Ces évaluations liées aux usages semblent valider, en partie, l'adéquation des méthodes proposées aux besoins de l'INA.

## Difficultés rencontrées

Les principales difficultés rencontrées ont concerné la paramétrisation et l'évaluation des algorithmes. Dans notre contexte d'étude où les utilisateurs finaux sont les documentalistes de l'INA, une contrainte forte est que les différents paramètres et options puissent être fixés sans intervention humaine, ou lorsque cela est inévitable, que la gestion puisse être intuitive pour l'utilisateur. Par ailleurs, l'utilisation d'outils d'analyse automatiques ou semi-automatiques au sein de la chaîne de documentation de l'INA requiert non seulement un faible taux d'erreur sur les résultats, mais aussi que la correction manuelle de ces erreurs soit rapide, ou du moins nécessite un temps inférieur à la réalisation manuelle de la tâche considérée.

En général, les paramètres peuvent être gérés selon l'une des trois approches suivantes :

- les paramètres sont fixés par défaut une fois pour toutes dans l'algorithme ;
- les paramètres sont fixés automatiquement en fonction du contexte. Ils peuvent être calculés au sein de l'algorithme en fonction de données présentées en entrée ou extraites par celui-ci. Ils peuvent aussi être fixés par le biais d'un prototypage de l'algorithme en fonction du contexte, dépendant par exemple d'une taxonomie des documents audiovisuels ;
- les paramètres peuvent être fixés par l'utilisateur.

Nous n'avons pas étudié la deuxième approche dans le cadre de notre étude. Nous nous sommes efforcés de proposer des valeurs de paramétrisation par défaut, ou, lorsque cela nous était impossible, de permettre une gestion des paramètres compatibles avec les usages. À l'issue de nos travaux, nous devons reconnaître que l'étude et l'optimisation de nombreux paramètres reste un problème sensible, qui nécessiterait des expérimentations complémentaires.

Nous avons tenté de définir des méthodologies d'évaluation fondées sur les usages. Nous avons d'ailleurs fait un certain nombre de propositions en ce sens, qui constituent selon nous un des apports de nos travaux. Toutefois, dans la perspective d'une intégration dans les outils de documentation de l'INA, nos corpus restent parcellaires. Nous nous sommes aussi heurtés aux limites d'une évaluation quantitative et aux difficultés de mise en œuvre d'une évaluation qualitative rigoureuse. Enfin,

nous n'avons pas pu valider nos méthodes dans le cadre d'un traitement de masse des documents audiovisuels.

Notons, pour conclure sur ce point, que ces deux écueils ne nous sont évidemment pas spécifiques, et concernent l'ensemble du domaine de recherche lié à l'indexation automatique des documents audiovisuels. On peut souligner néanmoins que notre travail constitue une des premières tentatives pour faire face à cette question d'une évaluation approfondie guidée par les usages.

## Perspectives

Les difficultés évoquées ci-dessus suggèrent une exploration plus complète de ces questions. Certaines pistes dépassent d'ailleurs le cadre du développement strict d'outils d'indexation automatique et mettent l'accent sur des besoins méthodologiques concernant la constitution et l'annotation manuelle de corpus audiovisuels de référence, la définition de stratégie d'évaluation reconnue, une analyse "sociologique" ou "ethno-méthodologique" des usages professionnels de l'audiovisuel, aussi bien au niveau de la prise de vue et de la production qu'à celui de l'annotation des documents audiovisuels par les documentalistes.

Les perspectives ouvertes par nos travaux sur la macro-segmentation sont double. Elles concernent d'une part la définition de descripteurs plus efficaces porteurs d'une information diversifiée sur le contenu, l'utilisation conjointe de différentes sources d'information, l'amélioration de la description de la similarité des plans au sein de la hiérarchie des classes, et d'autre part la prise en compte sous une forme à définir de contraintes liées à la para-documentation. D'un point de vue applicatif, les bons résultats obtenus sur certains types de documents ouvrent la voie à une intégration de notre outil dans la chaîne de documentation de l'INA.

Les méthodes proposées pour la classification du contenu dynamique des séquences restent davantage du domaine de la recherche. Des travaux restent à effectuer sur les questions de la réduction de la dimension des primitives, de la normalisation temporelle des séquences d'activités, ou l'extraction de descripteurs globaux pertinents à partir des cartes des mouvements construites. Par ailleurs, l'apprentissage des classifieurs fondés sur les SVM et la détection d'événements semblent être une perspective de recherche prometteuse.

En conclusion, il convient de noter que le contexte applicatif de nos travaux pourrait être profondément modifié dans les années à venir. En effet, la mise en place de normes comme la norme MPEG-7 permettra de rendre disponible, en même temps que le flux audiovisuel, un certain nombre d'annotations liées à la production du document audiovisuel. De plus, la mise en place à l'INA de la captation numérique directe du flux entrant permet d'envisager à moyen terme l'éventualité de l'intégration d'une documentation riche sur celui-ci, liée à la diffusion ou au contenu des documents. Cette double perspective ferait sensiblement évoluer le travail des documentalistes de l'INA, et par conséquent, leurs besoins en outils d'analyse automatique des documents audiovisuels.

# Annexes

# Annexe A

# Informations sur le corpus de documents utilisés

## A.1 Données générales

Les documents utilisés ont été numérisés au format MPEG-1, à 25 images par seconde. Ils sont issus de différents corpus élaborés par l'INA soit dans le cadre national de l'*Action Indexation Multimedia* [1] (AIM [2]) soit dans celui de projets menés [3] à la *Direction de la Recherche et Expérimentation* (DRE) de l'INA. Seules les séquences de déplacements humains et de mouvements de caméra, numérisés à 15 images par secondes, proviennent de sources extérieures.

Une description succincte du corpus des documents utilisés est proposée dans le tableau A.1.

## A.2 Annotations de référence pour le découpage en macro-segments

Nous présentons dans cette section plus en détail les résultats obtenus lors de l'annotation manuelle d'un corpus de référence dédié à l'évaluation de l'outil de macro-segmentation présenté dans la partie II. Ces travaux ont été menés au sein du GRAMM de l'INA par des chercheurs et des profesionnels de la documentation. L'étude a été plus particulièrement menée sur quatre documents : *aim1mb05*, *topa_gainsbourg*, *munich2* et *aim1mb08* (voir la description des documents dans le tableau A.1). Nous détaillerons les niveaux de macro-segmentation proposés au final pour chaque type de documents, ainsi que quelques commentaires sur les principales difficultés rencontrées.

### A.2.1 Document *aim1mb05* - journal télévisé

La macro-segmentation a été effectuée sans difficulté majeure conformément aux prescriptions établies. Nous avons effectué une première segmentation avec les séquences {*habillage graphique*, *studio*, *reportage*}, reprenant l'articulation classique de la structuration des journaux télévisés.

Nous avons proposé une ébauche de hiérarchisation en répartissant les segments *studio* en trois sous-classes {*lancement sujet*, *brèves*, *interview*}. Pour les reportages, lorsque cela a été possible, nous avons tenté d'identifier les principaux *décors*.

1. Pour plus de détails sur l'utilisation de ce corpus, voir `http://www-asim.lip6.fr/AIM`.
2. Action commune aux GDR CHM et ISIS, puis GT10 du GDR-ISIS.
3. Ces projets expérimentaux et interdisciplinaires sont présentés en ligne sur le site de l'INA : `http://www.ina.fr/Recherche/projets\_experimentaux.fr.html`.

| Id | Corpus origine | Intitulé | Genre | Durée | Date |
|---|---|---|---|---|---|
| aim1mb01 | AIM | *Histoires d'eau* | documentaire | 35'35" | ? |
| aim1mb05 | AIM | Soir 3, le journal du soir (France 3) | journal télévisé | 15'26" | 1996 |
| aim1mb06 | AIM | Le Journal de 20 heures (France 2) | journal télévisé | 14'49" | ? |
| aim1mb07 | AIM | Le Journal de 20 heures (TF1) | journal télévisé | 14'50' | ? |
| aim1mb08 | AIM | épisode de la série *Chapeau melon et bottes de cuir* | fiction | 50'39" | 1976 |
| albertville | AGIR | Albertville 92 : retransmission des championnats d'Europe de patinage artistique | magazine sportif | 32'55" | 1992 |
| estivales_1 | ? | Estivales sur le bassin d'Arcachon | magazine de reportages | 42'01" | 1994 |
| estivales_2 | ? | Estivales sur la Côte Bleue | magazine de reportages | 42'00" | 1993 |
| f2jt_220897 | OLIVE | Le Journal de 20 heures (France 2) | journal télévisé | 38'00" | 1997 |
| f3jt_010897 | OLIVE | Soir 3, le journal du soir (France 3) | journal télévisé | 13'50" | 1997 |
| f3jt_031097 | AGIR | brèves sportives, Soir 3, le journal du soir (France 3) | journal télévisé | 2'48" | 1997 |
| f3jt_110997 | AGIR | informations sportives, Soir 3, le journal du soir (France 3) | journal télévisé | 2'01" | 1997 |
| f3jt_130795 | DiVAN | Soir 3, le journal du soir (France 3) | journal télévisé | 25'02" | 1995 |
| f3jt_160795 | DiVAN | Soir 3, le journal du soir (France 3) | journal télévisé | 23'29" | 1995 |
| f3jt_230897 | AGIR | informations sportives, Soir 3, le journal du soir (France 3) | journal télévisé | 2'11" | 1997 |
| munich1 | AGIR | *Sport Dimanche* : championnat d'Europe d'athlétisme à Munich (1) | magazine sportif | 57'12" | 1997 |
| munich2 | AGIR | *Sport Dimanche* : championnat d'Europe d'athlétisme à Munich (2) | magazine sportif | 57'25" | 1997 |
| munich3 | AGIR | *Sport Dimanche* : championnat d'Europe d'athlétisme à Munich (3) | magazine sportif | 55'34" | 1997 |
| pasrever | ? | *Faut pas Rêver* | magazine de reportages | 55'47" | 1994 |
| stade2 | OLIVE | *Stade 2* | magazine sportif | 1h 00' 32" | 1997 |
| talence1 | AGIR | *Sport Dimanche* : décastar de Talence | magazine sportif | 56'55" | 1996 |
| topa_gainsbourg | DiVAN | *Top à* Serge Gainsbourg | émission de variétés | 53'21" | 1974 |
| topa_halliday | DiVAN | *Top à* Johnny Halliday | émission de variétés | 29'00" | ? |
| come down go left right up | IMAG | déplacement humain à caméra fixe | 30 séquences de test | - | - |
| div rot trans | INRIA Rennes | mouvement de caméra sur motif texturé | 9 séquences de test | - | - |

TAB. A.1: *Documents utilisés pour la constitution des corpus d'expérimentation*

Enfin, il nous a semblé intéressant du point de vue de la documentation de regrouper dans un niveau de segmentation particulier et lacunaire les plans relatifs aux interviews (en studio ou dans les reportages). Ajoutons qu'afin de délimiter les frontières de ces séquences, nous nous sommes fondés sur l'image et non sur le son.

### A.2.2 Document *topa_gainsbourg* - émission de variété

L'alternance des prestations des artistes a servi de guide pour la constitution d'un niveau de segmentation principal, organisé selon les critères : {*générique*, *interprétation*, *interview*, *archive*}. Les prestations peuvent être des séquences très différentes (ballets, chanson solo, duo ou trio, sketch, etc.) mais une analyse plus fine s'apparente à une classification des séquences et non à une hiérarchisation de la segmentation. Les séquences de type *archives* peuvent être des extraits de répétition, d'archives télévisuelles, de reportages tournés à l'avance, etc.

### A.2.3 Document *munich2* - émission sportive

Comme pour les journaux télévisés, nous avons commencé par mettre en évidence la structure des émissions *Sport Dimanche* étudiées. Ainsi, un premier niveau de segmentation se fonde sur les séquences {*habillage graphique*, *studio*, *événement sportif*}. De même, nous avons, ensuite, proposé une ébauche de structuration en sur-segmentant les séquences *studio* en trois sous-classes {*lancement sujet*, *brèves*, *sommaire*}. Pour les événements sportifs, nous avons identifié les types de segments suivants : {*épreuve*, *commentaire sur site*, *interview*, *divers* (regroupant les vues générales récurrentes du site et les tableaux de résultats)}.

Enfin, il nous a semblé pertinent de proposer un niveau de segmentation lacunaire regroupant les plans relatifs aux ralentis et aux photogrammes. Notons que pour ce type de documents, le repérage des *interviews* peut être intégré à une structuration hiérarchique et n'a pas été annoté sur un niveau à part.

Remarquons aussi que dans un premier temps, et sans doute un peu abusivement, nous avons qualifié d'*habillage* les génériques de début et de fin, les séquences enchaînées de films publicitaires et les bandes annonces promotionnelles. En effet, ces types de séquences ne présentant pas d'intérêt documentaire *a priori* ni ne nécessitant de traitement automatique spécifique, elles ont été regroupées en un seul macro-segment.

### A.2.4 Document *aim1mb08* - fiction

C'est sans conteste le type de document qui a posé le plus de questions en matière de repérage d'unités sémantiques. La connaissance de l'usage des documents de fiction ne nous a guère aidés dans la mesure où la recherche dans les documents de fiction à l'INA se limite à retrouver un acteur ou à localiser les œuvres d'un auteur, d'un scénariste.

Plusieurs options se présentaient à nous en matière de macro-segmentation du document de fiction :

- caler le découpage sur les changements de décors ;
- constituer des segments en fonction de la structure narrative en repérant des unités d'actions ;
- se fonder sur l'apparition-disparition des personnages.

Du point de vue algorithmique, les découpages *décor* et *action* semblent intuitivement liés, respectivement, à l'extraction des primitives de couleur et de mouvement, tandis qu'en l'état actuel de la recherche un découpage fondé sur les personnages semble plus difficile à mettre en œuvre[4]. Notons que nous retrouvons certaines des notions du théâtre classique (unité de lieu et d'action), et qu'une pièce de théâtre est effectivement segmentée selon la présence des personnages (scène), et les changements de décors (acte).

Nous avons donc, dans un premier temps, étudié le découpage *action/lieux/personnages* et observé, comme cela était prévisible, une forte interaction rythmique entre ces éléments (voir tableau A.2). Notre hypothèse initiale de structuration hiérarchique emboîtée a donc été infirmée, puisque nous avons mis en évidence des segmentations qui s'enchevêtrent, fondées sur des grilles de lecture complémentaires, quoique corrélées, et non sur une structure de niveaux hiérarchiques.

Nous avons ensuite été amenés, lors de la mise en œuvre, à ne pas valider le découpage *lieu* car, dans le cadre particulier étudié, celui-ci recouvrait sans grande pertinence soit la segmentation en plans, soit le découpage *action* selon le niveau de granularité retenu.

### A.2.5 Quelques commentaires sur la macro-segmentation manuelle réalisée

Les résultats présentés et rendus disponibles aux formats XML doivent être considérés comme une première proposition de macro-segmentation de référence. Il conviendrait de prendre en compte davantage de documents par genre afin de valider nos observations et de s'assurer que nous n'avons pas été victimes de biais dus au traitement d'un document particulier. De même, il serait intéressant, à terme, d'enrichir ce premier corpus d'autres genres de documents. Insistons une fois encore sur le fait qu'en dépit d'une approche que nous avons voulu rigoureuse, nous ne proposons qu'un découpage subjectif en accord, dans la mesure du possible, avec les usages en pratique ou prospectifs tels que nous les percevons à l'INA.

À titre d'exemple d'usage prospectif, remarquons que les outils de segmentation automatique et les documentalistes proposent un découpage des journaux télévisés sur le schéma *plateau/reportage*. Cependant, il est tout à fait possible d'imaginer que des utilisateurs professionnels dans le cadre du Département Droits et Archives ou des scientifiques dans le cadre de l'Inathèque estiment que l'unité temporelle pertinente soit constituée du reportage précédé de son lancement plateau, et suivi, le cas échéant, d'un retour plateau pour une interview. Un tel usage invaliderait l'annotation proposée et serait beaucoup plus difficile à prendre en compte par des traitements automatiques. Toutefois, une étude des usages associée à des réflexions prospectives de ce genre seraient sans doute très enrichissantes dans la perspective de proposer de nouveaux outils pour inventer de nouveaux usages. Vaste projet...

## A.3 Exemples de notices INA

À titre d'exemple et d'illustration, nous proposons, avec l'aimable autorisation du *Département Droits et Archives* (DDA), trois exemples de notices INA pour trois types de documents issus du corpus utilisé au cours de cette étude. Il s'agit des documents identifiés comme *topa_gainsbourg* (émission de variétés), *aim1mb05* (journal télévisé) et *munich2* (émission sportive). Notons que pour le journal télévisé, il y a une notice par reportage ; nous n'en présenterons que quelques-unes.

4. Si de nombreuses méthodes pour la détection de visages ont été proposées, la fiabilité de celles que nous avons expérimentées ne nous a pas pleinement satisfait. De plus, il faudrait proposer un suivi et une classification des visages, sujets de recherche sur lesquels les recherches sont encore en cours.

| Actions | Lieux | Personnages |
|---|---|---|
| Espionnage | Corridor (couloir, sortie de secours) | Steed, Purdey |
| Poursuite | Montagne, carrière | Steed, Purdey, Licorne |
| Générique série | | Steed, Purdey, Gambit |
| Générique épisode | Carrière | |
| Poursuite (fin) | Carrière | Steed, Purdey |
| Attentat | Église (parvis, rue, bureau) | Cardinal, Steed, Purdey |
| Aéroport | Aéroport (piste, téléphone, bureau) | Licorne, Gambit |
| Piège | Paris (champs élysées, rue, terrasse), hôtel particulier(1er étage, hall, ascenseur) | Steed, Purdey, Gambit, Licorne |
| Arrestation | Hôtel particulier (1er étage, salle du manège), terrasse | Steed, Purdey, Gambit, Licorne |
| Où est la licorne? | Paris (terrasse), péniche (extérieur, intérieur) | Riter, Marco, Grima |
| Arrestation (suite) | Hôtel particulier (salle du manège) | Steed, Purdey, Gambit, Licorne |
| Où est la licorne? (suite) | Paris (fenêtre), péniche (intérieur), vue extérieure salle du manège | Riter, Marco |
| Assassinat de la licorne | Hôtel particulier (salle du manège), Paris (fenêtre) | Steed, Purdey, Gambit, Licorne, Riter |
| Poursuite du tueur | Paris (terrasse, rue, bouche de métro) , hôtel particulier (ascenseur) | Riter, Purdey, Gambit |
| Stratagème de John Steed | Hôtel particulier (salle du manège, toit...), Paris (rues), voiture de Marco | Steed, Purdey, Gambit |
| Stratagème des amis de la Licorne | Péniche (intérieur) | Riter, Marco, Grima |
| Enlèvement du prince | Train (compartiment, couloir), voies ferrées | Riter, Grima |
| Tribulations de l'échange | Hôtel particulier (salle du manège), Paris (rue) | Steed, Purdey, Gambit, chef des services secrets |
| Détention du prince | Péniche (extérieur) | Riter, Marco, Grima |
| Tribulations de l'échange (suite) | Hôtel particulier (salle du manège) | Steed, Purdey, Gambit, chef des services secrets |
| L'indic | Paris (champs élysées, boulodrome), voiture de steed | Steed, Purdey, Henri Duval |
| Tribulations de l'échange (suite) | Parking garage | Chef des services secrets, Marco |
| Le doute | Péniche (intérieur) | Riter, Marco, Grima |
| Parachutage | Hôtel particulier (cuisine, salle du manège, salle du mort, toits, façade), Paris (rues, camion) | Grima, policier |
| Arrestation | Paris (rues, camionnette, camion), hôtel particulier (1er étage, escaliers, toits) | Grima, Gambit, Steed, Purdey |
| Où est Grima? | Péniche (extérieur) | Riter, Marco |
| Interrogatoire | Hôtel particulier (salle du manège) | Grima, Gambit, Steed, Purdey |
| Tribulations de l'échange (suite) | Hôtel particulier (salle du manège), péniche (intérieur) | Steed, Purdey, Gambit, chef des services secrets, Marco, Riter |
| Dispositifs de l'échange | Hôtel particulier (sous-sol : ascenseur, hall, ascenseur, atelier, escaliers), camionnette | Steed, Purdey, Gambit, Riter, Marco, Prince |
| Arrestation des 2 malfaiteurs | Hôtel particulier (ascenseurs, hall, camionnette, perron) | Steed, Purdey, Gambit, Riter, Marco |
| Conclusion | Paris (voiture, Moulin rouge) | Steed, Purdey, Gambit |
| Générique de fin | | |

TAB. A.2: *Découpage Actions/Décors/Personnages pour la fiction* aim1mb08

Les notices se présentent comme des fiches constituées de champs contraints ou libres. Les documentalistes ont accès pour les créer à des thésaurus et des "bibles d'indexation" ou guides d'usage. Dans les notices présentées, le découpage en séquences est parfois explicitement donné. Il est possible d'observer aussi le haut niveau de la documentation effectuée, son caractère parfois très synthétique, l'usage courant de notions très conceptuelles, les notations concernant la forme du document (notamment les valeurs de plan), l'utilisation de noms propres (personnes, lieux), et les transcriptions parcellaires d'interview ou de discours.

**Liste des abréviations utilisées dans les notices.** CPL : contre-plongée; DP : divers plans; INT : interprète; ITW : interview; GP : gros plan; PANO : panoramique; PAR : participant; PE : plan d'ensemble; PR : plan rapproché.

### A.3.1 Notice de *topa_gainsbourg*

Id notice: CPF86654021
Date de diffusion: 04.05.1974
Société de programmes: ORTF
Durée: 00H 54MIN 00SEC
Titre collection: TOP A
Titre sous-collection: SERGE GAINSBOURG
Forme: SPECTACLE TV
Genre: VARIETES
Descripteurs thématiques - Terme: INT, BIRKIN JANE; INT, GAINSBOURG SERGE; INT, HARDY FRANCOISE; INT, BEDOS GUY ; INT, DUTRONC JACQUES ; PAR , PLASSCHAERT ARTHUR
Oeuvres:
Jane Birkin chante *Bébé gai.*
Serge Gainsbourg chante *Dr Jeckyll.*
Françoise Hardy et Jane Birkin chantent *Les p'tits papiers.*
Gainsbourg chante *La Javanaise.*
Jane Birkin chante *C'est la vie qui veut ça.*
Sketch de Guy Bedos *Le tombeur de filles.*
Gainsbourg chante *Je suis venu te dire que je m'en vais.*
Jane Birkin chante *Something that happens.*
David Cassidy chante *I am just a day dreamer.*
Jane Birkin chante *Je plais aux G.I..*
Françoise Hardy chante *Je suis moi.*
Jacques Dutronc chante *L'espace.*
Gainsbourg chante *Le soleil est rare.*
Jane Birkin chante *My chérie Jane.*
Patronyme générique: BIRKIN JANE, GAINSBOURG SERGE, HARDY FRANCOISE, BEDOS GUY,DUTRONC JACQUES, PLASSCHAERT ARTHUR
Producteurs : PRD, PARIS : OFFICE NAT. DE RADIODIFFUSION ET TELEVIS. FR (ORTF), 1974
Résumé: Présence des ballets Arthur Plasschaert.

### A.3.2 Notices de *aim1mb05*

Id notice: CAC96039364
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: 00h 00min 45sec
Titre collection: SOIR 3
Titre propre: POLICIERS ARRETES ARGENTINE
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: ARGENTINE; BUENOS AIRES; TERRORISME; ANTISEMITISME; ATTENTAT; EXPLOSION; SIEGE SOCIAL; ASSOCIATION; RUINE; VICTIME; BLESSE; SECOURS; ENTERREMENT; RETROSPECTIVE; JUSTICE; ENQUETE; ARRESTATION; POLICIER
Résumé: Onze policiers de la province de Buenos Aires en Argentine ont été arrêtés dans le cadre de l'enquête sur l'attentat contre la mutuelle juive Amia, le 18 Juillet 1994, qui avait fait 86 morts. Images d'archives du 18/07/94 : PE lieu de l'explosion, immeubles détruits, DP secours évacuant les blessés.
PR secours portant une civière.
Ambulances
Chien sur les ruines de l'immeuble détruit à la recherche de survivants.
PE enterrement des victimes de l'attentat.

Id notice: CAC96039392
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: -
Titre collection: SOIR 3
Titre propre: NELSON MANDELA A PARIS
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: FRANCE; YVELINES; RAMBOUILLET; CHATEAU; RELATION DIPLOMATIQUE; VISITE; CHEF D'ETAT; MANDELA NELSON; CHIRAC JACQUES; POIGNEE DE MAINS; PERRON; PARC; AEROPORT; ARRIVEE; ARTHUIS JEAN
Résumé: Nelson MANDELA, actuellement en visite officielle en france, a été reçu par le chef de l'Etat Jacques CHIRAC dans la résidence présidentielle au chateau de Rambouillet.
PE Jacques CHIRAC et Nelson MANDELA sur le perron du chateau de Rambouillet, couvert d'un tapis rouge pour l'occasion et encadré par la garde Républicaine
PR Poignée de main entre les deux hommes.
GP Jacques CHIRAC et Nelson MANDELA.
PE chateau de Rambouillet.
Jacques CHIRAC et Nelson MANDELA se promenant dans le parc du chateau.
PE arrivée de Nelson MANDELA à l'aéroport dOrly.
PE foule saluant le chef d'Etat sud-africain.
Nelson MANDELA à l'aéroport où il a été reçu par Jean ARTHUIS, le ministre de l'Economie et des finances.

Id notice: CAC96039395
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: 00H 01MIN 40SEC

Titre collection: SOIR 3
Titre propre: SORTIE SPELEOS BLOQUES
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: FRANCE; ISERE; VERCORS-GEOGRAPHIE; GOUFFRE; SPELEOLOGIE; ACCIDENT; SAUVETAGE; SECOURS; SECOURISTE; TENTE; SPELEOLOGUE; GRANDE BRETAGNE; DECES; BLESSE; GRENOBLE; HOPITAL; CONFERENCE DE PRESSE; EMOTION
Résumé: Un spéléologue reste encore bloqué dans le gouffre du Berger dans le Vercors sous assistance médicale à moins 900 m en dessous du niveau de la mer.
PR gouffre Berger où les sauveteurs en cordée se relaient pour extraire le dernier survivant de l'expédition britannique.
ITW d'un sauveteur : "Il n'est pas blessé, il n'a pas de gros problème physique mais le moral est bien atteint".
PR descente dans le gouffre en cordée.
Sortie d'un spéléologue / DP tente médicale installée à proximité du gouffre.
Descente des sauveteurs dans le gouffre.
Hopital MICHALLON de Grenoble : conférence de presse des deux spéléologues extraits hier du gouffre, qui ont été placés sous surveillance médicale
PE de la conférence de presse
Un spélélologue, bouleversé à l'évocation de ses deux camarades morts dans le gouffre, quitte la salle de conférence.
Id notice: CAC96039399
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: -
Titre collection: SOIR 3
Titre propre: TOUT IMAGES : FAMILLE OUFKIR EN FRANCE
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: FRANCE; MAROC; CASABLANCA; POLITIQUE INTERIEURE; HASSAN II; LIBERATION; PRISONNIER D'OPINION; RESIDENCE FORCEE; COUP D'ETAT; OUFKIR MOHAMED; PASSEPORT; ENFANT; OUFKIR MALIKA; OUFKIR SOUKAINA; OUFKIR RAOUF; ARRIVEE; AEROPORT; OUFKIR MARIA INAN; EMOTION; RETROSPECTIVE
Résumé: BREVE : Trois enfants du général Mohamed OUFKIR, ancien ministre marocain de la Défense auteur d'une tentative de coup d'état contre Hassan II et décédé en 1972, sont arrivés à Paris,à l'aéroport d'Orly.
Malika, Soukaina et Raouf OUFKIR, qui avaient quitté Casablanca, arrivent à l'aéroport d'Orly.
Embrassade avec leur soeur Maria et plusieurs proches de la famille.
ITW de Soukaina OUFKIR : "On a eu notre visa hier, on ne réalise pas encore".
ITW de Raouf OUFKIR au sujet du Roi Hassan II : "Je suis et je reste un sujet du roi, c'est notre pays".
Réaction de Malika OUFKIR : "Je suis trop émue".
Archive INA : le général OUFKIR en uniforme militaire.
Id notice: CAC96039409
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: -

Titre collection: SOIR 3
Titre propre: TOUT IMAGES : TOUR DE FRANCE
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: CYCLISME; COURSE CYCLISTE; ETAPE DE MONTAGNE; PELOTON
Résumé: BREVE : 13ème étape du Tour de France entre Le Puy-en-Velay et Super-besse-Sancy.
trés bref plan du peloton sur une route de montagne.
Id notice: CAC96039418
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: -
Titre collection: SOIR 3
Titre propre: TOUT IMAGES : RAVE PARTIE A BERLIN
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: RFA; BERLIN; FETE; MUSIQUE; RASSEMBLEMENT; JEUNESSE; FOULE
Résumé: BREVE : Huitième édition de la Love Parade à Berlin où 700000 personnes se sont retrouvées pour danser la techno et parader dans la rue.
PE rue de Berlin envahit par les danseurs de techno
danseurs le corps peint.
PE chars et foule.
Id notice: CAC96039428
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: 00H 01MIN 20SEC
Titre collection: SOIR 3
Titre propre: ZOOM : INVITES ELYSEE
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: FRANCE; LIMOUSIN; JEUNESSE; FETE NATIONALE; PARIS; RESIDENCE OFFICIELLE; RECEPTION; GARDEN PARTY; CHIRAC JACQUES; INVITE; AGRICULTEUR; VACHE; ELEVAGE; LIMOGES; FACULTE; ETUDIANT; INTERVIEW
Résumé: A l'occasion de la traditionnelle Garden Party du 14 juillet dans les jardins de l'Elysée, plusieurs jeunes ont été invités par le chef de l'Etat, Jacques CHIRAC.
Rencontre avec Valérie, jeune agricultrice dans le Limousin; sa ferme où elle élève un troupeau de vaches.
DP troupeau de vaches dans un champ.
ITW de Valérie MIRAMONT qui a reçu une invitation pour la garden party de l'Elysée: "On est flatté de monter à l'Elysée avec mon fiancé. J'espère pouvoir parler à Jacques CHIRAC".
Rencontre avec Eric FLITTI, jeune étudiant en faculté de droit à Limoges? et ITW : "Je voudrais lui parler des problèmes des jeunes qui éprouvent de plus en plus de mal à
s'insérer dans la vie active et j'espère lui parler de la région".
DP faculté de droit et de sciences économiques de Limoges.
DP Valérie MIRAMONT dans son champ.
Id notice: CAC96039435
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: 00H 01MIN 50SEC

Titre collection: SOIR 3
Titre propre: SORTIR : LES SONNEURS DE CORS DE BRIANCON
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: FRANCE; HAUTES ALPES; NEVACHE; PAYSAGE MONTAGNEUX; TRADITION; MUSIQUE; MUSICIEN; COR; GROUPE MUSICAL
Résumé: Reportage dans les Hautes-Alpes près de Nevache, où un groupe de sonneurs de cors perpétue une tradition centenaire.
Arrivée des sonneurs de cor du groupe *les Briançonneurs* dans la montagne où ils exercent leur art.
PE paysage montagneux.
Sonneur jouant dans un champ de fleur.
GP joueur de cor.
Concert de cor au coeur des alpages.
ITW de Yannick : "C'est un plaisir égoïste dans la montagne".
Sonneurs de cor jouant sur un petit pont en bois.
ITW de Louis du groupe *les Chartreux* : "C'est un son très doux, qui n'agresse pas".
DP sonneurs en train de jouer en plein air.
Randonneurs dans la montagne arrivant dans un refuge.
Id notice: CAC96039437
Date de diffusion: 13.07.1996
Société de programmes: FR3
Durée: 00H 00MIN 20SEC
Titre collection: SOIR 3
Titre propre: VOILE A BREST
Forme: JOURNAL TELEVISE
Descripteurs thématiques - Terme: FRANCE; BRETAGNE; FINISTERE; BREST; PORT; RASSEMBLEMENT; YACHTING; REGATE; VOILIER DE TRADITION; PROUE
Résumé: Les vieux gréements se sont réunis dans la rade de Brest à l'occasion du rassemblement *Brest 96* et participeront à une régate jusqu'à Douarnenez.
PE rade de Brest où sont rassemblés de magnifiques voiliers pour la plupart en bois.
CPL proue d'un voilier sur l'eau.

### A.3.3 Notice de *munich2*

Id notice: CAC97114987
Date de diffusion: 22.06.1997
Titre propre: ATHLETISME : COUPE D'EUROPE
Durée: 01H 43MIN 00SEC
Titre collection: SPORTS DIMANCHE
Résumé: ATHLETISME, Coupe d'Europe des nations à MUNICH : Retransmission en direct des épreuves de la deuxième et dernière journée, sous la pluie. Les équipes de France féminine et masculine sauvent in extremis leur place en première division en obtenant chacune une sixième place rehaussée par l'excellente prestation de Christine ARRON dans un 200m dont Marie-Josée PEREC était la grande absente. Chez les dames, le titre revient aux Russes. Chez les messieurs, ce sont les Britanniques qui l'ont emporté. Retransmission entrecoupée à plusieurs reprises par celle de la coupe du monde d'aviron
100M HAIES Dames : Départ / Course / Arrivée : victoire de Svetlana LAUKHOVA (Rus 12"94) devant Patricia GIRARD (Fra 13"03,3) et Angela THORP (GB 13"16)

lépreuve au ralenti
ITW de Patricia GIRARD, transie après sa course sous la pluie “c’est dégueulasse...avant la course, j’étais frigorifiée”.
Tiérce et Coupure publicitaire.
TRIPLE SAUT Messieurs : Jonathan EDWARDS, vainqueur de l’épreuve, réalise un premier saut à 17,39m
Ralenti (épreuve vue de face.)
JAVELOT Messieurs : Steve BACKLEY, vainqueur de l’épreuve lance à 86,86m
Ralenti.
PERCHE - En raison du mauvais temps, les épreuves se déroulent en salle : Jean GALFIONE passe les 5,40m
Epreuve revue en ralenti
Galfione enfile un pantalon
ITW de Jean GALFIONE sur la contrainte de la salle : “ce n’est pas grave, c’est une question de points pour l’équipe”
ITW de Maurice HOUVION , son entraineur sur le même sujet.
JAVELOT Messieurs : Gatsioudos KONSTANTINOS (Grè), 2ème de l’épreuve
Ralenti du jet.
200M Dames : DP athlètes sur la ligne de départ, Course et arrivée : Christine ARRON arrive la première (22”89) devant Andrea PHILIPP (RFA) et Marina TRANDENKOVA
Ralenti de la fin de course
ITW de C. ARRON : ”je suis très contente, j’étais venue pour gagner..Les 15 derniers mètres ont été très durs”
PERCHE : L’Allemand Tim LOBINGER franchit 5,50m
Ralenti et PR de lathlète.
TRIPLE SAUT Messieurs : Jonathan EDWARDS réalise le saut de la victoire à 17,74m
Ralenti . Il lève les bras, heureux.
800M Messieurs : Départ, course et arrivée en tête de Vebjoern RODAL (Nor) en 1’47”54 devant Nico MOTCHEBON (RFA) et Mark SESAY (GB). Le Français Jimmy JEAN-JOSEPH prend la 4ème place
Ralenti de la fin de course.
SAUT EN LONGUEUR Dames : 2ème essai de la Française Linda FERGA à 6,27 (6,42m au 1er).
ITW de L. RODAL (en anglais) par Nelson MONTFORT : “habitué à ce climat”, les conditions lui ont convenu.
Coupe du monde daviron
PERCHE : Jean GALFIONE rate son premier essai à 5,60m
Ralenti
SAUT EN LONGUEUR Dames : Premier essai de Fiona MAY (Ita) à 6,53m
Ralenti.
PERCHE : Maksim TASAROV, vainqueur de l’épreuve, passe la barre des 5,60m
Ralenti
SAUT EN LONGUEUR Dames : Susen TIDEKE-GREENE, troisième, effectue un essai
Ralenti
1500M Dames : départ et intégrale de la course remportée par Kelly HOLMES (GB - 4’04”79) devant Gabriela SZABO (Rou - 4’06”25) et BIRIOUKOVA (Rus - 4’7”98)
Tribunes vides en raison de la pluie.
PERCHE Messieurs : Jean GALFIONE passe les 5,60m.

MARTEAU Dames : Olga KUZENKOVA avec un jet à 73,10m, pulvérise le record du monde.
TRIPLE SAUT Messieurs : ITW de Jonathan EDWARS peu après sa victoire (en anglais, traduction par Nelson MONTFORT) "très heureux", il parle de l'esprit d'équipe des
Anglais.
SAUT EN LONGUEUR Dames : 2ème essai de Fiona MAY à 6,52m
Ralenti de lépreuve
Coupure publicitaire.
3000M STEEPLE : après plusieurs minutes, la course est menée par le Britannique Robert HOUGH, suivi de A.LAMBRUSCHINI qui manque son dernier passage de haies
Arrivée en tête de R. HOUGH (8'35"03) devant l'Italien A.LAMBRUSCHINI (8'36"15) et le Russe V. PRONIN. Le Français Ali BELGHAZI prend la 4ème place.
PERCHE : Jean GALFIONE rate son premier essai à 5,75m.
HAUTEUR Dames : Heike BALCK, vainqueur, passe 1,94m.
PERCHE : M. TARASOV passe 5,80m
Ralenti de lépreuve.
110M HAIES Messieurs : DP athlètes au départ
Faux départ.
PERCHE : Jean GALFIONE passe 5,75m.
110M HAIES Messieurs : départ et course. Arrivée de l'Allemand SCHWARTHOFF(13"20) en tête devant Colin JACKSON (13"28,3) et A. KISLYKH. Le Français V. CLARICO prend
la 4ème place
F. SCHWARTHOFF, heureux, lève les bras
Colin JACKSON déçu
Ralenti de la course.
PERCHE : GALFIONE échoue à 5,85m au premier essai
Ralenti
ITW de SCHWARTHOFF : "it' very beautiful" vDISQUE Messieurs (Rediffusion de lépreuve de la veille) : Lars RIEDEL, vainqueur de l'épreuve, effectue un jet de 63,36m
Ralenti, PL. du lanceur
PR des jambes de L. RIEDEL au moment du lancer.
PERCHE : essai raté de Tim LOBINGER
Ralenti
DISQUE :Tableau des résultats
Lancer du Français Jean PONS .
200M Messieurs : Sprinters sur la ligne de départ
Faux départ
Course et arrivées en tête ex-aequo du Britannique Linford CHRISTIE (dont c'est la dernière course) et du Grec Georgios PANAYIOTOPOULOS en 20"56
Ralenti de larrivée.
PERCHE : J. GALFIONE échoue à 5,85m, et prend donc la 2ème place derrière TARASOV
Ralenti
ITW de Linford CHRISTIE (en anglais), alors qu'il a annoncé sa retraite prochaine "ça ne devrait pas faire de différence si on est noir ou blanc, chaque pays devrait être fier qu'il y ait des athlètes blancs ou noirs...J'étais là pour donner de la joie et du bonheur.."
AVIRON
PERCHE : M. TARASOV passe 5,95m.
3000M Dames : Départ et intégrale de la course remportée par l'Italienne Roberta BRUNET en

8'51"6 devant l'Allemande Kristina DA FONSECA-WOLHEIM et l'Anglaise Paula RADCLIFFE. La Française Blandine BITZNER prend la 5ème place
Ralenti de larrivée.
3000M Messieurs (diffusion en différé) : Arrivée de Dieter BAUMANN en 7'48 devant Manuel PANCARBO et Papoulias PANAYIOTIS.
Coupure publicitaire et direct AVIRON.
5000M Messieurs : départ et intégrale de la course menée en grande partie par KARTEN
Arrivée en tête de l'Italien Genaro DI NAPOLI (13'38"33) suivi de Anacleto JIMENEZ (13'39"42) et PanaXyiotis PAPOULIAS (13'40"02). Le Français M. ESSAID prend la 5è place.
VA du stade
AVIRON.
4X400M Dames : Intégrale du relais mené et remporté par les Russes KULIKOVA, BAKHVALOVA, KRUSHELEVA, et KOLTYAROVA . 2èmes les Allemandes FELLER, ROHLANDER, RIEGER, et RUCKER . 3èmes les Anglaises. Les Françaises DOMENECH, SCHOLENT, BEVIS et OPHELTES prennent la 7ème place
Ralenti de larrivée. v4X400M Messieurs : Faux départ
Intégrale de la course remportée par les Anglais (BLACK, BAULCH, THOMAS, RICHARDSON) suivis des Italiens (VACCARI, AIMAR, MORI, SABER) et des Russes. Les Français NORDAIN, MANGO, HILAIRE et CHEVAL se classent 4èmes
Ralenti de larrivée
Ralenti de RICHARDSON à la fin de sa course
Drapeaux britanniques s'agitant dans les tribunes
VA du stade.
ITW des relayeurs français, satisfaits d'avoir d'arraché la 6ème place pour leur équipe, évitant ainsi la relégation ( En médaillon, extrait de la course)
ITW Pierre Marie HILAIRE et Rodrigue NORDIN

# Annexe B

# Complément sur la macro-segmentation

## B.1 Algorithme de construction de la hiérarchie ascendante binaire, contrainte temporellement, et calcul de la mesure de cohérence

Nous présentons le pseudo-code de l'algorithme de macro-segmentation ; les notations sont celles définies à la section 5.3.

1. Initialisation : chaque plan $S_i$ est assigné à la classe $C_i$ et la matrice $\tilde{\mathcal{D}}$ est calculée en utilisant la relation 5.1.
2. Sélection de la distance $\tilde{d}_{min} = \min\{\tilde{\mathcal{D}}\}$
   - si $\tilde{d}_{min} = \infty$, la construction de la hiérarchie prend fin et le processus passe au point 4.
   - si plusieurs couples de classes $(C_i, C_j)$ vérifient $\tilde{\mathcal{D}}[i,j] = \tilde{d}_{min}$, le couple constitué des plans dont la somme des durées est minimale est sélectionné et le traitement passe au point 3.
   - sinon le couple de classes correspondant à $\tilde{d}_{min}$ est sélectionné et le traitement passe au point 3.
3. Les deux classes sélectionnées sont regroupées au sein de la hiérarchie, une nouvelle classe est créée et la matrice $\tilde{\mathcal{D}}$ est mise à jour d'après la formule de Lance et William (relation 5.2). Le procédé revient à l'étape 2.
4. La matrice cophénétique $\tilde{\mathcal{D}}_c$ est calculée.
5. La mesure de cohérence $d_m$ est calculée selon la relation 5.4.

## B.2 Un exemple simple

Considérons un ensemble de dix plans, notés $\{S_1, \ldots, S_{10}\}$, représentés sur une ligne temporelle sur la figure B.1 (en haut). Les signatures associées aux plans $\{S_i\}$ sont respectivement $\{10, 50, 20, 35, 40, 30, 20, 60, 10, 80\}$.

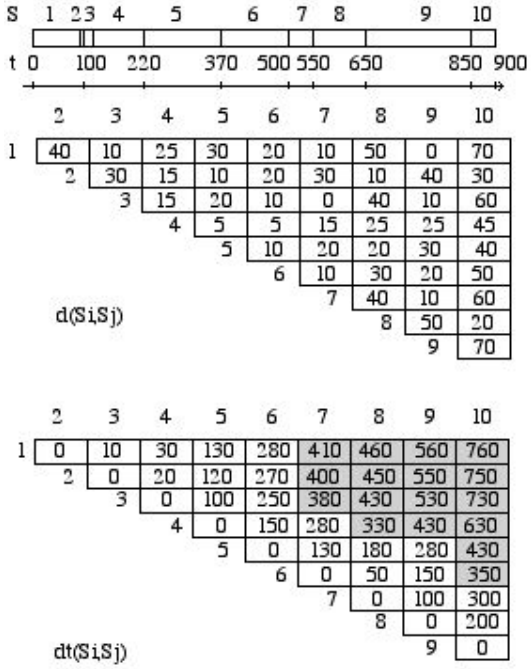



FIG. B.1: *Alignement temporel des segments (en haut), distances physiques (au milieu) et temporelles (en bas) entre plans (en grisé, les distances temporelles supérieures à $\Delta T$)*

L'ensemble des distances $d_{ij} = d(S_i, S_j)$ non contraintes par le temps entre les plans, ainsi que leurs distances temporelles $d_t(S_i, S_j)$ sont données figure B.1, respectivement au milieu et en bas. Dans cet exemple, nous fixons $\Delta T = 300$. Les cases en grisé représentent les plans dont l'éloignement temporel est supérieur à $\Delta T$, les dissimilarités résultantes contraintes temporellement seront alors infinies.

Au début de l'algorithme, chaque classe $C_i$ est initialisée avec un plan $S_i$, la matrice $\tilde{\mathcal{D}} = [\tilde{\delta}(i,j)]$ de dissimilarité contrainte temporellement est calculée à l'aide de l'équation 5.1. Pour cet exemple, la fonction $W$ est supposée constante. La matrice $\tilde{\mathcal{D}}$ est symétrique. Les valeurs $\tilde{d}(S_i, S_j)$ de la matrice $\tilde{\mathcal{D}}$ sont présentées à la figure B.2 sous forme d'une matrice triangulaire supérieure.

À la première itération, les deux classes les plus proches sont regroupées. Plusieurs couples de classes ayant une dissimilarité égale à la valeur minimale 5, le couple présentant l'étendue temporelle la plus faible $(C_4, C_6)$ est alors sélectionné et un premier regroupement est effectué au sein de la classe $C_{11}$ (figure B.2).

La matrice $\tilde{\mathcal{D}}$ est alors mise à jour par la formule de Lance & William (selon la méthode du lien maximal dans cet exemple) et le procédé est itéré comme indiqué à la section B.1. La figure B.3 présente les différentes mises à jour de $\tilde{\mathcal{D}}$ lors des itérations successives. Le traitement prend fin lorsque $\tilde{\mathcal{D}}$ ne contient plus que des valeurs infinies.

La figure B.4 (en haut) présente la hiérarchie binaire ascendante contrainte par le temps issue des regroupements successifs entre classes. La contrainte temporelle s'exprime notamment par la présence de trois sommets dans la hiérarchie (en grisé). Ils correspondent aux éléments dont la dissimilarité inter-classe est infinie. On parle parfois de forêt de hiérarchies. Cette hiérarchie peut être représentée par une matrice $\tilde{\mathcal{D}}_c$, dite matrice cophénétique, présentée figure B.4 (en bas). Les valeurs de $\tilde{\mathcal{D}}_c$ correspondent à la valeur de la distance intra de la classe où les deux plans considérés

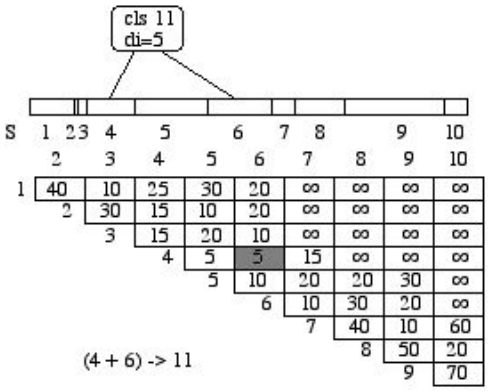



FIG. B.2: *Matrice $\tilde{\mathcal{D}}$ initiale (en bas) et premier regroupement entre les deux classes indiquées en grisé (en haut)*

sont regroupés pour la première fois. Ainsi, $S_1$ et $S_3$ sont regroupés pour la première fois au sein de la classe $C_{13}$ et $\tilde{d}_c(S_1, S_3) = 10$. De même, $S_7$ et $S_8$ sont regroupés pour la première fois au sein de la classe $C_{17}$ et $\tilde{d}_c(S_7, S_8) = 70$. Lorsque deux plans, tels $S_1$ et $S_2$, ne sont jamais regroupés, alors leur distance cophénétique contrainte temporellement est infinie.

Cette matrice $\tilde{\mathcal{D}}_c$ permet le calcul du critère $d_m$, comme suit (pour cet exemple, on fixe $\alpha = 0.0$):

- $d_m(S_1, S_2) = \min\{10, 25, 25, \infty, \dots\} = 10$
- $d_m(S_2, S_3) = \min\{10, 25, 25, 10, \infty, \dots\} = 10$
- $d_m(S_3, S_4) = \min\{25, 25, 10, 25, 25, \infty, \dots\} = 10$
- $d_m(S_4, S_5) = \min\{25, 10, 25, 5, \infty, \dots\} = 5$
- $d_m(S_5, S_6) = \min\{25, 25, 5, \infty, \dots\} = 5$
- $d_m(S_6, S_7) = \min\{\infty, \dots\} = 10$
- $d_m(S_7, S_8) = \min\{70, 10, 70, \infty, \dots\} = 10$
- $d_m(S_8, S_9) = \min\{10, 70, 70, 20, \} = 10$
- $d_m(S_9, S_{10}) = \min\{70, 20, 70, \infty, \dots\} = 20$

La valeur de $d_m$ ne dépend pas uniquement des deux plans frontière considérés. Ainsi, $d_m(S_1, S_2) = 10$, bien que les deux plans ne soient jamais regroupés dans la hiérarchie. La cohérence est forte à cette rupture de plans car il existe de part et d'autre des plans fortement similaires, en l'occurence $S_1$ et $S_3$. Par contre, $d_m(S_6, S_7) = \infty$ parce qu'aucun des plans de part et d'autre de cette rupture ne présente de similarité au sein de la hiérarchie. Ainsi, pour une valeur de seuil de $\delta = 50.0$, deux macro-segments seront proposés regroupant respectivement $\{S_1, S_2, S_3, S_4, S_5, S_6\}$ et $\{S_7, S_8, S_9, S_{10}\}$ (figure B.5).

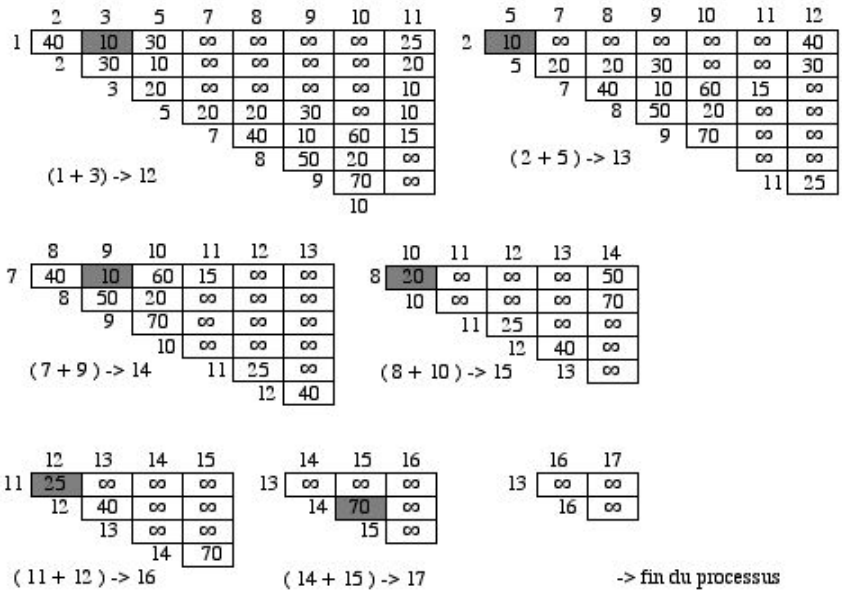
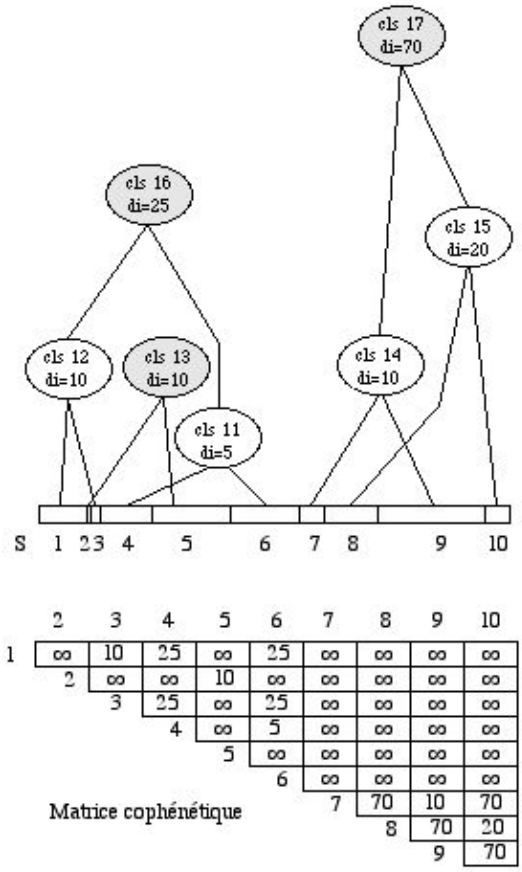



FIG. B.3: *États successifs de la matrice $\tilde{\mathcal{D}}$ lors de la construction de la hiérarchie (en grisé, les classes qui vont être regroupées)*




| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | ∞ | 10 | 25 | ∞ | 25 | ∞ | ∞ | ∞ | ∞ |
| 2 | | ∞ | ∞ | 10 | ∞ | ∞ | ∞ | ∞ | ∞ |
| 3 | | | 25 | ∞ | 25 | ∞ | ∞ | ∞ | ∞ |
| 4 | | | | ∞ | 5 | ∞ | ∞ | ∞ | ∞ |
| 5 | | | | | ∞ | ∞ | ∞ | ∞ | ∞ |
| 6 | | | | | | ∞ | ∞ | ∞ | ∞ |
| 7 | | | | | | | 70 | 10 | 70 |
| 8 | | | | | | | | 70 | 20 |
| 9 | | | | | | | | | 70 |

FIG. B.4: *Hiérarchie ascendante binaire contrainte par le temps (en haut) et matrice cophénétique $\tilde{\mathcal{D}}_c$ associée (en bas)*

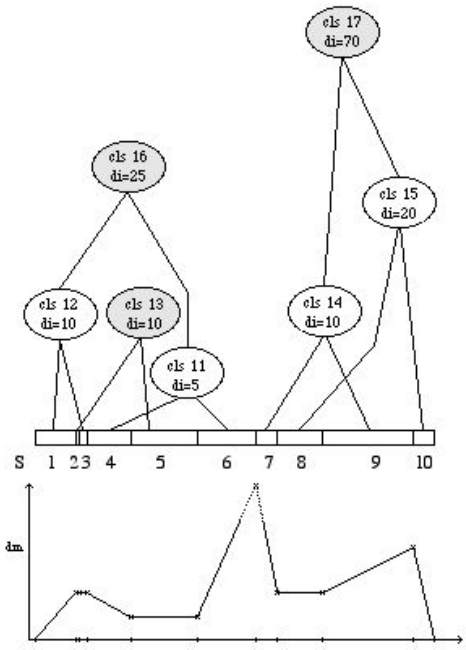



FIG. B.5: *Hiérarchie ascendante binaire contrainte par le temps (en haut) et mesure de cohérence $d_m$ associée (en bas)*

# Annexe C

# De l'apprentissage statistique à l'usage des machines à vecteurs de support

## C.1 Un cadre théorique pour l'apprentissage statistique

### C.1.1 Modélisation de l'apprentissage par l'exemple

L'apprentissage par l'exemple peut être modélisé par un schéma assez simple (voir figure C.1), à travers trois éléments principaux :

- un générateur $G$ de vecteurs aléatoires $x \in \mathbb{R}^n$, tirés indépendamment selon une probabilité de distribution $F(x)$ *fixée* mais *inconnue* ;
- un professeur (ou *supervisor*) $S$ dont la réponse à chaque vecteur d'entrée $x$ est une valeur de sortie $y$, fondée sur la fonction de distribution conditionnelle $F(y|x)$, elle aussi *fixée* et *inconnue* ;
- une machine d'apprentissage $LM$ pouvant mettre en œuvre un ensemble de fonctions $\mathcal{F} = \{f_\alpha = f(x, \alpha),\ \alpha \in \Lambda\}$, où $\Lambda$ est un ensemble de paramètres.

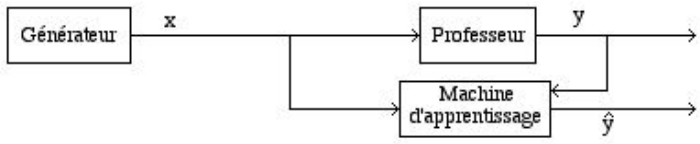



FIG. C.1: *Schéma synoptique de l'apprentissage par l'exemple (figure reprise de [Vapnik 95])*

L'objectif de l'apprentissage est alors de trouver la fonction $f_{\alpha_l}$ de l'ensemble $\mathcal{F}$ qui donne la meilleure approximation de la réponse du professeur $S$. Pour ce faire, il est nécessaire de se munir :

- d'un ensemble d'apprentissage $T = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, constitué de données supposées indépendantes et identiquement distribuées, sur lequel la réponse du professeur $S$ est connue ;

– d'une fonction de coût $L$, mesure de l'adéquation des réponses observées et de celle du professeur $S$.

Le choix de la fonction $f_{\alpha_l}$ se fera par minimisation de la fonction de risque $R$, définie par $R(\alpha) = \int L((y, f(x, \alpha))dF(x, y)$. La distribution $F(x, y) = F(x)F(y|x)$ étant elle aussi inconnue, la minimisation sera effectuée sur l'ensemble d'apprentissage $T$, en substituant à $R$ la fonction du risque empirique $R_{emp}$ définie par $R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(y_i, f(x_i, \alpha))$. C'est le principe de minimisation du risque empirique (*Empirical Risk Minimization principle* - ERM) (voir [Vapnik 95, Sec. 1.4]).

Insistons sur le fait que l'on procède à une double approximation. D'une part, la fonction $f_{\alpha_0}$, définie par $\alpha_0 = argmin\{R(\alpha), \alpha \in \Lambda\}$, est une approximation sur l'ensemble de fonctions $\mathcal{F}$ de la réponse $S$ du professeur. D'autre part, la fonction $f_{\alpha_l}$, définie par $\alpha_l = argmin\{R_{emp}(\alpha), \alpha \in \Lambda\}$, est une approximation de la fonction $f_{\alpha_0}$ sur l'ensemble d'apprentissage $T$.

Une fois la problématique de l'apprentissage par l'exemple modélisée par le principe *ERM*, les questions sont nombreuses : Quelles sont les conditions de convergence de $R(\alpha_l)$ et de $R_{emp}(\alpha_l)$ vers $R(\alpha_0)$ lorsque $l$ tend vers l'infini ? Lorsqu'il y a convergence, quel est alors le taux de convergence ? Celui-ci peut-il être prévu et contrôlé ? Enfin, comment construire des algorithmes fondés sur ces principes tout en contrôlant leur capacité de généralisation ?

Tous ces points, très théoriques, ont été traités par quelques auteurs, et, en premier lieu, Vladimir Vapnik dans [Vapnik 95]. Le lecteur pourra s'y référer, nous ne donnerons ici que les principaux résultats permettant de situer le cadre théorique des machines à vecteurs de support (*Support Vector Machines* - SVM).

### C.1.2 Principaux résultats de la théorie de l'apprentissage statistique

Vladimir Vapnik introduit une première grandeur statistique, liée à la fois à l'ensemble d'apprentissage $T$ et à l'ensemble de fonctions $\mathcal{F}$, appelée entropie de Vapnik-Chervonenkis (ou *VC entropy*) et notée $H^{\Lambda}(l)$. Une condition énoncée sur $H^{\Lambda}(l)$ permet de s'assurer de la consistence (ou *consistency*) de l'apprentissage, et donc de la convergence de $R(\alpha_l)$ et de $R_{emp}(\alpha_l)$ vers $R(\alpha_0)$ [Vapnik 95, Chap. 2].

Deux nouveaux concepts sont ensuite définis pour un ensemble de fonctions permettant de borner, et donc de prévoir et de contrôler, en théorie, le taux de convergence de l'apprentissage. Ces grandeurs statistiques sont l'entropie de Vapnik-Chervonenkis "recuite" (ou *Annealed VC entropy*), notée $H^{\Lambda}_{ann}(l)$ et la fonction de croissance (ou *Growth function*), notée $G^{\Lambda}(l)$. La première permet d'obtenir des résultats pour une distribution donnée, la seconde permet de généraliser les résultats indépendamment de la distribution [Vapnik 95, Chap. 3].

Ces résultats restant plus théoriques que réellement constructifs, V. Vapnik et A.J. Chervonenkis ont étudié plus précisément la fonction de croissance $G^{\Lambda}(l)$. Leur étude, dont le principal résultat est le théorème 1, a conduit à la définition d'une grandeur définie pour un ensemble de fonctions, appelée dimension de Vapnik-Chervonenkis (ou *VC dimension*) et notée $h$ [Vapnik 95, Sec. 3.5].

**Théorème 1** *Toute fonction de croissance $G^{\Lambda}$ soit satisfait l'égalité $G^{\Lambda}(l) = l \ln 2$, soit est bornée selon l'inégalité $G^{\Lambda}(l) \leq h(\ln \frac{l}{h} + 1)$, où $h$ est un entier tel que lorsque $l = h$ alors $G^{\Lambda}(h) = h \ln 2$ et $G^{\Lambda}(h + 1) < (h + 1) \ln 2$.*

Plus intuitivement, la dimension de Vapnik-Chervonenkis d'un ensemble de fonctions $\mathcal{F}$ est le nombre maximum $h$ de vecteurs pouvant être séparés en deux classes des $2^h$ façons différentes

possibles en utilisant les fonctions de $\mathcal{F}$. Si pour tout $n$ il existe un ensemble de $n$ vecteurs vérifiant cette condition, la dimension de Vapnik-Chervonenkis de $\mathcal{F}$ est alors infinie.

Cette grandeur nous assure de résultats sur la convergence de l'apprentissage, et permet, comme nous allons le voir dans la section suivante, de s'assurer de principes constructifs.

## C.2 Définition du principe de minimisation structurelle du risque (*SRM*)

### C.2.1 Introduction au concept de confiance de Vapnik-Chervonenkis

Parmi les résultats présentés succinctement dans la section précédente, l'équation la plus communément considérée dans le cadre de la reconnaissance des formes est l'inégalité C.1 vraie avec la probabilité $1-\eta$ (voir [Burges 98, Eq. 3]).

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}} \tag{C.1}$$

où $h$ est la dimension de Vapnik-Chervonenkis, $l$ le nombre de vecteurs d'apprentissage, et $\eta \in [0,1]$.

Afin de s'assurer de la convergence de $R(\alpha)$, il est nécessaire de minimiser à la fois le risque empirique $R_{emp}(\alpha)$ et le deuxième membre de la relation C.1 appelé confiance de Vapnik-Chervonenkis (ou *VC confidence*). Nous retrouvons, ici, l'idée que pour une tâche d'apprentissage donnée sur un ensemble d'apprentissage fini, il convient de s'assurer à la fois de l'adéquation des résultats obtenus sur l'ensemble d'apprentissage et des capacités de généralisation de la machine.

Le principe de minimisation structurelle du risque (ou *Structure Risk Minimization principle*) permet de gérer le compromis entre l'adéquation des résultats sur l'ensemble d'apprentissage et les capacités de généralisation [Vapnik 95, Chap. 4].

### C.2.2 Principe de minimisation structurelle du risque

Le principe SRM consiste en la définition d'une structure d'ensembles emboîtés de fonctions $(\mathcal{F}_n)$ permettant de choisir le meilleur compromis entre la qualité de l'approximation sur les données considérées et la complexité de la fonction d'approximation. La dimension de Vapnik-Chervonenkis $h_n$ étant croissante, par définition (voir [Burges 98, Sec. 8.2]), sur cette structure d'ensembles emboîtés, nous obtenons, lorsque l'indice $n$ augmente, un risque empirique $R_{emp}(\alpha)$ qui décroît tandis que l'intervalle de confiance de Vapnik-Chervonenkis croît (voir figure C.2).

Dans le cadre de la recherche de ce compromis, deux approches peuvent être envisagées pour minimiser le membre de gauche de la relation C.1. La première consiste à conserver l'intervalle de confiance fixé et de minimiser le risque empirique; les réseaux de neurones en sont une des applications (voir [Vapnik 95, Sec. 5.3]). La seconde propose de fixer le risque empirique (à une valeur nulle, par exemple pour un cas séparable), et de minimiser l'intervalle de confiance; c'est ce qui est réalisé par les SVM lors de la recherche de l'hyperplan séparateur optimal (ou *Optimal Separating Hyperplane - OSH*).

## C.3 Définition de l'hyperplan séparateur optimal (*OSH*)

Considérons qu'un ensemble d'apprentissage $T = \{(x_1, y_1), ..., (x_l, y_l),\ y_i \in \{+1, -1\}\}$ puisse être séparé par un hyperplan $\mathcal{H} : (w.x) + b = 0$. En modifiant $w$ et $b$, il est possible de mettre

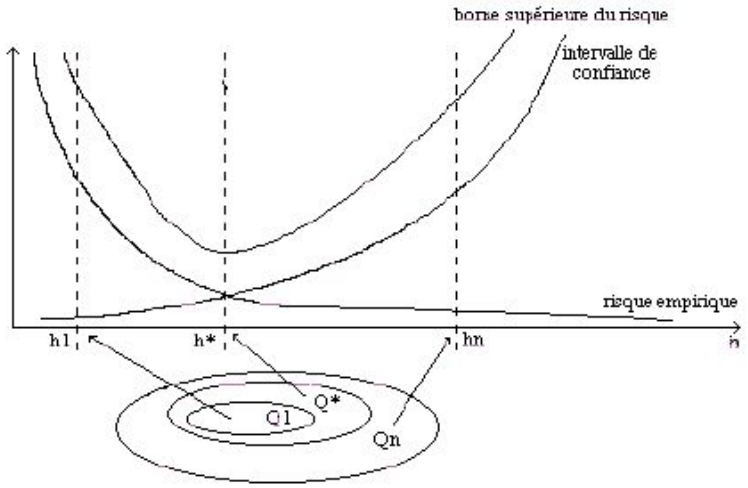



FIG. C.2: *Représentation du principe de minimisation structurelle du risque (SRM), compromis entre la qualité de l'approximation sur les données considérées et la complexité de la fonction d'approximation (figure reprise de [Vapnik 95])*

l'hyperplan $\mathcal{H}$ sous *forme canonique* (Def. 1), nécessitant la définition implicite des deux hyperplans $\mathcal{H}_1 : (w.x) + b = 1$ et $\mathcal{H}_2 : (w.x) + b = -1$.

**Définition 1** *Afin de décrire un hyperplan séparateur, la notation compacte de la* forme canonique *est telle que $w$ et $b$ vérifient: $y_i[(w.x_i) + b] \geq 1$, $i = 1, ..., l$.*

L'hyperplan séparateur optimal (ou *Optimal Separating Hyperplane - OSH*) est l'hyperplan séparateur, satisfaisant aux conditions de la définition 1, qui minimise pour tout $w$ et $b$ la fonction : $\Psi(w) = ||w||^2$.

Plus intuitivement, un ensemble de points est séparé par un *OSH* s'il est séparé sans erreur et si la distance (ou marge) entre le point le plus proche de l'hyperplan et celui-ci est maximale.

## Lien avec les SVM

Considérons les sous-ensembles de fonctions définis par $\mathcal{F}_A = \{f(x, w, b) = sign\{(w.x) + b\}, ||w|| \geq A\}$ paramétrés par $A$, nous obtenons alors, comme précédemment, une structure d'ensembles de fonctions emboîtés. Il a été démontré [Burges 98, Sec. 8.2] que lorsque $A$ croît, la dimension de Vapnik-Chervonenkis $h_A$ de l'ensemble de fonctions $\mathcal{F}_A$ croît aussi, ainsi que, par conséquent, l'intervalle de confiance de Vapnik-Chervonenkis. De plus, la grandeur $A$ est liée à la marge puisque pour tout hyperplan de $\mathcal{F}_A$ sa marge vérifie $\frac{1}{||w||} \geq \frac{1}{A}$.

Ainsi les SVM recherchent, parmi tous les hyperplans séparateurs qui minimisent le risque empirique, celui (i.e. l'*OSH*) pour lequel l'intervalle de confiance est minimal (i.e. la marge est maximale) [Osuna 97b].

Nous présentons à la figure C.3 un exemple illustrant les capacités théoriques de généralisation des SVM liées à la recherche de l'OSH. En noir, les points d'apprentissage d'un problème à deux classes. À gauche (figure C.3.a), un classifieur associé à un hyperplan séparateur quelconque de marge $M_1$. À droite (figure C.3.b), un classifieur associé à l'hyperplan séparateur optimal de marge $M_2$ ($M_2 > M_1$). Les deux classifieurs séparent parfaitement les données d'apprentissage. Si l'on

considère un exemple de test (en clair), légèrement atypique par rapport aux données d'apprentissage de sa classe d'origine, il a plus de chances d'être bien classé par le classifieur associé à l'OSH.

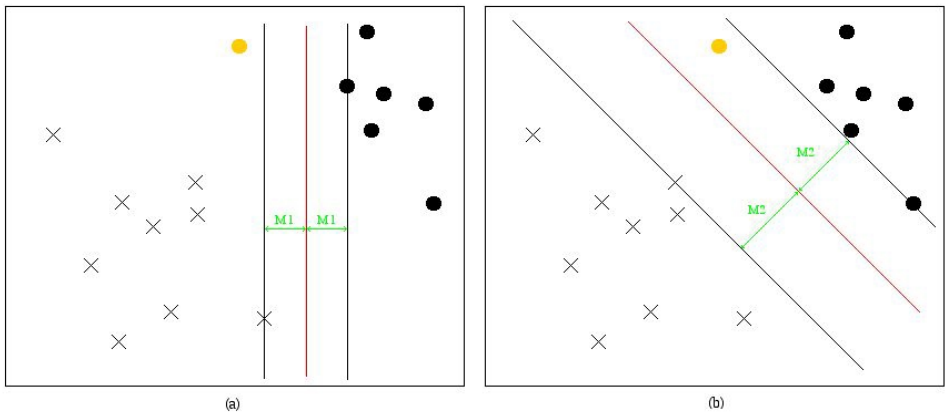



FIG. C.3: *Visualisation de la capacité théorique de généralisation des SVM liée à la recherche de l'hyperplan séparateur optimal (OSH). À gauche un hyperplan séparateur classique, à droite l'hyperplan séparateur optimal.*

Comme nous allons le voir dans la section suivante, la recherche de l'*OSH* nécessite la résolution d'un problème de programmation quadratique. L'algorithme de classification mise en œuvre s'appuie alors sur un nombre réduit de vecteurs $x$ de $T$, ce qui explique le nom donné à cette famille de classifieurs.

## C.4 Mise en œuvre des SVM

### C.4.1 Cas des données séparables

Dans le cadre de données d'apprentissage séparables, la recherche de l'*OSH* peut être formalisée comme suit :

**Problème 1** *Rechercher l'*OSH *qui correspond à la marge maximale en minimisant* $||w||^2$ *sous la contrainte* $y_i[(w.x_i) + b] \geq 1$, $i = 1, ..., l$.

Le problème 1 peut être considéré comme la résolution d'un problème de programmation quadratique. La solution de ce problème d'optimisation contrainte par $y_i[(w.x_i) + b] \geq 1$, $i = 1, ..., l$ sont les points de selle de la fonction de Lagrange (ou Lagrangien) $L$ donnée à l'équation C.2 où les $\alpha_i$ sont les multiplicateurs de Lagrange.

$$L(w, b, \alpha) = \frac{1}{2}(w, w) - \sum_{i=1}^{l} \alpha_i \{[(x.w) + b]y_i - 1\} \tag{C.2}$$

Notons qu'il y a un multiplicateur de Lagrange $\alpha_i$ pour chaque point $(x_i, y_i)$ de l'ensemble d'apprentissage $T$. Si $\alpha_i > 0$, le point correspondant est un vecteur de support appartenant à l'un

des hyperplans $\mathcal{H}_1$ ou $\mathcal{H}_2$ associés à l'*OSH*. En théorie, tous les points dont le multiplicateur de Lagrange est nul pourraient être supprimés de l'ensemble d'apprentissage $T$, sans que cela influe sur le résultat.

### C.4.2 Cas des données non séparables

Dans le cadre de données d'apprentissage non séparables, le problème 1 ne peut être résolu avec l'équation C.2 et il est nécessaire d'assouplir la contrainte en introduisant des variables positives de relachement $\xi_i \geq 0, i = 1, \ldots, l$, telles que :

$$\begin{array}{ll} x_i.w + b \geq 1 - \xi_i & \text{pour } y_i = +1 \\ x_i.w + b \leq -1 + \xi_i & \text{pour } y_i = -1 \end{array}$$

Une nouvelle formulation du problème 1, introduisant une fonction de coût, peut alors être proposée :

**Problème 2** *Minimiser la fonction* $\frac{1}{2}||w||^2 + C(\sum_i \xi_i)^k$, *sous la contrainte* $y_i(x_i.w+b) \geq 1-\xi_i, \forall i \in \{1, \ldots, l\}$, *où* $C$ *est un paramètre fixé représentant la pénalisation des erreurs.*

Comme dans le cas séparable, la résolution du problème est convexe $\forall k > 0$, mais ne reste quadratique que pour $k \in \{1, 2\}$ [Burges 98, Sec. 6.4]. La valeur $k = 1$ est donc retenue dans l'ensemble des travaux étudiés. La principale différence est que les multiplicateurs de Lagrange $\alpha_i$ sont bornés par $C$ : $0 \leq \alpha_i \leq C$. Les vecteurs de support sont toujours définis comme ceux associés à des multiplicateurs de Lagrange $\alpha_i$ non nuls. Toutefois, il est important de différencier les vecteurs de support dont les multiplicateurs de Lagrange associés vérifient $0 < \alpha_i < C$, de ceux pour lesquels $\alpha_i = C$. Les premiers sont de même nature que les vecteurs de support obtenus dans le cas séparable, ils se trouvent à une distance de $\frac{1}{||w||}$ de l'*OSH* et sont appelés vecteurs de marge (*margin vector*). Les seconds correspondent soit à des vecteurs mal classés lors de l'apprentissage, soit à des vecteurs du bon côté de l'*OSH*, mais à une distance de celui-ci inférieure à $\frac{1}{||w||}$, ils sont généralement considérés comme des erreurs [1] [Pontil 98a, Sec. 2.3]. La prise en compte des erreurs par les variables $C$ et $\xi_i$ nécessite la définition implicite d'hyperplan séparateur optimal généralisé (*Generalized Optimal Hyperplane*), ce qui a été fait par Cortes & Vapnik [Vapnik 95, p.132-133]. Notons enfin, une étude de l'influence du paramètre $C$ sur la définition de l'*OSH* dans [Pontil 98a, Sec. 3].

### C.4.3 Cas des SVM non linéaires

#### *C.4.3.1 Projection dans des espaces de dimension supérieure*

L'idée principale, sur laquelle est fondée la généralisation des SVM, est que les données n'apparaissent dans les équations à résoudre que sous forme d'un produit scalaire.

Ainsi, si les données sont projetées dans un autre espace de Hilbert $\mathcal{H}$ à l'aide de la fonction de projection $\Phi : \mathcal{L} = \Re^d \mapsto \mathcal{H}$, alors l'algorithme d'apprentissage n'utilisera les données d'apprentissage que sous la forme du produit scalaire dans $\mathcal{H}$ : $\Phi(x_i).\Phi(x_j)$. L'intérêt d'une telle stratégie est de traiter des données non séparables d'un espace de dimension $\mathcal{L}$ dans un espace $\mathcal{H}$ de plus grande dimension où elles pourraient devenir séparables. En d'autres termes, la projection dans un

1. Dans la partie III, nous y faisons référence sous le vocable de "vecteurs de support *problématiques*".

espace de dimension supérieure permet d'obtenir dans l'espace de départ des surfaces de décision non linéaires.

On appelle fonction noyau (ou *kernel function*), la fonction $K$ telle que $K(u,v) = \Phi(u).\Phi(v)$. Ainsi, seule la connaissance de $K$, et non celle de $\Phi$, sera nécessaire au passage de l'espace $\mathcal{L}$ vers l'espace $\mathcal{H}$. Toutefois, en accord avec la théorie de Hilbert-Schmidt, $K$ doit, malgré tout, être une fonction symétrique vérifiant les conditions générales, dites de Mercer [Cristianini 00, Sec. 3.3.1].

### *C.4.3.2 Exemples de noyaux*

Plusieurs familles de fonctions noyaux ont été proposées [Burges 98, Sec. 7.3], nous en citons quelques unes. Le plus souvent les fonctions $K$ sont données sans référence explicite à la fonction de projection $\Phi$, sauf dans le cas des fonctions de Fourier.

- fonction polynomiales: le produit scalaire est substitué à une fonction du type $K_d(u,v) = [(u.v)+1]^d$ [Vapnik 95, p.139]. On trouve quelques variantes comme $K_d(u,v) = [(u.v)+1]^d - 1$ [Pittore 00, Sec. 5.3.1] ou $K_d(u,v) = (u.v)^d$ [Burges 98, Sec. 8.1.1]. $K_d$ satisfait les conditions Mercer, lorsque $(u,v) \in [-a,a]^n$. Ces noyaux sont censés permettre un réglage assez fin de la dimension de Vapnik-Chervonenkis de la fonction de projection. La dimension de Vapnik-Chervonenkis des SVM utilisant ces noyaux est donnée par $h = C_d^{n+d-1} + 1$, où $n$ est la dimension de l'espace initial des données ($\mathcal{L} = \mathbb{R}^n$) ;

- fonctions à base radiale (*Radial Basis Functions - RBF*): ces fonctions sont définies par $K(u,v) = K_\gamma(|u-v|)$, où $K_\gamma(|u-v|)$ est, pour tout $\gamma$ fixé, une fonction monotone non négative. $K_\gamma(|u-v|) = e^{(-\gamma(u-v)^2)}$ en est un exemple assez répandu [Vapnik 95, p. 140]. Une application issue de cette famille de noyaux est le noyau gaussien (ou *Gaussian Kernel*) [Pittore 00, Sec. 5.3.2] où $\gamma = \frac{1}{2\sigma^2}$. Les vecteurs de support sont alors les centres des fonctions gaussiennes. Sous certaines conditions, la dimension de Vapnik-Chervonenkis peut alors être infinie [Burges 98, Sec. 8.1.2] ;

- réseaux de neurones à deux niveaux : la fonction utilisée est $K(u,v) = S[\kappa(u.v) + c]$ [Vapnik 95, p.141], où $S$ est une fonction sigmoïde. Le noyau sigmoïde, par exemple $K(u,v) = tanh(\kappa(u.v) + c)$, ne satisfait les conditions de Mercer que pour certaines valeurs de $(\kappa, c)$ ;

- splines (mono-dimensionel): le produit scalaire est remplacé par $K(u,v) = \sum_{p=0}^{n} u^p v^p + \sum_{q=1}^{N} (u-t_q)_+^n (v-t_q)_+^n$, où $n$ est le degré du *spline*, $t_q$ les nœuds de la partition, et $(x)_+ = \max(x,0)$ [Pittore 00, Sec. 5.3.3] ;

- noyaux de Dirichlet ou de Fourier : la fonction utilisée lors de la convolution du produit scalaire met en œuvre la transformée de Fourier de laquelle elle est déduite $K(u,v) = \frac{\sin[(N+\frac{1}{2})(u-v)]}{\sin(\frac{(u-v)}{2})}$ (voir [Pittore 00, Sec. 5.3.4] et [Burges 98, Sec. 7.2]).

Dans [Scholkopf 95], une expérimentation est menée sur la reconnaissance de nombres digitaux afin d'évaluer les résultats des divers noyaux proposés. Il semblerait que les meilleurs résultats soient alors obtenus avec des fonctions de la famille RBF, puis avec des fonctions sigmoïdes, et enfin avec des fonctions polynomiales de grande dimension.

## C.5 Construction des SVM

### C.5.1 Utilisation de la fonction de décision

L'ensemble des résultats précédemment évoqués permet de définir la fonction de décision dans l'espace des données initiales, pour les cas linéaire et non linéaire, comme suit :

$$f(x) = sign(w.x + b) = sign\Big( \sum_{i \text{ vecteurs de support}} \big(y_i \alpha_i K(x_i.x)\big) + b \Big)$$

Cette fonction indique la classe de la donnée $x$. Le calcul de $b$, effectué implicitement mais non numériquement lors de l'apprentissage, est donné par les conditions optimales de *Karush Khun Tucker* [Cristianini 00, Sec. 5.2]. En théorie, $b$ peut être calculé en utilisant un seul vecteur de support. Dans la pratique, la valeur de $b$ est obtenue comme une moyenne sur l'ensemble des vecteurs de support.

### C.5.2 Résolution numérique

Diverses méthodes ont été mises en œuvre pour résoudre le problème de programmation quadratique lié au principe des SVM. Nous n'avons pu trouver d'état de l'art synthétique de ces méthodes, et la réalisation d'une telle étude sortait du cadre fixé pour nos travaux. C'est pourquoi nous ne présentons que quelques méthodes rencontrées lors de nos lectures ; la plupart d'entre elles semblent dédiées aux grands ensembles de données et proposent des statégies permettant une résolution du problème quadratique ne nécessitant qu'une place de mémoire centrale modérée[2].

- les méthodes de gradient (*Constrained conjugate gradient ascent*), sont brièvement décrites dans [Burges 98, Sec. 6.8] ;
- les méthodes de projection sont aussi évoquées dans [Burges 98, Sec. 6.8] ;
- une présentation de la décomposition de Bunch-Kaufman est proposée dans cette même référence ;
- les méthodes des points intérieurs (*Interior Point methods*) y sont aussi citées ;
- dans [Burges 98, Sec. 6.8] sont enfin décrites les méthodes mixtes gradient - gradient conjugué (*Mixed gradient and conjugate gradient method*), qui sont utilisées par l'auteur ;
- l'algorithme à pivots complémentaires (*Complementary Pivoting Algorithm* - CPA) met en œuvre des SVM pour des problèmes de classification d'ensembles d'apprentissage de petite taille [Pittore 00, Sec. 5.5.1.1] ;
- la méthode du *chunking* est, au contraire, utilisée pour des ensembles d'apprentissage de grande taille [Pittore 00, Sec. 5.5.2.1] ;
- l'algorithme d'optimisation séquentielle minimale (*Sequential Minimal Optimization* - SMO), de J. Platt, semble avoir rencontré un certain succès dans la communauté concernée [Platt 98] ;

2. Néanmoins, le lecteur désireux d'en savoir plus pourra se reporter aux ouvrages déjà cités ainsi qu'au site `http://www.kernel-machines.org` où sont rassemblées de nombreuses références, et des applications développées par la communauté (voir notamment la contribution de J. Platt dans [Platt 98]).

– la méthode d'Osuna est présentée dans [Osuna 97b, Sec. 3] et plus succinctement dans [Osuna 97a, Sec. 3].

Des informations complémentaires sur les principales méthodes d'apprentissage peuvent être trouvées dans l'étude effectuée par C. Campbellon dans [Campbell 00], et E. Osuna présente diverses méthodes d'apprentissage pour les SVM dans [Osuna 97b, Sec. 3]. Une variante courante est celle des SVM pondérés (*weighted SVM*) permettant d'influencer l'importance relative des exemples d'apprentissage en fonction de leur classe d'origine [Chang 01, Sec. 6]. Enfin, considérant que les SVM sont "couramment nettement plus lents lors de la phase de test que les autres approches ayant des capacités de généralisation similaires", C. Burges propose dans [Burges 96] de modifier la phase d'apprentissage. Celle-ci est alors menée sur une sélection d'ensembles réduits de vecteurs, afin d'obtenir une approximation de la fonction de décision dont l'utilisation sera plus simple et plus rapide[3].

3. à propos de ces extensions de la méthode originelle, voir aussi [Burges 98].

# Annexe D

# Matrices de confusion pour la caractérisation des séquences

Cette annexe regroupe les résultats obtenus dans la partie III sur les bases d'expérimentation décrites au chapitre 9. Les matrices de confusion et les taux de classification correcte sont analysés en détail à la sous-section 9.2.2.

## D.1 Matrices de confusion et taux de classification correcte pour l'expérimentation 1

| $Map_{MHI}$ | *Présentateur* | *Voiture* | *Océan* | *Oiseau* | *Danseur* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Présentateur* | 5 | 0 | 0 | 0 | 0 | 100 |
| *Voiture* | 0 | 5 | 0 | 0 | 0 | 100 |
| *Océan* | 0 | 0 | 5 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 0 | 5 | 0 | 100 |
| *Danseur* | 0 | 0 | 0 | 0 | 5 | 100 |

$T_{pos} =$ 100 %

TAB. D.1: *Matrice de confusion et taux de classification correcte pour les 15 échantillons de test de l'expérimentation 1 avec le descripteur* $Map_{MHI}$ *(*$\tau = 30$*)*

| $Cooc_{MHI}$ | *Présentateur* | *Voiture* | *Océan* | *Oiseau* | *Danseur* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Présentateur* | 0 | 0 | 3 | 1 | 1 | 0 |
| *Voiture* | 0 | 0 | 0 | 4 | 1 | 0 |
| *Océan* | 1 | 1 | 0 | 0 | 3 | 0 |
| *Oiseau* | 0 | 0 | 0 | 5 | 0 | 100 |
| *Danseur* | 0 | 1 | 0 | 0 | 4 | 80 |

$T_{pos} =$ 36 %

TAB. D.2: *Matrice de confusion et taux de classification correcte pour les 15 échantillons de test de l'expérimentation 1 avec le descripteur* $Cooc_{MHI}$ *(*$\tau = 30$*)*

| $H_{Cooc_{MHI}}$ | *Présentateur* | *Voiture* | *Océan* | *Oiseau* | *Danseur* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Présentateur* | 5 | 0 | 0 | 0 | 0 | 100 |
| *Voiture* | 4 | 0 | 1 | 0 | 0 | 0 |
| *Océan* | 0 | 0 | 5 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 0 | 0 | 5 | 0 |
| *Danseur* | 0 | 0 | 0 | 0 | 5 | 100 |
| | | | $T_{pos} =$ | 60 % | | |

TAB. D.3: *Matrice de confusion et taux de classification correcte pour les 15 échantillons de test de l'expérimentation 1 avec le descripteur* $H_{Cooc_{MHI}}$ *(*$\tau = 30$*)*

| $DCT_{MHI}$ | *Présentateur* | *Voiture* | *Océan* | *Oiseau* | *Danseur* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Présentateur* | 5 | 0 | 0 | 0 | 0 | 100 |
| *Voiture* | 1 | 4 | 0 | 0 | 0 | 80 |
| *Océan* | 0 | 0 | 5 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 0 | 5 | 0 | 100 |
| *Danseur* | 0 | 0 | 0 | 0 | 5 | 100 |
| | | | $T_{pos} =$ | 96 % | | |

TAB. D.4: *Matrice de confusion et taux de classification correcte pour les 15 échantillons de test de l'expérimentation 1 avec le descripteur* $DCT_{MHI}$ *(*$\tau = 30$*,* $p = 253$*)*

| $H_{Map_{MHI}}$ | *Présentateur* | *Voiture* | *Océan* | *Oiseau* | *Danseur* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Présentateur* | 5 | 0 | 0 | 0 | 0 | 100 |
| *Voiture* | 3 | 2 | 0 | 0 | 0 | 40 |
| *Océan* | 0 | 2 | 3 | 0 | 0 | 60 |
| *Oiseau* | 0 | 0 | 0 | 5 | 0 | 100 |
| *Danseur* | 0 | 0 | 0 | 0 | 5 | 100 |
| | | | $T_{pos} =$ | 80 % | | |

TAB. D.5: *Matrice de confusion et taux de classification correcte pour les 15 échantillons de test de l'expérimentation 1 avec le descripteur* $H_{Map_{MHI}}$ *(*$\tau = 30$*)*

| $Map_{STG}$ | *Présentateur* | *Voiture* | *Océan* | *Oiseau* | *Danseur* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Présentateur* | 5 | 0 | 0 | 0 | 0 | 100 |
| *Voiture* | 0 | 5 | 0 | 0 | 0 | 100 |
| *Océan* | 0 | 0 | 5 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 0 | 5 | 0 | 100 |
| *Danseur* | 0 | 0 | 0 | 0 | 5 | 100 |
| | | | $T_{pos} =$ | 100 % | | |

TAB. D.6: *Matrice de confusion et taux de classification correcte pour les 15 échantillons de test de l'expérimentation 1 avec le descripteur* $Map_{STG}$ *(*$\tau = 30$*)*

## D.2 Matrices de confusion et taux de classification correcte pour l'expérimentation 2

| $Map_{MHI}$ | *S'approcher* | *Descendre* | *S'éloigner* | *À gauche* | *À droite* | *Monter* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|---|
| *S'approcher* | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| *Descendre* | 0 | 2 | 0 | 0 | 0 | 0 | 100 |
| *S'éloigner* | 0 | 0 | 2 | 0 | 0 | 0 | 100 |
| *À gauche* | 0 | 0 | 0 | 2 | 0 | 0 | 100 |
| *À droite* | 0 | 0 | 0 | 0 | 1 | 1 | 50 |
| *Monter* | 0 | 1 | 0 | 0 | 0 | 1 | 50 |

$T_{pos} = \quad 67\ \%$

TAB. D.7: *Matrice de confusion et taux de classification correcte pour les 12 échantillons de test de l'expérimentation 2 avec le descripteur* $Map_{MHI}$ *(*$\tau = 30$*)*

| $Cooc_{MHI}$ | *S'approcher* | *Descendre* | *S'éloigner* | *À gauche* | *À droite* | *Monter* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|---|
| *S'approcher* | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| *Descendre* | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| *S'éloigner* | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| *À gauche* | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| *À droite* | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| *Monter* | 0 | 0 | 0 | 0 | 2 | 0 | 0 |

$T_{pos} = \quad 0\ \%$

TAB. D.8: *Matrice de confusion et taux de classification correcte pour les 12 échantillons de test de l'expérimentation 2 avec le descripteur* $Cooc_{MHI}$ *(*$\tau = 30$*)*

| $H_{Cooc_{MHI}}$ | *S'approcher* | *Descendre* | *S'éloigner* | *À gauche* | *À droite* | *Monter* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|---|
| *S'approcher* | 1 | 0 | 1 | 0 | 0 | 0 | 50 |
| *Descendre* | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| *S'éloigner* | 0 | 0 | 2 | 0 | 0 | 0 | 100 |
| *À gauche* | 0 | 0 | 1 | 1 | 0 | 0 | 50 |
| *À droite* | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| *Monter* | 0 | 0 | 0 | 0 | 0 | 2 | 100 |

$T_{pos} = \quad 50\ \%$

TAB. D.9: *Matrice de confusion et taux de classification correcte pour les 12 échantillons de test de l'expérimentation 2 avec le descripteur* $H_{Cooc_{MHI}}$ *(*$\tau = 30$*)*

| $DCT_{MHI}$ | *S'approcher* | *Descendre* | *S'éloigner* | *À gauche* | *À droite* | *Monter* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|---|
| *S'approcher* | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| *Descendre* | 0 | 2 | 0 | 0 | 0 | 0 | 100 |
| *S'éloigner* | 0 | 0 | 2 | 0 | 0 | 0 | 100 |
| *À gauche* | 0 | 0 | 0 | 2 | 0 | 0 | 100 |
| *À droite* | 0 | 0 | 0 | 1 | 1 | 0 | 50 |
| *Monter* | 0 | 2 | 0 | 0 | 0 | 0 | 0 |

$T_{pos} = \quad 58\ \%$

TAB. D.10: *Matrice de confusion et taux de classification correcte pour les 12 échantillons de test de l'expérimentation 2 avec le descripteur* $DCT_{MHI}$ *(*$\tau = 30$*,* $p = 91$*)*

| $H_{Map_{MHI}}$ | *S'approcher* | *Descendre* | *S'éloigner* | *À gauche* | *À droite* | *Monter* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|---|
| *S'approcher* | 2 | 0 | 0 | 0 | 0 | 0 | 100 |
| *Descendre* | 0 | 2 | 0 | 0 | 0 | 0 | 100 |
| *S'éloigner* | 0 | 0 | 2 | 0 | 0 | 0 | 100 |
| *À gauche* | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| *À droite* | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| *Monter* | 0 | 2 | 0 | 0 | 0 | 0 | 0 |

$T_{pos} = \quad 50\ \%$

TAB. D.11: *Matrice de confusion et taux de classification correcte pour les 12 échantillons de test de l'expérimentation 2 avec le descripteur* $H_{Map_{MHI}}$ *(*$\tau = 30$*)*

| $Map_{STG}$ | *S'approcher* | *Descendre* | *S'éloigner* | *À gauche* | *À droite* | *Monter* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|---|
| *S'approcher* | 1 | 0 | 1 | 0 | 0 | 0 | 50 |
| *Descendre* | 0 | 2 | 0 | 0 | 0 | 0 | 100 |
| *S'éloigner* | 0 | 0 | 2 | 0 | 0 | 0 | 100 |
| *À gauche* | 0 | 0 | 0 | 1 | 1 | 0 | 50 |
| *À droite* | 1 | 0 | 0 | 0 | 1 | 0 | 50 |
| *Monter* | 0 | 0 | 0 | 0 | 0 | 2 | 100 |

$T_{pos} = \quad 67\ \%$

TAB. D.12: *Matrice de confusion et taux de classification correcte pour les 12 échantillons de test de l'expérimentation 2 avec le descripteur* $Map_{STG}$ *(*$\tau_0 = 0.005$*,* $\tau_1 = 0.01$*,* $\tau_2 = 0.025$*,* $\tau_3 = 0.05$*)*

## D.3 Matrices de confusion et taux de classification correcte pour l'expérimentation 3

| $Map_{MHI}$ | *Divergence* | *Rotation* | *Translation* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|
| *Divergence* | 6 | 0 | 0 | 100 |
| *Rotation* | 0 | 6 | 0 | 100 |
| *Translation* | 0 | 0 | 6 | 100 |

$T_{pos} = 100\ \%$

TAB. D.13: *Matrice de confusion et taux de classification correcte pour les 18 échantillons de test de l'expérimentation 3 avec le descripteur* $Map_{MHI}$ *(*$\tau = 30$*)*

| $Cooc_{MHI}$ | *Divergence* | *Rotation* | *Translation* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|
| *Divergence* | 0 | 2 | 4 | 0 |
| *Rotation* | 0 | 5 | 1 | 83 |
| *Translation* | 0 | 1 | 5 | 83 |

$T_{pos} = 56\ \%$

TAB. D.14: *Matrice de confusion et taux de classification correcte pour les 18 échantillons de test de l'expérimentation 3 avec le descripteur* $Cooc_{MHI}$ *(*$\tau = 30$*)*

| $H_{Cooc_{MHI}}$ | *Divergence* | *Rotation* | *Translation* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|
| *Divergence* | 6 | 0 | 0 | 100 |
| *Rotation* | 5 | 0 | 1 | 0 |
| *Translation* | 0 | 0 | 6 | 100 |

$T_{pos} = 67\ \%$

TAB. D.15: *Matrice de confusion et taux de classification correcte pour les 18 échantillons de test de l'expérimentation 3 avec le descripteur* $H_{Cooc_{MHI}}$ *(*$\tau = 30$*)*

| $DCT_{MHI}$ | *Divergence* | *Rotation* | *Translation* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|
| *Divergence* | 6 | 0 | 0 | 100 |
| *Rotation* | 4 | 0 | 2 | 0 |
| *Translation* | 0 | 0 | 6 | 100 |

$T_{pos} = 67\ \%$

TAB. D.16: *Matrice de confusion et taux de classification correcte pour les 18 échantillons de test de l'expérimentation 3 avec le descripteur* $DCT_{MHI}$ *(*$\tau = 30$, $p = 91$*)*

| $H_{Map_{MHI}}$ | *Divergence* | *Rotation* | *Translation* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|
| *Divergence* | 6 | 0 | 0 | 100 |
| *Rotation* | 3 | 3 | 0 | 50 |
| *Translation* | 0 | 0 | 6 | 100 |

$T_{pos} = 83\ \%$

TAB. D.17: *Matrice de confusion et taux de classification correcte pour les 18 échantillons de test de l'expérimentation 3 avec le descripteur* $H_{Map_{MHI}}$ *(*$\tau = 30$*)*

| $Map_{STG}$ | *Divergence* | *Rotation* | *Translation* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|
| *Divergence* | 6 | 0 | 0 | 100 |
| *Rotation* | 5 | 1 | 0 | 17 |
| *Translation* | 0 | 0 | 6 | 100 |

$T_{pos} = 72\ \%$

TAB. D.18: *Matrice de confusion et taux de classification correcte pour les 18 échantillons de test de l'expérimentation 3 avec le descripteur* $Map_{STG}$ *(*$\tau_0 = 0.005$, $\tau_1 = 0.01$, $\tau_2 = 0.025$, $\tau_3 = 0.05$*)*

## D.4 Matrices de confusion et taux de classification correcte pour l'expérimentation 4

| $Map_{MHI}$ | *Automobile* | *Ballet* | *Eau* | *Oiseau* | *Studio* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Automobile* | 6 | 0 | 0 | 0 | 0 | 100 |
| *Ballet* | 0 | 4 | 0 | 0 | 2 | 67 |
| *Eau* | 0 | 0 | 6 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 1 | 5 | 0 | 83 |
| *Studio* | 0 | 0 | 0 | 0 | 6 | 100 |

$T_{pos} = 90\ \%$

TAB. D.19: *Matrice de confusion et taux de classification correcte pour les 30 échantillons de test de l'expérimentation 4 avec le descripteur* $Map_{MHI}$ *(*$\tau = 30$*)*

| $Cooc_{MHI}$ | *Automobile* | *Ballet* | *Eau* | *Oiseau* | *Studio* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Automobile* | 1 | 3 | 0 | 2 | 0 | 17 |
| *Ballet* | 1 | 0 | 0 | 2 | 3 | 0 |
| *Eau* | 3 | 0 | 0 | 0 | 3 | 0 |
| *Oiseau* | 2 | 2 | 0 | 2 | 0 | 33 |
| *Studio* | 1 | 2 | 0 | 1 | 2 | 33 |

$T_{pos} = 17\ \%$

TAB. D.20: *Matrice de confusion et taux de classification correcte pour les 30 échantillons de test de l'expérimentation 4 avec le descripteur* $Cooc_{MHI}$ *(*$\tau = 30$*)*

| $H_{Cooc_{MHI}}$ | *Automobile* | *Ballet* | *Eau* | *Oiseau* | *Studio* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Automobile* | 0 | 0 | 0 | 4 | 2 | 0 |
| *Ballet* | 0 | 6 | 0 | 0 | 0 | 100 |
| *Eau* | 0 | 1 | 1 | 1 | 3 | 17 |
| *Oiseau* | 0 | 0 | 1 | 5 | 0 | 83 |
| *Studio* | 0 | 0 | 0 | 0 | 6 | 100 |

$T_{pos} = 60\ \%$

TAB. D.21: *Matrice de confusion et taux de classification correcte pour les 30 échantillons de test de l'expérimentation 4 avec le descripteur* $H_{Cooc_{MHI}}$ *(*$\tau = 30$*)*

| $DCT_{MHI}$ | *Automobile* | *Ballet* | *Eau* | *Oiseau* | *Studio* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Automobile* | 3 | 1 | 1 | 0 | 1 | 50 |
| *Ballet* | 0 | 3 | 0 | 0 | 3 | 50 |
| *Eau* | 0 | 0 | 6 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 1 | 5 | 0 | 83 |
| *Studio* | 0 | 0 | 0 | 0 | 6 | 100 |

$T_{pos} = 77\ \%$

TAB. D.22: *Matrice de confusion et taux de classification correcte pour les 30 échantillons de test de l'expérimentation 4 avec le descripteur* $DCT_{MHI}$ *(*$\tau = 30$, $p = 253$*)*

| $H_{Map_{MHI}}$ | *Automobile* | *Ballet* | *Eau* | *Oiseau* | *Studio* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Automobile* | 2 | 0 | 1 | 0 | 3 | 33 |
| *Ballet* | 0 | 0 | 6 | 0 | 0 | 100 |
| *Eau* | 3 | 1 | 0 | 2 | 0 | 0 |
| *Oiseau* | 0 | 0 | 0 | 6 | 0 | 100 |
| *Studio* | 0 | 5 | 0 | 0 | 1 | 17 |

$T_{pos} = 50\ \%$

TAB. D.23: *Matrice de confusion et taux de classification correcte pour les 30 échantillons de test de l'expérimentation 4 avec le descripteur* $H_{Map_{MHI}}$ *(*$\tau = 30$*)*

| $Map_{STG}$ | *Automobile* | *Ballet* | *Eau* | *Oiseau* | *Studio* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Automobile* | 6 | 0 | 0 | 0 | 0 | 100 |
| *Ballet* | 0 | 6 | 0 | 0 | 0 | 67 |
| *Eau* | 3 | 0 | 3 | 0 | 0 | 100 |
| *Oiseau* | 0 | 0 | 0 | 5 | 1 | 83 |
| *Studio* | 0 | 0 | 0 | 0 | 6 | 100 |

$T_{pos} = 87\ \%$

TAB. D.24: *Matrice de confusion et taux de classification correcte pour les 30 échantillons de test de l'expérimentation 4 avec le descripteur* $Map_{STG}$ *(*$\tau_0 = 0.005$, $\tau_1 = 0.01$, $\tau_2 = 0.025$, $\tau_3 = 0.05$*)*

## D.5 Matrices de confusion et taux de classification correcte pour l'expérimentation 5

| Sans compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $Map_{MHI}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 9 | 1 | 0 | 0 | 0 | 90 |
| *Cyclisme* | 0 | 10 | 0 | 0 | 0 | 100 |
| *Formule 1* | 0 | 1 | 6 | 1 | 2 | 60 |
| *Football* | 0 | 0 | 1 | 8 | 1 | 80 |
| *Patinage* | 0 | 0 | 1 | 0 | 9 | 90 |

$T_{pos} = 84\ \%$

| Avec compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $Map_{MHI}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 10 | 0 | 0 | 0 | 0 | 100 |
| *Cyclisme* | 0 | 10 | 0 | 0 | 0 | 100 |
| *Formule 1* | 0 | 1 | 9 | 0 | 0 | 90 |
| *Football* | 0 | 0 | 0 | 10 | 0 | 100 |
| *Patinage* | 0 | 0 | 1 | 0 | 9 | 90 |

$T_{pos} = 96\ \%$

TAB. D.25: *Matrice de confusion et taux de classification correcte pour les 50 échantillons de test de l'expérimentation 5 avec le descripteur* $Map_{MHI}$ *(*$\tau = 50$*)*

| Sans compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $Cooc_{MHI}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 2 | 2 | 1 | 1 | 4 | 20 |
| *Cyclisme* | 0 | 0 | 2 | 4 | 4 | 0 |
| *Formule 1* | 5 | 1 | 0 | 4 | 0 | 0 |
| *Football* | 2 | 6 | 1 | 0 | 1 | 0 |
| *Patinage* | 5 | 0 | 2 | 0 | 3 | 30 |

$T_{pos} = 10\ \%$

| Avec compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $Cooc_{MHI}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 1 | 0 | 2 | 0 | 7 | 10 |
| *Cyclisme* | 3 | 0 | 1 | 1 | 5 | 0 |
| *Formule 1* | 2 | 0 | 0 | 8 | 0 | 0 |
| *Football* | 0 | 2 | 8 | 0 | 0 | 0 |
| *Patinage* | 2 | 2 | 1 | 3 | 2 | 20 |

$T_{pos} = 6\ \%$

TAB. D.26: *Matrice de confusion et taux de classification correcte pour les 50 échantillons de test de l'expérimentation 5 avec le descripteur* $Cooc_{MHI}$ *(*$\tau = 50$*)*

| Sans compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $H_{Cooc_{MHI}}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 8 | 0 | 0 | 0 | 2 | 80 |
| *Cyclisme* | 4 | 0 | 0 | 3 | 3 | 0 |
| *Formule 1* | 1 | 0 | 1 | 6 | 2 | 10 |
| *Football* | 2 | 0 | 1 | 7 | 0 | 70 |
| *Patinage* | 3 | 0 | 0 | 4 | 3 | 30 |

$T_{pos} = 38\ \%$

| Avec compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $H_{Cooc_{MHI}}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 9 | 0 | 1 | 0 | 0 | 90 |
| *Cyclisme* | 3 | 1 | 2 | 4 | 0 | 10 |
| *Formule 1* | 2 | 3 | 0 | 4 | 1 | 0 |
| *Football* | 0 | 0 | 0 | 10 | 0 | 100 |
| *Patinage* | 2 | 6 | 1 | 0 | 1 | 10 |

$T_{pos} = 42\ \%$

TAB. D.27: *Matrice de confusion et taux de classification correcte pour les 50 échantillons de test de l'expérimentation 5 avec le descripteur* $H_{Cooc_{MHI}}$ *(*$\tau = 50$*)*

| Sans compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $DCT_{MHI}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 10 | 0 | 0 | 0 | 0 | 100 |
| *Cyclisme* | 5 | 5 | 0 | 0 | 0 | 50 |
| *Formule 1* | 2 | 0 | 0 | 5 | 3 | 0 |
| *Football* | 0 | 0 | 0 | 8 | 2 | 80 |
| *Patinage* | 0 | 0 | 0 | 1 | 9 | 90 |

$T_{pos} = 64\ \%$

| Avec compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $DCT_{MHI}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 10 | 0 | 0 | 0 | 0 | 100 |
| *Cyclisme* | 4 | 4 | 0 | 2 | 0 | 40 |
| *Formule 1* | 1 | 1 | 6 | 2 | 0 | 60 |
| *Football* | 0 | 0 | 0 | 10 | 0 | 100 |
| *Patinage* | 0 | 0 | 0 | 0 | 10 | 100 |

$T_{pos} = 80\ \%$

TAB. D.28: *Matrice de confusion et taux de classification correcte pour les 50 échantillons de test de l'expérimentation 5 avec le descripteur* $DCT_{MHI}$ *(*$\tau = 50$, $p = 253$*)*

| Sans compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $H_{Map_{MHI}}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 8 | 0 | 1 | 0 | 1 | 80 |
| *Cyclisme* | 5 | 0 | 0 | 2 | 3 | 0 |
| *Formule 1* | 1 | 0 | 1 | 5 | 3 | 10 |
| *Football* | 0 | 1 | 0 | 7 | 2 | 70 |
| *Patinage* | 0 | 0 | 0 | 2 | 8 | 80 |

$T_{pos} = 48\ \%$

| Avec compensation du mouvement de la caméra | | | | | | |
|---|---|---|---|---|---|---|
| $H_{Map_{MHI}}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
| *Aviron* | 10 | 0 | 0 | 0 | 0 | 100 |
| *Cyclisme* | 3 | 1 | 1 | 2 | 3 | 10 |
| *Formule 1* | 2 | 0 | 1 | 5 | 2 | 10 |
| *Football* | 1 | 0 | 0 | 9 | 0 | 90 |
| *Patinage* | 0 | 0 | 0 | 3 | 7 | 70 |

$T_{pos} = 56\ \%$

TAB. D.29: *Matrice de confusion et taux de classification correcte pour les 50 échantillons de test de l'expérimentation 5 avec le descripteur* $H_{Map_{MHI}}$ *(*$\tau = 50$*)*

| $Map_{STG}$ | *Aviron* | *Cyclisme* | *Formule 1* | *Football* | *Patinage* | $T_{pos}(c)$ (%) |
|---|---|---|---|---|---|---|
| *Aviron* | 10 | 0 | 0 | 0 | 0 | 100 |
| *Cyclisme* | 3 | 6 | 1 | 0 | 0 | 60 |
| *Formule 1* | 2 | 1 | 6 | 0 | 1 | 60 |
| *Football* | 0 | 0 | 0 | 8 | 2 | 80 |
| *Patinage* | 0 | 0 | 3 | 0 | 7 | 70 |

$T_{pos} = 74\ \%$

TAB. D.30: *Matrice de confusion et taux de classification correcte pour les 50 échantillons de test de l'expérimentation 5 avec le descripteur* $Map_{STG}$ *(*$\tau_0 = 0.005$, $\tau_1 = 0.01$, $\tau_2 = 0.025$, $\tau_3 = 0.05$*)*

# Annexe E

# Réalisations et développements

## E.1 Environnement de développement

Après quelques travaux dans un environnement *UNIX*, diverses contraintes, dont la participation aux projets *DiVAN* et *AGIR*, nous ont amenés à effectuer nos développements sous *Windows NT*. La programmation des différents algorithmes a été réalisée en langage *C* sous *Visual C++*.

## E.2 Outils mis en œuvre

Le développement des algorithmes, qui représentent plus de 20 000 lignes de code, constitue une part importante de nos travaux. De plus, nous avons collaboré à la réalisation de certains outils d'analyse automatique au sein du GRAMM de l'INA. Enfin, nous avons utilisé et intégré à nos logiciels des algorithmes produits à l'extérieur, et notamment des librairies développées par l'équipe VISTA de l'INRIA Rennes.

Nous proposons, dans cette annexe, quelques indications sur ces développements, partie intégrante du travail de recherche mené, et une brève description des librairies utilisées.

### E.2.1 Outils externes

Parmi les librairies que nous n'avons pas développées nous-même mais que nous avons dû intégrer ou faire coopérer avec nos algorithmes, nous citerons :

***Mpeg1Movie & Mpeg1Encode*** Ces deux librairies ont été développées au GRAMM de l'INA, et nous ont permis de décoder et d'encoder le flux MPEG. *Mpeg1Movie* est inspirée de l'outil *Berkeley MPEG Tools* présenté à l'adresse `http://bmrc.berkeley.edu/frame/research/mpeg/index.html`. *Mpeg1Encode* se fonde sur le développement du *PVRG-MPEG CODEC*, disponible à l'adresse `ftp://mm-ftp.cs.berkeley.edu/pub/multimedia/mpeg/mpeg-2.0.tar.Z`.

***MD_Shots*** Cet outil de l'équipe VISTA de l'INRIA-Rennes [Bouthemy 99b] permet le découpage de la vidéo en plans élémentaires.

***RMR & libMotion-2D*** Ces logiciels d'estimation de modèles paramétriques de mouvement 2D ont été développés au sein de l'équipe VISTA de l'INRIA-Rennes [Odobez 95]. Nous nous en sommes servis plus particulièrement pour la construction des pyramides d'images, et pour l'estimation et la compensation du mouvement dominant.

***C_Motion & S_Activity*** Nous avons utilisé ces logiciels, fournis par l'équipe VISTA de l'INRIA-Rennes [Bouthemy 99b], afin d'extraire les informations liées au mouvement de la caméra et à l'activité en entrée de notre algorithme de macro-segmentation.

***LibSVM 2.1*** *LibSVM*, disponible à l'adresse `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, permet d'effectuer des classifications avec des machines à vecteurs de support (SVM). Les noyaux les plus classiques sont proposés afin d'obtenir des surfaces de décisions non linéaires [Chang 01].

***Intel Image Processing Library*** Nous avons utilisé pour calculer la DCT la librairie *IPLib* fournie par *Intel* à l'adresse `http://developer.intel.com/software/products/perflib/ipl/index.htm`.

***EFEMlib*** Cette librairie propose une implémentation d'une méthode classique de Canny pour l'extraction de contours. Son développement s'inspire largement des travaux de M. Heath *et al.* [Heath 96], disponibles à l'adresse `http://marathon.csee.usf.edu/edge/edge_detection.html`.

***Content Provider Application*** *COPA* est un outil de visualisation de descriptions au format XML. Développé au sein du GRAMM de l'INA [Agir 01], il nous a permis de visualiser les résultats de l'algorithme de macro-segmentation, ainsi que certains descripteurs numériques extraits du flux, en même temps que la vidéo.

***HierarchyVisu*** La visualisation des hiérarchies de classes construites par l'outil de macro-segmentation a été possible *via* cette interface, entièrement développée par le GRAMM de l'INA.

***Phylogeny Inference Package*** Nous avons utilisé l'outil *PHYLIP* pour effectuer les consensus de hiérarchies. Ce programme est distribué par l'auteur sous la référence : Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle.

### E.2.2 Outils développés en collaboration

***KElib*** *KElib* permet d'extraire des images-clefs sur un segment donné. Outre les bornes temporelles du segment et le fichier source MPEG, il est possible d'indiquer en entrée le nombre maximal $k$ d'images-clefs souhaitées. Le principe de fonctionnement est le suivant : (i) un histogramme de couleur est calculé pour chaque image du segment temporel considéré ; (ii) la mesure $L_1$ sur les histogrammes et un algorithme de k-moyennes permet le regroupement des images en au plus $k$ classes par itérations successives ; (iii) les images-clefs sont définies comme les images du segment les plus proches (au sens de $L_1$) des centroïdes des classes obtenues.

***TFEMlib*** L'extraction de descripteurs de textures mise en œuvre est décrite dans [Manjunath 96]. Nous avons simplifié cette méthode fondée sur des filtres de Gabor bidimensionnels en réduisant le banc de filtres utilisé à quatre filtres correspondant à $\sigma_g = 4.0$ et $\omega_s = 0.5$ pour les quatre orientations suivantes : $\{0; \frac{\pi}{4}; \frac{\pi}{2}; \frac{3\pi}{4}\}$.

### E.2.3 Outils développés intégralement

***CFEMlib*** L'extraction de nombreuses signatures de couleur à partir d'une image insérée dans un flux MPEG a été implémentée dans ce module. Les signatures disponibles sont : les histogrammes simples, les histogrammes cumulatifs, les histogrammes localisés par région, les

vecteurs de couleurs cohérentes (CCV), les auto-corrélogrammes, les moments statistiques sur la distribution des couleurs, la liste des couleurs dominantes, les histogrammes des luminances, les moments statistiques sur la distribution des luminances [Swain 91,Huang 98]. L'ensemble de ces signatures sont paramétrables par l'utilisateur. En outre, un certain nombre de mesures sur celles-ci sont disponibles, notamment les distances $L_1$, $L_2$, $L_{inf}$ et leurs variantes [Swain 91], la mesure du $\chi_2$ sur les histogrammes [Brunelli 99], la distance de Hausdorff orientée [Huttenlocher 93] pour les histogrammes localisés par région et la mesure de déplacement de terrain (*Earth Mover's distance* - *EMD*) [Rubner 98] pour les couleurs dominantes et les histogrammes localisés par région.

***HLSlib & HLSprocess*** Ces outils correspondent à la méthode de macro-segmentation définie dans la partie II.

***SFlib & SeqSignComputation*** Cette librairie et l'exécutable associé permettent l'extraction des descripteurs du mouvement, présentés à la section 8.2, sur une base de séquences d'images.

***SVMBuildClassifier*** *SVMBuildClassifier*, qui utilise les fonctions de la librairie *LibSVM*, permet la construction et l'évaluation des classifieurs évoqués à la section 8.3.

***SVMSeqCharacterisation*** Comme *SVMBuildClassifier*, cet outil intègre les fonctions de la librairie *LibSVM*. Il détermine, comme indiqué à la section 8.3, les résultats de la classification des séquences de test, à partir des signatures extraites de celles-ci et des modèles des classifieurs construits précedemment.

#### *Vitesse de calcul*

Nous n'avions pas de contraintes fortes sur les temps de traitement, dans la mesure où l'éventuelle intégration de nos méthodes à la chaîne de documentation de l'INA nécessiterait un développement spécifique de nos prototypes. Nous donnons toutefois, à titre indicatif, les temps de traitement obtenus sur un PC (512 Mb de mémoire physique, processeur PENTIUM III à 1GHz), pour les outils que nous avons développés.

Les temps d'extraction des primitives sont relativement élevés, même s'ils incluent les temps d'accès aux vidéos à travers le réseau du GRAMM à l'INA, et le temps de décodage du flux MPEG. L'extraction des signatures de couleurs sur 550 images-clefs (soit environ une heure de vidéo) prend entre 15 et 30 minutes en fonction de la nature de la signature extraite, ce qui correspond environ à 60 fois le temps réel. Le calcul des MHI est obtenu aux alentours de 20 fois le temps réel. Le traitement des séquences par les filtres de Gabor est extrêmement lent : nous dépassons 4 000 fois le temps réel ! Le traitement est, en effet, très lourd, dans la mesure où il faut effectuer la convolution de la séquence d'images avec 24 filtres (soit 4 triades de filtres à phase sinus et cosinus), sur 4 niveaux de résolution d'image, puis calculer en chaque point les réponses des filtres d'énergie, les réponses des triades, et enfin concaténer l'information associée aux différents niveaux de la pyramide et aux différentes orientations. De plus, nous n'avons pas mené de recherches sur la synthèse de ces filtres et notre implémentation est très "naïve". Notons, toutefois que la méthode n'est pas pour autant condamnée, dans la mesure où des mises en œuvre efficaces et rapides ont été proposées au travers d'implémentations récursives des filtres [Spinei 98b] ou par le recours technique à des DSP (*Digital Signal Processor*) [Spinei 00].

Les autres modules sont nettement plus rapides. La segmentation en séquences d'un document d'une heure est effectuée en cinq minutes, une fois les primitives extraites. L'apprentissage des classifieurs et leur évaluation sur les exemples de tests tournent en moins de cinq minutes pour une

base de 150 séquences dont ont été extraits des descripteurs de taille $65 \cdot 10^3$. Lors de l'utilisation conjointe des classifieurs sur cette même base pour la caractérisation des séquences, les résultats sont aussi obtenus en près de cinq minutes.

# Bibliographie

[Adams 00] Brett Adams, Chitra Dorai, Svetha Venkatesh. – Novel approach to determining tempo and dramatic story sections in motion pictures. – *International Conference on Image Processing*, Vancouver, septembre 2000.

[Adelson 85] Edward H. Adelson, James R. Bergen. – Spatiotemporal energy models for the perception of motion. *Journal of Optical Society of America*, 2(2):284–299, 1985.

[Adelson 86] Edward H. Adelson, James R. Bergen. – The extraction of spatio-temporal energy in human and machine vision. – *Workshop on Motion: Representation and Analysis*, pp. 151–155, Charleston, mai 1986.

[AFNOR 96] AFNOR. – Information et documentation, principes généraux pour l'indexation des documents (AFNOR NF Z: 47- 102). *Documentation, présentation des publications, traitements documentaire et gestion de bibliothèques*, pp. 509–518. – Association Française de Normalisation, 1996.

[Aggarwal 99] J.K. Aggarwal, Q. Cai. – Human motion analysis: a review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999.

[Agir 01] Consortium Agir. – *Rapport de fin de phase I.* – F052, Projet RNRT AGIR, 2001.

[Aigrain 94] Philippe Aigrain, Philippe Joly. – The automatic real-time analysis of film editing and transition effects and its applications. *Computers & Graphics*, 18(1):93–103, 1994.

[Aigrain 96] Philippe Aigrain, Hongjiang Zhang, Dragutin Petkovic. – Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications*, 3(3):179–202, 1996.

[Aigrain 97] Philippe Aigrain, Philippe Joly, Véronique Longueville. – Medium knowledge-based macro-segmentation of video into sequences. *Intelligent Multimedia Information Retrieval*, éd. par Mark T. Maybury, pp. 159–173. – AAAI/MIT Press, 1997.

[Aksoy 98] Selim Aksoy, Robert M. Haralick. – Textural features for image database retrieval. – *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 45–49, Santa Barbara, juin 1998.

[Aksoy 00] Selim Aksoy, Robert M. Haralick, Faouzi A. Cheikh, Moncef Gabbouj. – A weighted distance approach to relevance feedback. – *International Conference on Pattern Recognition*, vol. 4, pp. 812–815, Barcelone, septembre 2000.

[Anandan 95] P. Anandan, Michal Irani, Rakesh Teddy Kumar, Jim Bergen. – Video as an image data source : Efficient representations and applications. – *International Conference on Image Processing*, vol. 1, pp. 318–321, Washington, octobre 1995.

[Aoki 96] Hisashi Aoki, Shigeyoshi Shimotsuji, Osamu Hori. – A shot classification method of selecting effective key-frames for video browsing. – *ACM Multimedia Conference*, pp. 1–10, Boston, novembre 1996.

[Ardizzone 96] Edoardo Ardizzone, Marco LaCascia. – Video indexing using optical flow field. – *International Conference on Image Processing*, vol. 3, pp. 831–834, Lausanne, septembre 1996.

[Ariki 96] Y. Ariki, Y. Saito. – Extraction of TV news articles based on scene cut detection using dct clustering. – *International Conference on Image Processing*, vol. 3, pp. 847–850, Lausanne, septembre 1996.

[Auffret 00] Gwendal Auffret. – *Structuration de documents audiovisuels et publication électronique*. – Thèse de doctorat, Université Technologique de Compiègne, 2000.

[Bachimont 98] Bruno Bachimont. – Bibliothèques numériques audiovisuelles : des enjeux scientifiques et techniques. *Document numérique*, 2(3-4):219–242, 1998.

[Barabino 99] N. Barabino, M. Pallavicini, A. Petrolini, M. Pontil, A. Verri. – Support vector machines vs multi-layer perceptrons in particle identification. – *European Symposium on Artificial Neural Networks*, pp. 257–262, Bruges, avril 1999.

[Baras 02] Claude Baras, Alexandre Allauzen, Lori Lamel, Jean-Luc Gauvain. – Transcribing audio video archives. – *International Conference on Acoustics, Speech and Signal Processing*, Orlando, mai 2002.

[Barron 94] J.L. Barron, D.J. Fleet, S.S. Beauchemin. – Performances of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.

[Bigun 94] Josef Bigun. – Speed, frequency, and orientation tuned 3-d gabor filter banks and their design. – *International Conference on Pattern Recognition*, pp. 184–187, Jerusalem, octobre 1994.

[Bimbo 95] Alberto Del Bimbo, Enrico Vicario, Daniele Zingoni. – Symbolic description and visual querying of image sequences using spatio-temporal logic. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):609–621, 1995.

[Bimbo 00] Alberto Del Bimbo. – Issues and directions in visual information retrieval. – *International Conference on Pattern Recognition*, vol. 4, pp. 31–38, Barcelone, septembre 2000.

[Black 95] Michael J. Black, Yaser Yacoob. – Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1995.

[Black 97] Michael J. Black, Yaser Yacoob, Shanon X. Ju. – Recognizing human motion using parameterized models of optical flow. *Motion-based Recognition*, éd. par Shah & Jain. – Kluwer Academic Publishing, 1997.

[Black 98] Michael J. Black, David J. Fleet, Yaser Yacoob. – A framework for modelling appearance change in images sequences. – *International Conference on Computer Vision*, pp. 660–667, Bombay, janvier 1998.

[Blanz 96] V. Blanz, B. Scholkopf, H. Bulthoff, C. Burges, V. N. Vapnik, T. Vetter. – Comparison of view-based object recognition algorithms using realistic 3d models. – *International Conference on Artificial Neural Networks*, vol. LNCS 1112, pp. 251–256, Berlin, juillet 1996.

[Bobick 96] Aaron F. Bobick, James W. Davis. – An appearance-based representation of action. – *International Conference on Pattern Recognition*, pp. 307–312, Vienne, août 1996.

[Bobick 97] Aaron F. Bobick. – Movement, activity and action: the role of knowledge in the perception of motion. *Phil. Trans. Royal Society London B*, 352:1257–1265, 1997.

[Bobick 98] Aaron F. Bobick, Y.A. Ianov. – Action recognition using probabilistic parsing. – *International Conference on Computer Vision and Pattern Recognition*, pp. 196–202, Santa Barbara, juin 1998.

[Bobick 01] Aaron F. Bobick, James W. Davis. – The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[Bolle 98] R. M. Bolle, B.-L. Yeo, M. M. Yeung. – Video query : Research directions. *IBM Journal of Research and Development*, 42(2):233–252, 1998.

[Boreczky 98] John S. Boreczky, Lynn D. Wilcox. – A hidden Markov model framework for video segmentation using audio and image features. – *International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, pp. 3741–3744, Seattle, mai 1998.

[Bouthemy 99a] Patrick Bouthemy, Christophe Garcia, Rémi Ronfard, G. Tziritas, Emmanuel Veneau, Didier Zugaj. – Scene segmentation and image feature extraction for video indexing and retrieval. – *International Conference on Visual Information Systems*, vol. LNCS 1614, pp. 245–252, Amsterdam, juin 1999.

[Bouthemy 99b] Patrick Bouthemy, Marc Gelgon, François Ganansia. – A unified approach to shot change detection and camera motion characterization. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(7):1030–1044, 1999.

[Bradski 00] Gary R. Bradski, James Davis. – Motion segmentation and pose recognition with motion history gradients. – *IEEE Workshop on Applications of Computer Vision*, pp. 238–244, Palm Springs, décembre 2000.

[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*, 10(2):78–112, 1999.

[Bruno 01] Eric Bruno, Denis Pellerin. – Global motion model based on b-spline wavelets: Application to motion estimation and video indexing. – *International Symposium on Image and Signal Processing and Analysis*, Pula, juin 2001.

[Burges 96] Christopher J. C. Burges. – Simplified support vector decision rules. – *International Conference on Machine Learning*, pp. 71–77, Bari, juillet 1996.

[Burges 98] Christopher J.C. Burges. – A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2):121–167, 1998.

[Campbell 00] Colin Campbell. – Algorithmic approaches to training support vector machines : A survey. – *European Symposium on Artificial Neural Networks*, pp. 27–36, Bruges, avril 2000.

[Canny 86] J. Canny. – A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[Carrive 98] Jean Carrive, François Pachet, Rémi Ronfard. – Using description logics for indexing audiovisual documents. – *International Workshop on Description Logics*, pp. 116–120, Trento, juin 1998.

[Carrive 00] Jean Carrive. – *Classification de séquences audiovisuelles.* – Thèse de doctorat, Université de Paris VI, 2000.

[Castel 96] Charles Castel, Laurent Chaudron, Catherine Tessier. – What's going on? a high level interpretation of sequences of images. – *Workshop on Conceptual Descriptions from Images*, Cambridge, avril 1996.

[Cedras 95] Claudette Cedras, Mubarak Shah. – Motion-based recognition : a survey. *Image and Vision Computing*, 12(2):129–155, 1995.

[Chanal 93] Marc Chanal, Gilbert Pineau, Maté Rabinovsky, Régis Brugière, Sylvie Dargnies, Rosine Gautier, Yves Turquier. – *Les Techniques audiovisuelles : vidéo & film (principes - outils - pratiques).* – Economica, Institut National de l'Audiovisuel - Polytechnica, Paris, 1993.

[Chandler 94] Daniel Chandler. – *The 'Grammar' of Television and Film.* – Rapport de recherche, University of Wales, 1994.

[Chang 87] Shi-Kuo Chang, Qing-Yun Shi, Cheng-Wen Yan. – Iconic indexing by 2-D strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(3):413–428, 1987.

[Chang 96] Yuh-lin Chang, Wenjun Zeng, Ibrahim Kamel, Rafael Alonso. – Integrated image and speech analysis for content-based video indexing. – *International Conference on Multimedia Computing and Systems*, pp. 306–313, Tokyo, mai 1996.

[Chang 97] Shih-Fu Chang, William Chen, Horace J. Meng, Hari Sundaram, Di Zhong. – VideoQ : An automated content based video search system using visual cues. – *ACM Multimedia Conference*, pp. 313–324, Seattle, novembre 1997.

[Chang 98] S.-F. Chang, W. Chen, H. Meng, Hari sundaram, Di Zhong. – A fully automated content based video search engine supporting spatio-temporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):602–615, 1998.

[Chang 01] Chih-Chung Chang, Chih-Jen Lin. – *LIBSVM: a library for support vector machines.* – Rapport technique, Department of Computer Science and Information Engineering, National Taiwan University, 2001.

[Cherfaoui 94] Mourad Cherfaoui, Christian Bertin. – Two-stage strategy for indexing and presenting video. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 2185, pp. 174–184, février 1994.

[Cherfaoui 95] Mourad Cherfaoui, Christian Bertin. – Temporal segmentation of videos : a new approach. – *IS&T Symposium on Electronic Imaging Science and Technology (Digital Video Compression and Processing on Personal Computers : Algorithms and Technologies)*, pp. 38–47, San Jose, février 1995.

[Chomat 99a] Olivier Chomat, James L. Crowley. – Probabilistic recognition of activity using local appearance. – *International Conference on Computer Vision and Pattern Recognition*, pp. 104–109, Fort Collins, juin 1999.

[Chomat 99b] Olivier Chomat, James L. Crowley. – Utilisation de champs réceptifs spatio-temporels pour la reconnaissance de l'apparence locale d'activités. – *Journées des jeunes chercheurs francophones en vision par ordinateur*, Aussois, avril 1999.

[Chomat 00] Olivier Chomat. – *Caractérisation d'éléments d'activités par la statistique conjointe de champs réceptifs.* – Thèse de doctorat, Institut National Polytechnique de Grenoble, 2000.

[Clergue 95] E. Clergue, M. Goldberg, N. Madrane, B. Merialdo. – Automatic face and gestual recognition for video indexing. – *International Workshop on Automatic Face and Gesture Recognition*, pp. 110–115, Zurich, juin 1995.

[Courtney 97] Jonathan D. Courtney. – Automatic video indexing via object motion analysis. *Pattern Recognition*, 30(4):607–625, 1997.

[Cristianini 00] Nello Cristianini, John Shawe-Taylor. – *An introduction to Support Vector Machines.* – Cambridge University Press, 2000.

[Cui 95] Yuntao Cui, Daniel L. Swets, John J. Weng. – Learning-based hand sign recognition using SHOSLIF-M. – *International Conference on Computer Vision*, pp. 45–58, Boston, 1995.

[Dailianas 95] Apostolos Dailianas, Robert B. Allen, Paul England. – Comparison of automatic video segmentation algorithms. – *Photonics West*, vol. SPIE 2615, pp. 2–16, Philadelphie, octobre 1995.

[Davenport 91] Glorianna Davenport, Thomas Aguierre Smith, Natalio Pincever. – Cinematic primitives for multimedia. *IEEE Computer Graphics & Applications*, 11(4):67–74, 1991.

[Davis 96] James W. Davis. – *Appearance-Based Motion Recognition of Human Actions.* – Master of Science n° 387, M.I.T. Media Lab Perceptual Computing Group, juillet 1996.

[Davis 97] James W. Davis, Aaron Bobick. – The representation and recognition of action using temporal templates. – *International Conference on Computer Vision and Pattern Recognition*, pp. 928–934, Puerto Rico, juin 1997.

[Davis 98] James W. Davis, Aaron F. Bobick. – A robust human-silhouette extraction technique for interactive virtual environments. – *IFIP Workshop on Modeling and Motion Capture Techniques for Virtual Environments*, vol. LNAI 1537, pp. 12–25, Heidelberg, 1998.

[Davis 99a] James W. Davis. – *Recognizing Movement using Motion Histograms.* – Rapport technique n° 487, MIT Media Lab, Perceptual Computing Group, mars 1999.

[Davis 99b] James W. Davis, Gary Bradski. – Real-time motion template gradients using Intel CVLib. – *ICCV Workshop on Frame-rate Vision*, septembre 1999.

[Deriche 87] Rachid Deriche. – Using Canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, 1(2):167–187, 1987.

[Dimitrova 95] Nevenka Dimitrova. – The myth of semantic video retrieval. *ACM Computing Surveys*, 27(4):584–586, 1995.

[Duhen ] Jacques Duhen. – *Production et réalisation TV.* – Manuel de formation (années 70), Centre de Formation des Personnels, Technique et de Production, Office de Radiodiffusion Télévision Française (ORTF).

[Dumais 98a] S. Dumais. – Using SVMs for text categorization. *IEEE Intelligent Systems, Trends and controversies on support vector machines*, 13(4):18–28, 1998.

[Dumais 98b] S. T. Dumais, J. Platt, D. Heckerman, M. Sahami. – Inductive learning algorithms and representations for text categorization. – *ACM Conference on Information and Knowledge Management*, pp. 148–155., Bethesda, novembre 1998.

[Fablet 01] Ronan Fablet. – *Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images; application à l'indexation vidéo.* – Thèse de doctorat, Université de Rennes I, IRISA, 2001.

[Felsenstein 89] J. Felsenstein. – PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[Fischer 95] Stephan Fischer, Rainer Lienhart, Wolfgang Effelsberg. – Automatic recognition of film genres. – *ACM Multimedia Conference*, pp. 295–304, San Francisco, novembre 1995.

[Garrett 84] George P. Garrett, O. B. Hardison, Jane R. Gelfman. – Introduction - film terms in context. *Darling, A Hard Day's Night, The Best Man*, pp. 24–35. – New-York, Irvington, 1984.

[Gauvain 90] Jean-Luc Gauvain, Lori F. Lamel, Maxime Esknazi. – Design considerations and text selection for BREF, a large french read-speech corpus. – *International Conference on Speech and Language Processing*, vol. 2, pp. 1097–2000, Kobe, novembre 1990.

[Gauvain 98] Jean-Luc. Gauvain, Lori Lamel, Gilles Adda. – Partitioning and transcription of broadcast news data. – *International Conference on Speech and Language Processing*, vol. 4, pp. 1335–1338, Sydney, dècembre 1998.

[Gelgon 98] Marc Gelgon, Patrick Bouthemy. – Determining a structured spatio-temporal representation of video content for efficient visualisation and indexing. – *European Conference on Computer Vision*, vol. 1, pp. 595–609, Fribourg, juin 1998.

[Gelgon 99] Marc Gelgon, Patrick Bouthemy, T. Dubois. – A region tracking method with failure detection for an interactive video indexing environment. – *International Conference on Visual Information Systems*, vol. LNCS 1614, pp. 261–268, Amsterdam, juin 1999.

[Gong 95] Yihong Gong, Lim Teck Sin, Chua Hock Chuan, Hongjiang Zhang, Masao Sakauchi. – Automatic parsing of TV soccer programs. – *International Conference on Multimedia Computing and Systems*, pp. 167–174, Vancouver, mai 1995.

[Gordon 99] Allan D. Gordon. – *Classification.* – Chapman and Hall/CRC, 2nd edition édition, 1999.

[Gotlieb 90] Calvin C. Gotlieb, Herbert E. Kreyszig. – Texture descriptors based on co-occurrence matrices. *Computer Vision, Graphics, and Image Processing*, 51(1):70–86, 1990.

[Gould 89] Kristine Gould, Mubarak Shah. – The trajectory primal sketch: A multi-scale scheme for representing motion characteristics. – *Conference on Computer Vision and Pattern Recognition*, pp. 79–85, San Diego, juin 1989.

[Gourdon 99] Anne-Marie Gourdon. – La coupe du monde 98 (la technologie au service du spectaculaire). *Des arts et des spectacles à la télévision. Le regard du téléspectateur*, éd. par Anne-Marie Gourdon, p. 200. – Paris, CNRS Éditions, 1999.

[Gudivada 95] Venkat N. Gudivada, Vijay V. Raghavan. – Content-based image retrieval systems. *IEEE Computer*, 28(9):18–22, 1995.

[Gunsel 98a] Bilge Gunsel, A. Mfit Ferman, A. Murat Tekalp. – Temporal video segmentation using unsupervised clustering and semantic object tracking. *Journal of Electronic Imaging*, 7(3):592–604, 1998.

[Gunsel 98b] Bilge Gunsel, A. Murat Tekalp, Peter J.L. Van Beek. – Content-based access to video objects : Temporal segmentation, visual summarization, and feature extraction. *Signal Processing*, 66(2):261–280, 1998.

[Guo 00] Guodong Guo, Stan Z. Li, Kap Luk Chan. – Learning similarity for texture image retrieval. – *European Conference on Computer Vision*, pp. 178–190, Dublin, juin 2000.

[Gutschoven 00] Bert Gutschoven, Patrick Verlinde. – Multi-modal identity verification using support vector machines. – *International Conference on Information Fusion*, pp. 25–30, Paris, juillet 2000.

[Haering 99] Niels Haering, Richard J. Qian, M. Ibrahim Sezan. – A semantic event detection approach and its application to detecting hunts in wildlife video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6):857–868, 1999.

[Hammoud 98] Riad Hammoud, Liming Chen, Dominique Fontaine. – An extensible spatial-temporal model for semantic video segmentation. – *International Forum on Multimedia and Image Processing*, Anchorage, mai 1998.

[Hampapur 95] Arun Hampapur, Ramesh Jain, Terry E. Weymouth. – Production model based digital video segmentation. *Multimedia Tools and Applications*, 1(1):9–46, 1995.

[Hampapur 97] Arun Hampapur, Amarnath Gupta, Bradley Horowitz, Chiao-Fe Shu, Chuck Fuller, Jeffrey Bach, Monika Gorkani, Ramesh Jain. – Virage video engine. – *Conference on Storage and Retrieval for Images and Video Databases*, vol. SPIE 3022, pp. 188–197, San Jose, février 1997.

[Hanjalic 99] Alan Hanjalic, Reginald L. Lagendijk, Jan Biemond. – Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Transactions on Circuits and Systems for video technology*, 9(4):580–587, 1999.

[Haralick 73] Robert M. Haralick, K. Shanmugan, Its'hak Dinstein. – Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973.

[Hauptmann 95] Alexander G. Hauptmann, Michael A. Smith. – Text, speech, and vision for video segmentation : the Informedia project. – *AAAI Fall Symposium, Computational Models for Integrating Language and Vision*, Boston, novembre 1995.

[Hauptmann 98] Alexander G. Hauptmann, Michael J. Witbrock. – Story segmentation and detection of commercial in broadcast news video. – *Advances in Digital Librairies Conference*, Santa Barbara, avril 1998.

[Heath 96] M. Heath, S. Sarkar, T. Sanocki, K. Bowyer. – Comparison of edge detectors: a methodology and initial study. – *International Conference on Computer Vision and Pattern Recognition*, pp. 143–148, San Francisco, juin 1996.

[Heeger 87] David J. Heeger. – Model for the extraction of image flow. *Journal of Optical Society of America*, 4(8):1455–1471, 1987.

[Heeger 88] David J. Heeger. – Optical flow using spatiotemporal filters. *International Journal of Computer Vision*, 1:279–302, 1988.

[Horn 81] B. Horn, B. Schunck. – Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[Hsieh 00] Ing-Sheen Hsieh, Kuo-Chin Fan. – An adaptative clustering algorithm for color quantization. *Pattern Recognition Letters*, 21:337–346, 2000.

[Hu 62] M. Hu. – Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[Huang 97] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, Ramin Zabih. – Image indexing using color correlograms. – *International Conference on Computer Vision and Pattern Recognition*, pp. 762–768, San Juan, juin 1997.

[Huang 98] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, Ramin Zabih. – Spatial color indexing and applications. *International Journal of Computer Vision*, pp. 245–268, 1998.

[Hunter 99] Jane Hunter, Liz Armstrong. – A comparison of schemas for video metadata representation. – *International World Wide Web Conference*, pp. 353–373, Toronto, mai 1999.

[Huttenlocher 93] Daniel P. Huttenlocher, Gregory A. Klanderman, William J. Rucklidge. – Comparing images using the hausdorff-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[Ide 98] Ichiro Ide, Koji Yamamoto. – Automatic video indexing based on shot classification. – *International Conference on Advanced Multimedia Content Processing*, pp. 87–102, Osaka, novembre 1998.

[Idris 97] F. Idris, S. Panchanathan. – Review of image and video indexing techniques. *Journal of Visual Communication and Image Representation*, 8(2):146–166, 1997.

[Intille 94] Stephen S. Intille, Aaron Bobick. – *Tracking using a local closed-world Assumption: tracking in the football domain.* – Master of Science n° 296, M.I.T. Media Lab, Perceptual Computing Group, août 1994.

[Ioka 92] M. Ioka, M. Kurokawa. – A method for retrieving sequences of images on the basis of motion analysis. – *Conference on Storage and Retrieval for Images and Video Databases*, vol. SPIE 1662, pp. 35–46, San Jose, février 1992.

[Iyengar 98] Giridharan Iyengar, Andrew Lippman. – Models for automatic classification of video sequences. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 3312, pp. 216–227, San Jose, janvier 1998.

[Jain 88] Anil K. Jain, Richard C. Dubes. – *Algorithms for Clustering Data.* – Prentice Hall, 1988.

[Jain 99] Anil K. Jain, Aditya Vailaya, Xiong Wei. – Query by video clip. *Multimedia Systems*, 7(5):369–384, 1999.

[Jasinschi 91] R.S. Jasinschi. – Energy filters, motion uncertainty, and motion sensitive cells in the visual cortex: a mathematical analysis. *Biological Cybernetic*, 65:515–523, 1991.

[Jolion 98] Jean-Michel Jolion. – *Indexation d'images : nouvelle problématique ou vieux débats ?* – Rapport de recherche n° 9805, Laboratoire Reconnaissance de Formes et Vision, INSA Lyon, mai 1998.

[Jolion 00] Jean-Michel Jolion. – Feature similarity. *Principles of Visual Information Retrieval*, éd. par M. Lew eds. – Springer, 2000.

[Joly 96a] Philippe Joly. – *Consultation et analyse des documents en image animée numérique*. – Thèse de doctorat, 1996.

[Joly 96b] Philippe Joly, Hae-Kwang kim. – Efficient automatic analysis of camera work and microsegmentation of video using spatiotemporal images. *Signal Processing*, 8(4):295–307, 1996.

[Ju 96] Shanon X. Ju, Michael J. Black, Yaser Yacoob. – Cardboard people : A parameterized model of articulated image motion. – *International Conference on Automatic Face and Gesture Recognition*, pp. 38–44, Killington, octobre 1996.

[Ju 98] Shanon X. Ju, Michael J. Black, Scott Minneman, Don Kimber. – Summarization of video-taped presentations: Automatic analysis of motion and gesture. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):686–696, 1998.

[Karlsen 00] Robert E. Karlsen, David J. Gorsich, Grant R. Gerhart. – Target classification via support vector machines. *Optical Engineering*, 39(3):704–711, 2000.

[Kender 98] John R. Kender, Boon-Lock Yeo. – Video scene segmentation via continuous video coherence. – *International Conference on Computer Vision and Pattern Recognition*, pp. 367–373, Santa Barbara, juin 1998.

[Kim 96] Hae-Kwang Kim. – Efficient automatic text location method and content-based indexing and structuring of video database. *Journal of Visual Communication and Image Representation*, 7(4):336–344, 1996.

[Kobla 97] V. Kobla, D. Doermann, D. Faloutsos. – Video trails : Representing and visualizing structure in video sequences. – *ACM Multimedia Conference*, pp. 335–346, Seattle, novembre 1997.

[Kobla 99] Vikrant Kobla, Daniel DeMenthon, David Doermann. – Detection of slow-motion replay sequences for identifying sports videos. – *IEEE Workshop on Multimedia Signal Processing*, pp. 135–140, Copenhague, septembre 1999.

[Konrad 92] Janusz Konrad, Eric Dubois. – Bayesian estimation of motion vector fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):910–927, 1992.

[Kurita 97] Takio Kurita, Yasuo Kobayashi, Taketoshi Mishima. – Higher order local autocorrelation features of parcor images for gesture recognition. – *International Conference on Image Processing*, pp. 722–725, Washington, octobre 1997.

[Lacoste 02] Corinne Lacoste, Ronan Fablet, Patrick Bouthemy, Y.-F. Yao. – Création de résumés de vidéos par une approche statistique. – *Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, Angers, janvier 2002.

[LeRoux 02] Estelle LeRoux, Karine Lespinasse. – *Les possibilités de structurer des termes et des informations pour décrire des documents audiovisuels.* – Rapport de recherche, Institut National de l'Audiovisuel, janvier 2002.

[Lienhart 97] Rainer Lienhart, Silvia Pfeiffer, Wolfgang Effelsberg. – Video abstracting. *Communications of the ACM*, 40(12):55–62, 1997.

[Lienhart 98] R. Lienhart, W. Effelsberg, R. Jain. – Towards a Visual Grep : A systematic analysis of various methods to compare video sequences. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 3312, pp. 271–282, Ishwar, janvier 1998.

[Lienhart 99] Rainer Lienhart, Silvia Pfeiffer, Wolfgang Effelsberg. – Scene determination based on video and audio features. – *Conference on Multimedia Computing and Systems*, pp. 685–690, Florence, juin 1999.

[Llach 99] Joan Llach, Philippe Salembier. – Analysis of video sequences: table of contents and index creation. – *International Workshop on Very Low Bitrate Video*, Kyoto, octobre 1999.

[Low 96] Chien Yon Low, Qi Tian, Hongjiang Zhang. – An automatic news video parsing, indexing and browsing system. – *ACM Multimedia*, pp. 425–426, 1996.

[Manjunath 96] B.S. Manjunath, W.Y. Ma. – Texture features for browsing and retrieval of image data. *Pattern Analysis and Machine Intelligence*, 18(8):836–842, 1996.

[Marques 97] Ferran Marques, Cristina Molina. – An object tracking technique for content-based functionalities. – *Visual Communication and Image Processing Conference*, vol. SPIE 3024, pp. 190–198, San Jose, février 1997.

[Marsicoi 97] M. De Marsicoi, L. Cinque, S. Levialdi. – Indexing pictorial documents by their content : a survey of current techniques. *Image and Vision Computing*, 15(2):119–141, 1997.

[McCane 98] B. McCane, B. Galvin, K. Novins. – On the evaluation of optical flow algorithms. – *International Conference on Control, Automation, Robotics and Vision*, Singapore, décembre 1998.

[Merialdo 98] Bernard Merialdo. – Utilisation de corpus pour l'indexation vido de journaux tlviss. – *Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, pp. 37–46, Clermont Ferrand, janvier 1998.

[Merialdo 99] Bernard Merialdo, Kyung Tak Lee, Dario Luparello, Jérémie Roudaire. – Automatic construction of personalized TV news programs. – *ACM Multimedia Conference*, pp. 323–331, Orlando, novembre 1999.

[Merlino 97] Andrew Merlino, Daryl Morey, Mark Maybury. – Broadcast news navigation using story segmentation. – *ACM Multimedia Conference*, pp. 381–391, Seattle, novembre 1997.

[Metz 68] Christian Metz. – *Essais sur la signification au cinéma.* – Klincksieck, vol. 1, Paris, 1968.

[Mitiche 96] Amar Mitiche, Patrick Bouthemy. – Computation and analysis of image motion: a synopsis of current problems and methods. *International Journal of Computer Vision*, 19(1):29–55, 1996.

[Miyamori 98] Hisashi Miyamori, Tomio Echigo, Shun-Ichi Iisaku. – Proposal of query by short-time action descriptions in a scene. – *IAPR Workshop on Machine Vision Applications*, pp. 111–114, Makuhari, novembre 1998.

[Moeslund 01] Thomas B. Moeslund, Erik Granum. – A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231–268, 2001.

[Mohan 98] Rakesh Mohan. – Video sequence matching. – *International Conference on Acoustic, Speech and Signal Processing*, pp. 3697–3700, Seattle, mai 1998.

[Nagasaka 91] Akio Nagasaka, Yuzuru Tanaka. – Automatic video indexing and full-video search for object appearances. – *IFIP Conference on Visual Database Systems*, vol. 2, pp. 113–126, Budapest, octobre 1991.

[Nagel 87] Hans-Hellmut Nagel. – On the estimation of optical flow: relations between different approaches and some new results. *Artificial Intelligence*, 33(3):299–324, 1987.

[Nagel 88] Hans-Hellmut Nagel. – From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59–74, 1988.

[Nagel 01] Hans-Hellmut Nagel. – Image sequence evaluation: 30 years and still going strong. – *International Conference on Pattern Recognition*, vol. 1, pp. 149–158, Barcelone, septembre 2001.

[Nakano 00] Hiroki Nakano. – Classifying the motions of human body by using Gabor wavelets. – *International Conference on Pattern Recognition*, pp. 691–694, Barcelone, septembre 2000.

[Nam 97] Jeho Nam, Ahmed H. Tewfik. – Combined audio and visual streams analysis for video sequence segmentation. – *International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. 2665–2668, Munich, avril 1997.

[Naphade 98] M.R. Naphade, T. Kristjansson, B. Frey, T.S. Huang. – Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. – *International Conference on Image Processing*, vol. 3, pp. 536–540, Chicago, octobre 1998.

[Naphade 00a] M. Naphade, T. Huang. – Semantic video indexing using a probabilistic framework. – *International Conference on Pattern Recognition*, vol. 3, pp. 83–88, Barcelone, septembre 2000.

[Naphade 00b] Milind R. Naphade, Minerva M. Yeung, Boon-Lock Yeo. – A novel scheme for fast and efficient video sequence matching using compact signatures. – *Conference on Storage and Retrieval for Media Databases*, vol. SPIE 3972, pp. 564–572, San Jose, janvier 2000.

[Nastar 98a] Chahab Nastar, Matthias Mitschke, Christophe Meilhac. – Efficient query refinement for image retrieval. – *International Conference on Computer Vision and Pattern Recognition*, pp. 547–552, Santa Barbara, juin 1998.

[Nastar 98b] Chahab Nastar, Matthias Mitschke, Christophe Meilhac, Nadia Boujemaa. – Surfimage: a flexible content-based image retrieval system. – *ACM Multimedia Conference*, pp. 339–344, Bristol, septembre 1998.

[Nelson 92] Randal C. Nelson, Ramprasad Polana. – Qualitative recognition of motion using temporal texture. – *DARPA Image Understanding Workshop*, pp. 555–559, San Diego, janvier 1992.

[Neumann 83] Bernd Neumann, Hans-Joachim Novak. – Event models for recognition and natural language description of events in real-world image sequences. – *International Joint Conference on Artificial Intelligence*, pp. 724–726, Karlsruhe, août 1983.

[Ngo 01] C.W. Ngo, T.C. Pong, R.T. Chin. – Video partitioning through temporal slices analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2001.

[Niblack 93] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, G. Taubin. – The QBIC project: Querying images by content using color, texture, and shape. – *Storage and Retrieval for Image and Video Databases*, vol. SPIE 1908, pp. 173–181, San Jose, février 1993.

[Odobez 95] Jean-Marc Odobez, Patrick Bouthemy. – Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, 1995.

[Ohba 97] Kohtaro Ohba, Katsushi Ikeuchi. – Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9):1043–1048, 1997.

[Osuna 97a] Edgar Osuna, Robert Freund, Federico Girosi. – Training support vector machines : an application to face detection. – *International Conference on Computer Vision and Pattern Recognition*, pp. 130–136, Porto Rico, juin 1997.

[Osuna 97b] Edgar E. Osuna, Robert Freund, Federico Girosi. – *Support Vector Machines: Training and Applications.* – A.I. Memo n° 1602, MIT Artificial Intelligence Lab, mars 1997.

[Papageorgiou 97] Constantine Phaedon Papageorgiou. – *Object and Pattern Detection in Video Sequences.* – Master of science, Department of Electrical Engineering and Computer Science, mai 1997.

[Papageorgiou 99] Constantine Papageorgiou. – A pattern classification approach to dynamical object detection. – *International Conference on Computer Vision*, pp. 1223–1228, Corfu, septembre 1999.

[Pass 96] Greg Pass, Ramin Zabih, Justin Miller. – Comparing images using color coherence vectors. – *ACM Multimedia Conference*, pp. 65–73, Boston, novembre 1996.

[Pass 99] Greg Pass, Ramin Zabih. – Comparing images using joint histograms. *Journal of Multimedia Systems*, 7(3):234–240, 1999.

[Pentland 94] A. Pentland, R. W. Picard, S. Sclaroff. – Photobook : Content-based manipulation of image databases. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 2185, San Jose, février 1994.

[Pfeiffer 99] Silvia Pfeiffer. – The importance of perceptive adaptation of sound features in audio content processing. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 3656, pp. 328–337, San Jose, janvier 1999.

[Pittore 00] Massimiliano Pittore. – *A Trainable system for event detection and recognition.* – Thèse de doctorat, Università di Genova, 2000.

[Platt 98] J. Platt. – How to implement SVMs. *IEEE Intelligent Systems, Trends and controversies on support vector machines*, 13(4):18–28, 1998.

[Pontil 98a] Massimiliano Pontil, Alessandro Verri. – Properties of support vector machines. *Neural Computation*, 10:955–974, 1998.

[Pontil 98b] Massimiliano Pontil, Alessandro Verri. – Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.

[Prie 98] Yannick Prie, Alain Mille, Jean-Marie Pinon. – AI-STRATA : A user-centered model for content-based description and retrieval of audiovisual sequences. – *International Advanced Multimedia Content Processing Conference*, Osaka, novembre 1998.

[Qian 99] Richard Qian, Niels Haering, Ibrahim Sezan. – A computational approach to semantic event detection. – *International Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 200–206, Fort Collins, juin 1999.

[Rabiner 93] L. R. Rabiner, B-H. Juang. – *Fundamentals of speech recognition.* – Prentice-Hall, 1993.

[Rigoll 96] G. Rigoll, A. Kosmala, M. Schuster. – A new approach to video sequence recognition based on statistical methods. – *International Conference on Image Processing*, vol. 3, pp. 839–842, Lausanne, septembre 1996.

[Roobaert 99] D. Roobaert, M.M. Van Hulle. – View-based 3D object recognition with support vector machines. – *International Workshop on Neural Networks for Signal Processing*, pp. 77–84, Wisconsin, août 1999.

[Rowe 94] Lawrence A. Rowe, John S. Boreczky, Charles A. Eads. – Indexes for user access to large video databases. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 2185, pp. 150–161, San Jose, février 1994.

[Rubner 98] Yossi Rubner, Carlo Tomasi, Leonidas J. Guibas. – *The Earth Mover's Distance as a Metric for Image Retrieval.* – Technical Report n° STAN-CS-TN-98-86, Department of Computer Science, Stanford University, septembre 1998.

[Rui 98] Yong Rui, Thomas S. Huang, Sharad Mehrotra. – Exploring video structure beyond the shots. – *International Conferrence on Multimedia Computing and Systems*, pp. 237–240, Austin, juin 1998.

[Rui 99a] Yong Rui, Thomas S. Huang, Shih-Fu Chang. – Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, 1999.

[Rui 99b] Yong Rui, Thomas S. Huang, Sharad Mehrotra. – Constructing table-of-content for videos. *Multimedia Systems*, 7(5):359–368, 1999.

[Ruiloba 99] Rosa Ruiloba, Philippe Joly, Stéphane Marchand-Maillet, Georges Quénot. – Towards a standard protocol for the evaluation of video-to-shots segmentation algorithms. – *Content-Based Multimedia Indexing Workshop*, pp. 41–48, Toulouse, octobre 1999.

[Saarela 99] Janne Saarela, Bernard Merialdo. – Using content models to build audio-video summaries. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 3656, pp. 338–347, San Jose, janvier 1999.

[Sahouria 98] Emile Sahouria, Avideh Zakhor. – Content analysis of video using principal components. – *International Conference on Image Processing*, vol. 3, pp. 541–545, Chicago, octobre 1998.

[Salembier 01] Philippe Salembier. – An overview of the MPEG-7 standard and of future challenges for visual information analysis. – *Workshop on Image Analysis for Multimedia Services*, pp. 75–82, Tampere, mai 2001.

[Saraceno 97] Caterina Saraceno, Riccardo Leonardi. – Audio as a support to scene change detection and characterization of video sequences. – *International Conference on Acoustics, Speech, and Signal Processing*, pp. 2597–2600, Munich, mai 1997.

[Saraceno 98] Caterina Saraceno, Riccardo Leonardi. – Identification of story units in audio-visual sequences by joint audio and video processing. – *International Conference on Image Processing*, pp. 363–367, Chicago, octobre 1998.

[Scassellati 94] B. Scassellati, S. Alexopoulos, M. Flickner. – Retrieving images by 2D shape: a comparison of computation methods with human perceptual judgments. – *Storage and Retrieval for Image and Video Databases*, vol. SPIE 2185, pp. 2–14, San Jose, février 1994.

[Scheirer 97] Eric Scheirer, Malcolm Slaney. – Construction and evaluation of a robust multi-features speech/music discriminator. – *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1331–1334, Munich, avril 1997.

[Schmid 97] Cordelia Schmid, Roger Mohr. – Object recognition using local characterization and semi-local constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):530–534, 1997.

[Scholkopf 95] Bernhard Scholkopf, Chris Burges, Vladimir Vapnik. – Extracting support data for a given task. – *International Conference on Knowledge Discovery and Data Mining*, pp. 252–257, Montréal, août 1995.

[Scholkopf 98] Bernhard Scholkopf. – SVMs - a practical consequence of the learning theory. *IEEE Intelligent Systems, Trends and controversies on support vector machines*, 13(4):18–28, 1998.

[Shibata 92] Masahiro Shibata. – A temporal segmentation method for video sequence. – *Conference on Visual Communication and Image Processing*, vol. SPIE 1818, pp. 1194–1205, Boston, novembre 1992.

[Simoncelli 91] Eero P. Simoncelli, Edward H. Adelson. – *Computing optical flow distributions using spatio-temporal filters.* – Rapport technique n° 165, MIT Media Lab, Vision and Modelling, mars 1991.

[Slype 87] Georges Van Slype. – *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* – Editions d'organisation, Paris, 1987.

[Smith 92] Thomas G. Aguierre Smith, Glorianna Davenport. – The stratification system : A design environment for random access video. – *International Workshop on Network and Operating System Support for Digital Audio and Video*, San Diego, novembre 1992.

[Smith 96] John R. Smith, Shih-Fu Chang. – Tools and techniques for color image retrieval. – *Storage and Retrieval for Image and Video Databases*, vol. SPIE 2670, pp. 426–437, San Jose, mars 1996.

[Smith 97] Michael A. Smith, Takeo Kanade. – Video skimming and characterization through the combination of image and language understanding techniques. – *International Conference on Computer Vision and Pattern Recognition*, pp. 775–781, San Juan, juin 1997.

[Smola 96] Alexander Smola. – *Regression Estimation with Support Vector Learning Machines.* – Master of science, Technische Universität München, décembre 1996.

[Soffer 98] A. Soffer, H. Samet. – Using negative shape features for logo similarity matching. – *International Conference on Pattern Recognition*, pp. 571–573, Brisbane, août 1998.

[Spina 96] M.S. Spina, V.W. Zue. – Automatic transcription of general audio data: Preliminary analysis. – *International Conference on Spoken Language Processing*, vol. 2, pp. 594–597, Philadelphie, octobre 1996.

[Spinei 98a] Adrian Spinei. – *Estimation du mouvement par triades de filtres de Gabor. Application au mouvement d'objets transparents.* – Thèse de doctorat, Institut National Polytechnique de Grenoble, 1998.

[Spinei 98b] Adrian Spinei, Denis Pellerin, Jeanny Hérault. – Spatiotemporal energy-based method for velocity estimation. *Signal Processing*, 65(3):347–362, 1998.

[Spinei 00] Andrian Spinei, Denis Pellerin. – Fast hardware implementation of gabor filter based motion estimation. *Integrated Computer-Aided Engineering*, 7(1):67–77, 2000.

[Spinei 01] Adrian Spinei, Denis Pellerin. – Motion estimation of opaque or transparent objects using triads of gabor filters. *Signal Processing*, 81(4):845–853, 2001.

[Stricker 95] Markus Stricker, Markus Orengo. – Similarity of color images. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 2420, pp. 381–392, San Diego, février 1995.

[Stricker 96] Markus Stricker, Alexander Dimai. – Color indexing with weak spatial constraints. – *Conference on Storage and Retrieval for Image and Video Databases*, vol. SPIE 2570, pp. 29–40, San Jose, février 1996.

[Swain 91] Michael J. Swain, Dana H. Ballard. – Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[Swanberg 93] Deborah Swanberg, Chiao-Fe Shu, Ramesh Jain. – Knowledge guided parsing in video databases. – *Conference on Storage and Retrieval for Image and Video Database*, vol. 1908, pp. 13–24, San Jose, février 1993.

[Szummer 96] Martin Szummer, Rosalind W. Picard. – Temporal texture modeling. – *International Conference on Image Processing*, pp. 823–826, Lausanne, septembre 1996.

[Taki 98] Tsuyoshi Taki, Jun-Ichi Hasegawa, Teruo Fukumura. – Group motion features for teamwork evaluation and its application to soccer games. – *International Conference on Pattern Recognition*, vol. 1, Brisbane, août 1998.

[Tamura 78] Hideyuki Tamura, Shunji Mori, Takashi Yamawaki. – Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6):460–473, 1978.

[Taniguchi 97] Yukinobu Taniguchi, Akihito Akutsu, Yoshinobu Tonomura. – Panorama excerpts: Extracting and packing panoramas for video browsing. – *ACM Multimedia Conference*, pp. 427–435, Seattle, novembre 1997.

[Troncy 01] Raphaël Troncy. – *Études du Manuel d'Indexation commun à tous les documentalistes.* – Rapport de recherche, Institut National de l'Audiovisuel, avril 2001.

[Vapnik 95] Vladimir N. Vapnik. – *The Nature of Statistical Learning Theory.* – Springer-Verlag, New-York, 1995.

[Vasconcelos 97a] Nuno Vasconcelos, Andrew Lippman. – A Bayesian video modeling framework for shot segmentation and content characterization. – *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp. 59–67, San Juan, juin 1997.

[Vasconcelos 97b] Nuno Vasconcelos, Andrew Lippman. – Content-based pre-indexed video. – *International Conference on Image Processing*, Washington, octobre 1997.

[Vasconcelos 97c] Nuno Vasconcelos, Andrew Lippman. – Towards semantically meaningful feature spaces for the characterization of video content. – *International Conference Image Processing*, vol. 1, Washington, octobre 1997.

[Vasconcelos 98a] Nuno Vasconcelos, Andrew Lippman. – A Bayesian framework for semantic content characterization. – *International Conference on Computer Vision and Pattern Recognition*, pp. 566–571, Santa-Barbara, juin 1998.

[Vasconcelos 98b] Nuno Vasconcelos, Andrew Lippman. – Bayesian modeling of video editing and structure : Semantic features for video summarization and browsing. – *International Conference on Image Processing*, Chicago, octobre 1998.

[Vehel 00] Jacques Lévy Vehel, Bertrand Guilheneuf. – Texture based video indexing. – *International Conference on Signal and Image Processing*, pp. 919–928, Las Vegas, novembre 2000.

[Veneau 00] Emmanuel Veneau, Rémi Ronfard, Patrick Bouthemy. – From video shot clustering to sequence segmentation. – *International Conference on Pattern Recognition*, pp. 254–257, Barcelone, septembre 2000.

[Veneau 01] Emmanuel Veneau, Julien Pépy, Laurent Vinet. – Toward a semantic indexing based on images color features similarity ? – *International Conference on Image and Signal Processing*, Agadir, mai 2001.

[Vertan 00] C. Vertan, N. Boujemaa. – Upgrading color distributions for image retrieval: can we do better ? – *International Conference on Visual Information Systems*, vol. LNCS 1929, pp. 178–188, Lyon, novembre 2000.

[Vinod 98] V.V. Vinod. – Activity based video shot retrieval and ranking. – *International Conference on Pattern Recognition*, pp. 682–684, Brisbane, août 1998.

[Wildes 98] Richard P. Wildes. – A measure of motion salience for surveillance applications. – *International Conference on Image Processing*, Chicago, octobre 1998.

[Wildes 00] Richard P. Wildes, James R. Bergen. – Qualitative spatiotemporal analysis using an oriented energy representation. – *European Conference on Computer Vision*, vol. 2, pp. 768–784, Dublin, juin 2000.

[Wilson 95] Andrew Wilson, Aaron Bobick. – Learning visual behavior for gesture analysis. – *IEEE Symposium on Computer Vision*, pp. 229–234, Coral Gables, novembre 1995.

[Wilson 96] Andrew D. Wilson, Aaron A. Bobick, Justine Cassell. – Recovering the temporal structure of natural gesture. – *International Conference on Automatic Face and Gesture Recognition*, Killington, octobre 1996.

[Wilson 98] Andrew D. Wilson, Aaron F. Bobick. – Recognition and interpretation of parametric gesture. – *International Conference on Computer Vision*, Bombay, janvier 1998.

[Yacoob 96] Yaser Yacoob, Larry S. Davis. – Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[Yacoob 99] Yaser Yacoob, Michael J. Black. – Parameterized modeling and recognition of activities. *Computer Vision and image Understanding*, 73(2):232–247, 1999.

[Yang 01] Ming-Hsuan Yang, David Kriegman, Narendra Ahuja. – Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[Yeo 95] Boon-Lock Yeo, Bede Liu. – A unified approach to temporal segmentation of motion JPEG and MPEG compressed video. – *International Conference on Multimedia Computing and Systems*, pp. 2–13, Washington, mai 1995.

[Yeung 95] Minerva M. Yeung, Boon-Lock Yeo, Wayne Wolf, Bede Liu. – Video browsing using clustering and scene transitions on compressed sequences. – *Conference on Multimedia Computing and Networking*, vol. SPIE 2417, San Jose, février 1995.

[Yeung 96] Minerva M. Yeung, Boon-Lock Yeo, Bede Liu. – Extracting story units from long programs for video browsing and navigation. – *International Conferrence on Multimedia Computing and Systems*, pp. 296–305, Tokyo, juin 1996.

[Yeung 97] Minerva M. Yeung, Boon-Lock Yeo. – Video content characterization and compaction for digital library applications. – *Conference on Storage and Retrieval for Image and Video Database*, vol. SPIE 3022, San Jose, février 1997.

[Yeung 98] Minerva M. Yeung, Boon-Lock Yeo. – Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, 71(1):94–109, 1998.

[Yoshitaka 97] Atsuo Yoshitaka, Takakazu Ishii, Masahito Hirakawa, Tadao Ichikawa. – Content-based retrieval of video data by the grammar of film. – *Symposium on Visual Languages*, pp. 314–321, Darmstadt, septembre 1997.

[Yow 95] Dennis Yow, Boon-Lock Yeo, Minerva M. Yeung, Bede Liu. – Analysis and presentation of soccer highlights from digital video. – *Asian Conference on Computer Vision*, vol. 2, pp. 499–503, Singapore, décembre 1995.

[Zabih 96] Ramin Zabih, Justin Miller, Kevin Mai. – Video browsing using edges and motion. – *International Conference on Computer Vision and Pattern Recognition*, pp. 439–306, San Francisco, juin 1996.

[Zabih 99] Ramin Zabih, Justin Miller, Kevin Mai. – A feature-based algorithm for detecting and classifying scene breaks. *Multimedia Systems*, 7(2):119–128, 1999.

[Zhang 94] Hong Jiang Zhang, Yihong Gong, Stephen W. Smoliar, Shang Yeo Tan. – Automatic parsing of news video. – *International Conference on Multimedia Computing and Systems*, pp. 45–54, 1994.

[Zhang 95] Hong Jiang Zhang, C. Y. Low, S. W. Smoliar, Di Zhong. – Video parsing, retrieval and browsing: An integrated and content-based solution. – *ACM Multimedia Conference*, San Francisco, novembre 1995.

[Zhong 96] Di Zhong, HongJiang Zhang, Shih-Fu Chang. – Clustering methods for video browsing and annotation. – *Conference on Storage and Retrieval for Image and Video Database*, vol. SPIE 2670, pp. 239–246, San Jose, février 1996.

[Zupan 82] Jure Zupan. – *Clustering of Large Data Sets.* – John Wiley & Sons Ltd., 1982.

**Résumé**

Cette étude s'inscrit dans le cadre général de l'indexation automatique des documents audiovisuels. Deux axes de recherche y sont privilégiés : la segmentation temporelle d'un document vidéo et la caractérisation du contenu dynamique de celui-ci.

Dans le cadre de la segmentation temporelle, nous nous sommes intéressés à la segmentation en séquences ou macro-segmentation. Pour mener à bien cette tâche, nous avons proposé une description des plans vidéos par des descripteurs de nature composite, et une classification de ceux-ci utilisant une représentation hiérarchique de l'information visuelle contrainte par le temps. La méthode de détection des macro-segments est fondée sur le calcul et le seuillage de la distance cophénétique associée à la construction de la hiérarchie.

Afin de caractériser le contenu des documents, nous nous sommes placés dans un cadre d'indexation général que nous avons appliqué à un domaine particulier - le mouvement dans les documents de sport. Plus précisément, nous avons utilisé une méthode de classification par apprentissage fondée sur les machines à vecteurs de support (SVM) et étudié des descripteurs extraits de séquences courtes. Pour l'extraction de ceux-ci, nous avons mis en uvre deux familles de méthodes, la première liée aux images de l'historique du mouvement (MHI), la seconde issue de l'utilisation de filtres de Gabor spatio-temporels tridimensionnels.

Dans le contexte de nos travaux, nous avons porté une attention particulière à la question de l'évaluation des méthodes au regard des usages et des pratiques des documentalistes de l'audiovisuel. Nous avons défini une méthodologie d'évaluation qualitative et quantitative des performances de l'outil de macro-segmentation, et nous avons construit un ensemble de bases de séquences de tests pour l'analyse des résultats des différentes méthodes de caractérisation du contenu dynamique des séquences.

**Mots clefs :** indexation vidéo, segmentation temporelle, mouvement, classification hiérarchique, classification par apprentissage, évaluation des algorithmes, filtres de Gabor, reconnaissance d'activités.

**Abstract**

This thesis lies within the general context of automatic indexing of audiovisual documents. Two subjects of research are concerned : temporal segmentation of a video documents and motion content characterization of the latter.

In the scope of temporal segmentation, we deal with high-level segmentation in sequences. To achieve this task, we propose to describe video shots with various features, and to classify shots by using a hierachical time-constrained representation of visual information. The high-level segments detection method is based on both the claculus and the thresholding of the cophenetic distance linked to the hierarchy building.

In order to characterize documents content, we fall within a general indexing framework which we apply to a particular area - activity in sports videos. More precisely, we use a classification method by learning based on support vector machines (SVM). We also study features extraction for short sequences. In order to do so, we implement two families of methods : the first is linked to motion history images (MHI), the second is based on the use of tridimensional spatio-temporal banks of Gabor filters.

Within our study, we pay a careful attention to a particular issue, that of evaluating the process in comparison with usages and practices of audiovisual professionals. We define a methodology for qualitative and quantitative evaluation of the high-level segmentation tool performances. We build a set of video testing sequences databases for analyzing the results of the various motion content based classification of sequences methods.

**Keywords:** Video indexing, temporal segmentation, motion, hierarchical clustering, classification by learning, algorithms evaluation, Gabor filters, activity recognition.