UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE
ÉCOLE DOCTORALE EDITE
IRCAM – CENTRE POMPIDOU
COMMISSARIAT À L'ÉNERGIE ATOMIQUE

THÈSE DE DOCTORAT

spécialité
ACOUSTIQUE, TRAITEMENT DU SIGNAL ET INFORMATIQUE
APPLIQUÉS À LA MUSIQUE

présentée par
BERTRAND DELEZOIDE

pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE

Modèles d'indexation multimedia pour la description automatique de films de cinéma

soutenue le 24 Avril 2006
devant le jury composé de

Xavier RODET	Université Paris VI	Directeur de thèse
Christian FLUHR	CEA/LIC2M	Co-directeur de thèse
Françoise PRÉTEUX	INT d'Évry	Rapporteuse
François PACHET	Sony CSL-Paris	Rapporteur
Liming CHEN	École Centrale de Lyon	Examinateur
Patrick GALLINARI	Université Paris VI	Examinateur
Marcin DETYNIECKI	Université Paris VI	Examinateur

Fluctuat nec mergitur.

Résumé

Modèles d'indexation multimedia pour la description automatique de films de cinéma

Depuis une quinzaine d'année, l'analyse automatique des films de cinéma se focalise sur la description de l' « histoire » transmise par le support cinématographique. La recherche et l'estimation de cette information constituent un travail d'indexation de données. Il existe deux types d'information attachés à un document : la structure et le contenu. L'extraction de la structure d'un film met en évidence l'organisation temporelle de la narration. La détermination du contenu vise à décrire les caractéristiques physiques et sémantiques des éléments de la structure. Les premières méthodes d'indexation existantes, fondées sur l'analyse spatiale et temporelle des caractéristiques numériques du signal image et son fournissent des résultats satisfaisants pour des modèles d'index assez simples. Mais généralement leurs performances en termes de temps de calcul et de résultats se dégradent lorsque la structure et le contenu deviennent plus complexes.

Notre hypothèse suppose que ces méthodes ne prennent suffisamment en compte les relations particulières de dépendance pouvant exister entre les données de l'index (numériques, textuelles, sonores et visuelles). Nous proposons d'y remédier par la création d'un modèle de structure et de contenu spécifique au film, fondé sur la fusion de l'information fournie par ces différentes descriptions.

Dans ce travail, nous justifions cette hypothèse par l'étude des informations pertinentes présentes dans le contenu des films et de leur exploitation par les méthodes existantes de classification et de segmentation. Inspiré par les modèles d'analyse hiérarchique par réseau bayésien (RB) et support vector machine (SVM), nous construisons ensuite un modèle probabiliste de fusion adapté aux différentes relations entre descripteurs. Afin de valider notre modèle, nous donnons quelques exemples expérimentaux de détermination automatique de contenu audio, image et multimedia nécessaires à la description d'un film. Enfin, nous utilisons la coopération des descriptions ainsi extraites pour la segmentation de la structure hiérarchique des films.

Mots-clefs : indexation automatique multimedia, segmentation de la structure, extraction automatique du contenu, fusion des données, films de cinéma

Abstract

Models of multimedia indexing for the automatic description of motion-picture films.

For about fifteen years, the study of motion-picture films principally focuses on the description of ‘stories’ conveyed by the medium of cinema. The search for such information and its retrieval constitute the task of data indexing. It is generally assumed that two types of information can proceed from any given document : structure and content. Retrieving the structure of a film reveals the temporal organization of its narrative. Determining the content aims at describing the physical and semantic characteristics of the elements of a film’s structure. The existing methods, based on a spatial and temporal analysis of the digital characteristics of the sound and image signals, give good results for the simpler index models. However, their efficiency, in terms of computing time and accuracy of the results, deteriorates as the structure and content become more complex.

Our hypothesis supposes that such methods do not take fully into account the interdependent relations that may exist between the indexed data (numerical, textual, sonorous and visual). We propose to remedy this by creating a model of structure and content entirely specific to film, based on the merging of data obtained through various description processes.

In this work, we justify our premise by studying the relevant information found in the content of films, and the manner in which it is exploited by the existing classification and segmentation methods. Next, we build a probabilistic model for merging the contents of the film’s based on their specific relations that is inspired by Bayesian network models for hierarchical analysis and support vector machine. We then give a number of experimental examples of the automatic analyses of audio, image, and multimedia content necessary to the description of a film. Finally, we process the descriptors thus retrieved in order to perform segmentation of hierarchical structures in films.

Keywords : automatic multimedia indexing, structural segmentation, extraction of content, data fusion, motion-picture films.

Remerciements

Je tiens avant tout à remercier Xavier Rodet pour m'avoir proposé de travailler sur ce sujet passionnant, à la fois scientifique et proche de la création artistique cinématographique et musicale. Par son expérience exceptionnelle de la recherche appliquée en audio, il a su à la fois me conseiller et me laisser libre de creuser mes propres directions, ce dont je lui suis hautement reconnaissant.

Je remercie ensuite Christian Flhur pour son soutien scientifique et financier qui m'ont permis de mener à bien ce travail. Par ses connaissances de la recherche multimedia et du secteur de l'industrie, il m'a aidé à appréhender en profondeur les problématiques de mon sujet de thèse et à valoriser mon expérience auprès du monde scientifique et industriel. Je remercie Patrick Hede pour son encadrement au sein du laboratoire LIST/LIC2M du CEA. Ses qualités de communication, sa bonne humeur et son ouverture d'esprit furent des atouts appréciables pour mener à bien mes recherches.

Je remercie les fondateurs, enseignants et gestionnaires du DEA ATIAM. L'existence du DEA, la qualité de la formation transmise et le lien créé avec le monde artistique ont été décisifs dans mon orientation. Merci de même à tous les enseignants de lycée, classes préparatoires et Ecole Normale Supérieure qui ont éveillé ma curiosité scientifique bien avant.

Merci beaucoup à tous les collègues de l'IRCAM et du CEA qui m'ont éclairé par des discussions scientifiques de qualité et m'ont appris à programmer correctement, écrire un article ou présenter des slides. Merci en particulier à mes voisins de bureaux, aux membres de l'équipe Analyse-Synthèse de l'IRCAM et aux membres du LIC2M du CEA/FAR.

Merci également aux membres du jury de l'intérêt qu'ils ont bien voulu porter à mon travail.

Mes remerciements vont aussi à tous ceux qui m'ont permis d'effectuer ce travail dans de bonnes conditions et d'avoir plus de temps pour la recherche proprement dit. Merci en particulier au personnel administratif de l'IRCAM, du CEA et de l'école doctorale EDITE.

Merci au Ministère de l'Education Nationale et de la Recherche, qui a financé la plus grande partie de mes études supérieures, et au CEA, qui m'a permis par son financement de poursuivre mon travail de thèse au sein de plusieurs laboratoires de renommée internationale. Merci aussi aux collègues enseignants du laboratoire de bio-informatique de Jussieu avec qui j'ai collaboré dans le cadre de travaux dirigés. Merci à tous les élèves qui ont assisté à ces TD pour leur curiosité et leur enthousiasme.

Merci aux artistes d'horizons divers qui m'ont donné de bonnes musiques, de bonnes peintures et de bons films quand j'en avais besoin. Mention spéciale à tous ceux qui ont favorisé l'accès à ces oeuvres à des prix abordables pour ma maigre bourse de thésard, au Centre Pompidou, à l'IRCAM, aux développeurs des plates-formes d'échange de fichiers, à mon vidéoclub.

J'exprime toute ma gratitude pour leur soutien et leur bonne humeur à ceux que j'ai côtoyé durant tout ce temps en dehors du boulot : mes parents, mes grands parents, mes frères, mon amour, et tous mes amis d'ici et d'ailleurs.

Table des matières

Liste des notations

Objets usuels

Nous adoptons les notations générales suivantes.

Les vecteurs seront notés en caractères majuscule. Les matrices seront notés en caractères majuscule gras. Les scalaires seront notés en caractères standard. Les parenthèses désigneront des suites, les bornes des indices n'étant pas précisées lorsqu'elles sont définies auparavant. Les accolades désigneront des ensembles.

Fonctions usuelles

$\mathbf{a}^t$	Transposée du vecteur ou de la matrice $\mathbf{a}$
$\exp(\mathbf{a})$	Exponentielle point-à-point du vecteur $\mathbf{a}$
$\log(\mathbf{a})$	Logarithme point-à-point du vecteur $\mathbf{a}$
$P(a)$	Probabilité ou densité de probabilité marginale de la variable a
$P(a\|b)$	Probabilité ou densité de probabilité conditionnelle de la variable a sachant l'événement b
$P(a, b)$	Probabilité ou densité de probabilité jointe des variables a et b

Indices

i	Objet multimedia (de 1 à n)
j	Description d'un objet (de 1 à m)
k	Classe d'un concept (de 1 à p)
t	Trame temporelle (de 1 à T)

Variables utilisées et paragraphe de définition

Π, O_i, Q_i, Δ, D_j, C_j, $\mathcal{M}$, T, c_k, X, Ψ	§2.2
Υ	§2.3
$\mathbf{U}$	§2.3.2
λ_j	§2.3.3
S_j, s_j^k	§2.4.1
P^k, P_j^k	§2.4.2
B_j, M_j, m_k, H_j, h_k	§4.1.1

Abbréviations et paragraphe de définition

MPEG-7	Standard de normalisation des descriptions	§1.1.3
MFCC	Mel frequency cepstral coefficients du signal audio	§2.1.4
SPF	Flux spectral du signal audio	§2.1.4
4ME	Modulation à 4Hz du signal audio	§2.1.4
TLEP	Histogramme de texture et de couleur d'une image	§2.1.5
GMM	Modèle de classification par mélange de gaussiennes	§2.2.2
SVM	Machine à support de vecteur	§2.2.3
OPC	Modèle de classification SVM, "un par classe"	§2.2.3
PPV	Modèle de classification des plus proches voisins	§2.2.3
ACP	Analyse en composantes principales	§2.3.2
ICA	Analyse en composantes indépendantes	§2.3.2
LDA	Analyse discriminante linéaire	§2.3.2
MI	Information mutuelle	§2.3.3
RB	Réseaux bayésiens	§2.5.2
MMC	Modèles de Markov cachés	§2.6.4

Liste des figures

Liste des tableaux

Introduction

De l'utilité supposée des modèles d'indexation

Tout a commencé au milieu des années 80 avec l'arrivée sur terre d'un cyborg humanoïde envoyé du futur pour une macabre mission, tuer Sarah Connor et anéantir tout espoir de survie de l'espèce humaine. Bien que construit de métal, ce robot possède une intelligence artificielle hors du commun et montre des capacités perceptives étonnantes. Son aptitude à interpréter et à comprendre tout ce qui l'entoure lui permet d'agir en fonction des situations rencontrées. Pourvu à l'origine d'un dictionnaire de connaissance fixe, il est doué d'une faculté d'apprentissage qui lui permet de réagir à de nouveaux concepts lorsque leur sens lui ont été explicités. Ces fonctions font de lui un adversaire redoutable auquel sera confronté le personnage principal du film, le fils de Sarah Connor, lui aussi venu du futur, pour la sauver. Après moult dialogues, poursuites, explosions, etc., l'intrigue se dénoue dans la scène finale par la mort du Terminator et affiche ainsi la victoire de l'homme sur la *machine*. Mais, dès lors, de nombreux scientifiques tentèrent de reproduire ces fonctions à l'aide des calculateurs disponibles à cette époque. Il se sont vite rendu compte de l'immense et complexe tâche qui les attendaient et de la quantité de fonds nécessaires à l'aboutissement de leur desseins. C'est pourquoi, afin d'attirer des investisseurs, ils durent trouver des projets intermédiaires et lucratifs.

Tout ceci n'est, bien sûr, qu'une fiction, mais elle illustre bien les problématiques liées à l'avènement du numérique et du multimedia ces vingt dernières années. La croissance exponentielle de données de tous ordres, qu'elles soient textuelles, en images, vidéo, etc. a résulté en la création de bases de données gigantesques dans lesquelles il est devenu impossible de rechercher de manière exhaustive et manuelle une information donnée. Devant ces difficultés de recherche et de réutilisation des informations, il est devenu nécessaire de reproduire la capacité du cyborg à organiser les données perceptives, et surtout l'information qu'elles contiennent, de façon à pouvoir y accéder rapidement et directement, dans un but de réutilisation pour d'autres applications. Il s'agit donc ici d'un problème d'indexation de données multimedia.

Dans ce travail, nous considérons plus particulièrement le cas des films de cinéma, c'est à dire le mélange temporel d'images et de sons dans un but narratif. L'étude de ces documents nécessite la mise en œuvre de la plupart des caractéristiques de la perception : la vue, l'audition, mais aussi la compréhension de la parole et des concepts présents dans tous les médias.

Nous étudions ces films sous deux angles applicatifs différents et liés : la classification et la segmentation. La classification vise à décrire un document par un ensemble de concepts utiles pour la compréhension de l'action (présence d'un objet ou d'un personnage, localisation, compréhension de la parole, etc.) à partir de données numériques simples issus des capteurs perceptifs. La segmentation cherche à extraire la structure spatio-temporelle des signaux perçus par un spectateur correspondant aux éléments sémantiques et narratifs de l'histoire véhiculée par le film.

Les premières méthodes pour résoudre ces questions sont apparues il y a une quinzaine d'années. Elles exploitent les informations colorimétriques de l'image et spectro-temporelles du son présentes dans

les films, à l'aide de modèles statistiques ou paramétriques. La plupart de ces méthodes utilisent des modèles numériques dévoués à une seule modalité (en général visuelle), les autres étant traités comme des illustrations ou des éléments complémentaires. Leurs résultats sont globalement satisfaisants sur des enregistrements simples générés par le montage synchrone et réaliste d'équipements de surveillance ou de caméras fixes (reproduction de conférence), mais se dégradent sur des enregistrements plus complexes, comme c'est le cas dans les films. L'information disponible devient souvent insuffisante pour distinguer les cas ambigus ou bruités. Les dernières recherches dans ce domaine ont prouvé qu'une indexation efficace nécessite une approche multimodale, dans laquelle la collaboration de différents media est utilisée.

Le but essentiel de ce travail est de montrer que l'utilisation de modèles d'intégration des médias spécifiques peut aider à classer et à segmenter des films de cinéma dont l'analyse est habituellement considérée comme difficile. Nous étudions, en particulier pour la classification, une famille de modèles probabilistes appelés modèles génératifs du contenu, dont les paramètres sont appris sur des bases de données de films de cinéma.

Un autre but important est de montrer que l'utilisation de données textuelles, de type concepts, pour la description des éléments multimédia facilite les taches d'indexation du contenu comme de la structure. Nous proposons en particulier une méthode adaptée à la fusion des concepts et des descripteurs numériques habituellement employés pour ces applications.

Plan du document

A la suite de cette introduction, le document est découpé en huit chapitres, une conclusion et une annexe qui se présentent comme suit.

Nous commençons dans le chapitre 1 par définir les tâches d'indexation de la structure et du contenu des documents multimédia. Nous discutons des problématiques liées à la fusion des médias pour ce type d'application.

Le chapitre 2 présente les méthodes existantes du traitement du signal permettant l'indexation multimédia. Un "état de l'art" montre comment estimer les descripteurs de l'information issus du signal sonore et visuel des films et comment segmenter ces signaux pour en extraire la structure. L'étude de ces techniques et de leurs limitations nous amène à sélectionner un certain nombre d'entre elles.

Nous proposons ensuite dans le chapitre 3 un cadre assez général pour construire des modèles d'indexation représentant le contenu et la structure temporelle des films. Nous expliquons comment l'étude de la description du son et de l'image permet de segmenter les films. Nous mettons ainsi en lumière des liens importants entre classification et segmentation.

Le chapitre 4 traite un modèle probabiliste de classification multimédia basé sur une analyse des relations de dépendances entre les descripteurs du signal. Nous construisons un modèle à plusieurs niveaux sémantiques et nous décrivons les algorithmes de classification et d'apprentissage associés.

Le chapitre 5 montre quelques exemples de classification par des descripteurs numériques du signal. Nous testons la performance des techniques de modification de l'espace de description pour la classification du contenu monomédia.

Le chapitre 6 explique comment combiner l'information provenant de descripteurs numériques et textuels pour l'indexation du contenu monomédia. Nous évaluons la performance de classification d'un système de détermination du "lieu" reproduit par une photographie.

Le chapitre 7 étend ces résultats aux cas multimédia. Nous proposons plusieurs techniques de fusion des descriptions audio et visuelles d'un plan de film. Nous donnons les performances de localisation de plans pour ces modèles et nous comparons les résultats à ceux des méthodes de classification monomédia correspondantes.

Le chapitre 8 décrit une application du modèle de contenu à la segmentation hiérarchique de la structure des films de cinéma. Nous donnons ensuite les performances de segmentation de l'algorithme développé pour quelques exemples de films.

Nous concluons en proposant des pistes de recherche pour améliorer les modèles d'indexation construits et en construire d'autres semblables.

L'annexe A présente l'algorithme de classification "Suport Vector Machine" utilisé pour estimer les descripteurs textuels et décrit l'algorithme de segmentation fondé sur les "Modèles de Markov Cachés" employé pour segmenter le signal des films.

Chapitre 1

Présentation des tâches considérées

Le travail de thèse présenté dans cet exposé est le résultat de la collaboration de deux centres de recherche : l'Institut de Recherche et Coordination Acoustique/Musique (IRCAM) à Paris et le Commissariat à l'Energie Atomique (CEA) à Fontenay-Aux-Roses. Il concerne l'indexation, la classification et la segmentation de documents multimedia et plus particulièrement de films de cinéma.

Ce premier chapitre est consacré à la définition des problèmes étudiés. Dans le paragraphe 1.1 nous rappelons le cadre de l'indexation multimedia. Nous décrivons ensuite dans les paragraphes 1.2 et 1.3 les tâches d'indexation du contenu et de la structure des films.

1.1 Cadre de l'indexation multimedia

1.1.1 Multimedia

L'évolution des techniques a peu à peu conduit à ce que textes et images, puis sons enregistrés et enfin images animées soient stockés et reproduits sur des supports différents : papier, piste magnétique, pellicule photographique. La **numérisation** de tous ces éléments sous forme d'informations stockables et traitables par ordinateur a permis leur intégration sur un même fichier, dans une mémoire d'ordinateur ou sur des périphériques spécifiques comme les lecteurs de Cd-Rom.

L'accroissement des performances des ordinateurs permet non seulement le stockage, mais aussi le **traitement** des images et des sons, ce qui renouvelle les possibilités d'applications variées. Le développement des réseaux de télécommunications et de télévision par câble autorise même le traitement des images et des sons à distance, ce qui démultiplie ces possibilités. Le traitement informatique des images et des sons permet de réaliser par exemple des encyclopédies qui associent au texte des images fixes ou animées, des sons proposant l'enregistrement du cri d'un animal, des extraits musicaux d'un compositeur, un discours d'homme politique. L'enseignement utilise pour ses didacticiels le multimedia interactif : l'élève peut manipuler à l'écran les représentations des objets d'une expérience de chimie, voir le résultat et éventuellement l'entendre. Les mêmes procédés sont employés pour des jeux vidéo, pour des maquettes virtuelles d'appartements proposés au choix des locataires, pour des modèles virtuels d'avions ou d'automobiles. On entre ici dans le domaine de la réalité virtuelle.

Un fichier **multimedia** fait appel quasi simultanément à trois types de perception (vision d'image, audition et lecture). Ce point constitue à la fois une originalité et une limite du multimedia. Les œuvres encyclopédiques ou muséographiques, en particulier, trouvent là un support fort intéressant compte tenu de ces modalités spécifiques de présentation, de recherche et de navigation. La souplesse de la numérisation informatique permet en effet de préétablir des parcours à l'intérieur des informations, parcours que l'utilisateur pourra emprunter au gré de ses recherches. L'information textuelle, sonore et visuelle

y est accessible de façon interactive par le biais de commandes qui permettent la navigation dans des ensembles complexes d'informations. On peut ainsi, par exemple, faire défiler les images représentant successivement plusieurs pièces d'un musée, sélectionner un tableau, demander l'affichage d'un texte explicatif ou sa lecture par une voix préenregistrée, passer d'un mot à un autre lorsque l'hypertexte le permet.

Les techniques multimedia doivent leur richesse à la superposition sur un même sujet de l'image, du texte et du son, dans des contextes d'utilisation bien déterminés. En ce sens, les techniques multimedia s'ajoutent à la gamme des moyens de communication déjà disponibles.

1.1.2 Traitement du multimedia

Les traitements numériques des fichiers multimedia par ordinateur sont très nombreux et variés. La première utilisation des données est la diffusion : visualisation d'une photo, écoute d'une musique, projection d'un film, enseignement assisté par ordinateur ou jeux. La deuxième est la modification (et création) de données : montage et modification d'images, de sons, ou de vidéos.

La **diffusion** des données nécessite le passage du support original : bande magnétique… vers un support de type fichier, lisible par un ordinateur : c'est la numérisation. Les données stockés sous cette forme sont décodées par un logiciel afin de permettre la diffusion du media par le périphérique approprié : moniteur, enceintes. La numérisation des images au moyen de scanners ou de caméras CCD (charged coupled device), et des sons par échantillonnage et codage des impulsions, est une technique relativement ancienne. Ce n'est que depuis le milieu des années 1980 que des micro-ordinateurs suffisamment puissants, à des prix abordables, en ont permis l'usage dans de nombreux domaines, notamment l'éducation, les jeux, l'art, l'édition, le cinéma et la télévision. En effet, une image numérisée complexe qui puisse être reproduite avec une qualité suffisante sur un écran d'ordinateur occupe autant de place en mémoire qu'un livre d'une centaine de pages de texte. Une séquence vidéo doit avoir une cadence de 25 images par seconde, dix minutes de vidéo numérique occuperont l'espace de plus de 4 milliards de caractères. Seules les techniques de compression de données permettent le stockage de cette séquence sur un Cd-Rom, qui a une capacité de 660 millions de caractères. Encore cette séquence vidéo est-elle muette ; si l'on veut ajouter du son, l'espace nécessaire croîtra d'autant.

La digitalisation et la compression ont apporté des solutions à certains verrous technologiques, notamment sur l'accès et la **modification** des éléments qui composent un document. Pour la vidéo par exemple, de nombreux logiciels inspirés des bancs de montage cinématographiques permettent d'ajouter, de modifier, de supprimer ou de déplacer des images ou des sons à l'intérieur d'une vidéo. Parmi les outils disponibles, les logiciels commerciaux Final Cut et Adobe Premiere sont très utilisés dans le domaine de la vidéo personnelle. De nombreux logiciels semblables sont aujourd'hui sur le marché, ils permettent de modifier des sons et des images, de leur appliquer des filtres ou des effets : les possibilités sont illimitées.

Le développement récent de ces technologies de diffusion et de création des documents multimedia a entraîné la création de grandes archives de documents mono et multimedia : photos, sons, textes. L'énorme quantité de données contenues dans ces bibliothèques numériques rend très complexe l'accès et donc l'utilisation de cette information. C'est pourquoi il est nécessaire de créer des systèmes de gestion de ces bases de données multimedia. Quatre fonctionnalités majeures sont attendues. Ils doivent permettre :

1. de créer et de gérer de grandes bases de données contenant plusieurs documents de taille variable ;
2. de rechercher, dans ces bases de données, les documents ou extraits de documents qui correspondent à un critère de recherche donné ;

3. de créer dynamiquement et automatiquement un document « résumé » des données recherchées ;
4. de garantir des niveaux de qualité et de performance acceptables pour l'utilisateur.

Pour les tâches (2) et (3) de recherche et de résumé de documents, il est nécessaire de modéliser et de représenter l'information contenue dans les documents sous la forme d'une liste ordonnée de données significatives, appelée **index**.

1.1.3 L'indexation de vidéo

Les index sont au cœur des applications de traitement automatique des documents multimedia et plus particulièrement des vidéos. Pour supporter efficacement l'utilisation de l'information, et les demandes de l'utilisateur, ces index doivent être aussi riches et complets que possibles. Nous cherchons ici à modéliser "l'histoire" véhiculée par les films de cinéma. De manière générale, trois types d'informations sont attachées à un document :

- Un niveau bas qui décrit les caractéristiques simples des éléments d'un document comme les couleurs ou la texture d'une image, ou l'enveloppe d'un son ;
- Un niveau sémantique qui fournit une description de haut niveau de ce que contient la vidéo, qu'il s'agisse de personnages, de lieux, ou d'objets, et de leurs interactions ;
- Un niveau structurel qui met en évidence une organisation du document.

Le travail de recherche et d'extraction des caractéristiques, de la structure, et de la sémantique à des fins de modélisation constitue le travail d'indexation de données. Le résultat de l'indexation d'un document est une description numérique (exploitable par une machine) dans un ou plusieurs formalismes qui permet l'accès, la recherche, le résumé, la classification et la réutilisation de tout ou partie du document. Selon le niveau de représentation ciblé, il est possible ou non d'extraire automatiquement l'information recherchée. En un sens, la tâche d'indexation correspond donc à une nouvelle compression du signal des documents, réalisée afin d'extraire les informations utiles pour leur traitement.

Plusieurs formalismes de représentation ont été proposés pour les trois niveaux évoqués ci-dessus. En ce qui concerne la vidéo, l'indexation est souvent traitée comme l'inverse du processus de création. Un film est issu des actions de l'auteur, du réalisateur, ou du monteur qui sont orientées par la nécessité de créer une narration et par les caractéristiques physiques des media utilisés. Cette approche structurelle est essentiellement développée afin de proposer des applications de recherche par le contenu. Mais, ces descriptions peuvent aussi servir de base pour les manipulations de documents audiovisuels que nécessite une classe très large d'applications : résumé de documents, représentation simplifiée de la base, etc... Ces descriptions sont non seulement le moyen d'attacher des informations à différents niveaux de contenus des documents audiovisuels, mais elles permettent aussi de définir des structures d'organisation des documents sur lesquelles les applications peuvent s'appuyer.

Le standard **MPEG-7** [Com01], aujourd'hui reconnu, fédère et supporte différentes descriptions telles que les caractéristiques du contenu de niveau bas (forme, taille, texture, couleur, mouvement, position...), les informations sur le contenu sémantique (personnage, lieu, action...), ou encore les descripteurs culturels (auteur, date de création, format...). Il est également possible, en utilisant ce standard, de décrire les relations de structure temporelles et spatiales entre les objets qui composent la vidéo. Ces descriptions MPEG-7, appelées « Description Schemes », sont formalisées en utilisant le langage XML Schéma et peuvent être instanciées comme des documents XML.

1.2 Indexation du contenu

1.2.1 Historique des systèmes de recherche par le contenu

La connaissance est habituellement définie comme l'ensemble des faits du monde et est souvent représentée par des **concepts** et des relations entre ces concepts : les réseaux sémantiques. Les concepts sont des abstractions d'objets, des événements de situations ou des modèles perceptuels dans le monde (par ex. un modèle de couleur et le concept "Voiture"). Les relations représentent des interactions entre les concepts, par exemple un modèle couleur est visuellement similaire à un autre modèle, et "la berline" est une spécialisation du concept "Voiture".

Les premiers systèmes de recherche de documents sont donc basés sur des **index textuels** du contenu, dans lesquels les documents sont annotés manuellement par des concepts et la recherche est exécutée sur l'information ainsi extraite (voir [Cha92] pour une étude sur les systèmes de recherche d'image basés sur le texte). L'usage de mots clés pose plusieurs problèmes : l'annotation manuelle de documents est très coûteuse et en soi incomplète. La relation entre les mots et les concepts est parfois complexe en raison de phénomènes tels que synonymie (les mots différents indiquent le même concept) ou homonymie (le même mot indique des concepts différents) et quelques concepts ne peuvent pas être décrits par mots clés.

Afin de surmonter ces difficultés (principalement le coût), les premiers systèmes de recherche basés sur l'extraction automatique du contenu sont proposés dans les années 1990. Au lieu d'annotations basées sur le texte, les images sont indexées, et recherchées par traitement de **descripteurs numériques** visuels, tels que la couleur, la texture ou la forme. Plusieurs systèmes sont introduits pour chercher des images et des vidéos en utilisant leurs contenus visuels (voir [Idr97, For02] pour des études sur l'indexation d'image et de vidéo et les technologies de recherche de documents).

Jusqu'à aujourd'hui, la plupart des méthodes recherchent des documents par traitement de descripteurs numériques. A l'exception de systèmes qui peuvent identifier des objets caractéristiques comme les visages, les voitures [Sch00b], les individus [Fle96], ou les piétons [Ore97], l'indexation automatique n'est pas d'habitude dirigée vers des objets sémantiques.

Cependant, l'usage du contenu bas a aussi des limitations ; ces systèmes basés sur les caractéristiques de bas niveau ne satisfont pas les besoins des utilisateurs, car les études [Ens93, Mar00] montrent qu'ils semblent être surtout intéressés par la sémantique des documents. En raison du "fossé sémantique" entre les requêtes des utilisateurs et les descripteurs utilisés, ils éprouvent une difficulté pratique à formuler des recherches visuelles.

Pour pallier ces limitations, l'assignation automatique de concepts est nécessaire. Afin d'exploiter entièrement le potentiel de ces descripteurs textuels, une relation de compréhension entre les mots clés et le contenu numérique doit être définie. Cette relation est établie par catégorisation automatique du contenu du signal et l'annotation sémantique doit ainsi être approchée comme un problème d'apprentissage de modèles de classifications.

1.2.2 Catégorisation automatique du contenu

La classification permet d'associer à un élément du document la valeur prise par un concept. On appelle **ontologie** de classification, l'ensemble des concepts sélectionnés pour décrire les documents, et **taxonomie**, l'ensemble des valeurs possibles pour un concept.

Le choix de l'ontologie de classification est essentielle pour les systèmes d'indexation. "C'est cette désignation du visible qui, par une sorte de tri prélinguistique, lui permet de se transcrire dans le langage"[Fou66]. L'idée consiste à choisir une structure limitée (un ensemble de concepts) à partir de laquelle on étudiera,

dans tous les documents qui se présentent, les identités et les différences. Dès lors, toute différence ou identité ne relevant pas de ce caractère ne devra pas être prise en compte. Par exemple, lorsque Linné (botanique) choisit pour note caractéristique "toutes les parties de la fructification", les différences de feuille, de tige ou de racine devront être systématiquement négligées. Le choix de cette structure fixe permet d'effectuer la même description au sujet d'un document. Par exemple, tout un chacun pourra vérifier qu'une fleur donnée est circulaire ou hexagonale, que sa tige a telle taille,...etc. Foucault souligne ce point en écrivant qu' "en cette articulation fondamentale du visible, le premier affrontement du langage et des choses pourra s'établir d'une manière qui exclut toute incertitude". Il ajoute encore : "la structure, en limitant et en filtrant le visible, lui permet de se transcrire dans le langage". De nombreuses ontologies fixes ont été créées pendant la dernière décennie. Celles-ci concernent plusieurs domaines de la recherche, notamment la science médicale [Rec99], le Web [Lee01], le multimedia [ace], et le traitement de la vidéo [Loz98].

En général, tout débute par la définition d'un ensemble de **concepts moyens** ou « atomiques » qui, on l'assume, est assez large pour couvrir un espace de recherche d'intérêt. Par concept moyen, nous entendons des concepts comme le ciel, la musique, l'eau, la parole etc..., qui ne peuvent pas être décomposés ou représentés directement en termes d'autres concepts. L'estimation de ces descripteurs textuels est réalisée par classification d'un ou de plusieurs éléments de niveau bas.

Les concepts qui peuvent être décrits en termes d'autres concepts, les **concepts hauts**, sont déterminés par classification de la description basse et moyenne. Ils vont ainsi dépendre de la variété des descripteurs bas et des concepts atomiques définis, ainsi que du domaine de la base de données.

Nous notons que ces concepts sont définis indépendamment de la modalité dans laquelle ils s'expriment naturellement. Un concept atomique peut être multimodal et un concept de plus haut niveau peut être unimodal. Notons aussi que certains mots clés trouvés dans les annotations manuelles ne se réfèrent pas à des caractéristiques du signal, même si pour quelques bases de données spécifiques ils peuvent servir à annoter des documents partageant quelques caractéristiques communes du signal ; leur association avec le contenu peut produire de faux résultats de recherche. C'est pourquoi il est nécessaire de choisir avec soin l'ontologie du système de classification.

En pratique ce modèle d'ontologie est intéressant puisqu'il envisage la construction de nouveaux systèmes de connaissances comme un assemblage de composants réutilisables. Ainsi la connaissance peut être réemployée et le seul effort qu'un programmeur doit produire lorsqu'il crée un nouveau système est de définir les connaissances spécifiques qui n'ont pas encore été apprises. Les ontologies ainsi modelées entraînent d'autres avantages. Les systèmes sont alors capables de communiquer entre eux des connaissances et des raisonnements ; ce qui leur permet d'accomplir des tâches qui étaient jusque là hors de leurs capacités. Cette caractéristique a eu pour conséquence l'élaboration de systèmes basés sur le contenu plus vaste et complet avec une quantité minimum de coût et d'effort. Ces particularités essentielles pour le traitement de l'information, le partage des connaissances, l'extensibilité, l'interopérabilité et l'intercommunication entre machines, a intéressé beaucoup de chercheurs provenant de champs différents : aujourd'hui il subsiste encore beaucoup d'efforts sur la construction et l'utilisation d'ontologies (e.g. Worldnet).

1.2.3 L'indexation du contenu multimedia

Un aspect caractéristique de l'indexation de la vidéo par rapport à d'autres types d'indexation est la présence d'informations provenant de différents canaux d'informations ou modalités. Pour les séquences de films, on peut identifier trois modalités premières :

- **visuel** : naturel ou artificiel, ce qui est vu dans une vidéo,
- **audio** : parole, musique, et sons environnementaux, qui peuvent être entendus dans la vidéo.
- **textuel** : ressources textuelles, transcription des paroles et labels textuels qui décrivent le contenu de la vidéo.

Dans l'analyse multimedia, la plupart des outils sont habituellement dévoués à une seule modalité, les autres étant traités comme des illustrations ou comme des éléments complémentaires. Par exemple, les logiciels de recherche sur le Web n'utilisent pas l'image, les systèmes de recherche d'images mélangent à peine les descriptions visuelles et textuelles, l'analyse de vidéo est habituellement faite séparément sur le son et l'image. Une des raisons de cela est que ces media concernent des champs scientifiques différents et parfois très séparés. Le principal raccourci de ces approches est leur inaptitude à intégrer les informations provenant de plusieurs modalités. Une indexation efficace nécessite une approche multimodale dans laquelle la collaboration de différentes modalités est utilisée. L'intégration des trois modalités est courante dans le cadre de l'interprétation sémantique des informations de la vidéo par des indexeurs humains. Actuellement, les recherches s'orientent vers une automatisation de cette tâche. Plusieurs études ont montré que les performances des analyses et de la compréhension automatique du multimedia (notamment en terme de robustesse) peuvent être grandement améliorées par la combinaison de différentes modalités [Cas98, Lin03]. Une étude récente de l'indexation multimodale automatique de la vidéo est donnée dans [Bru99]. Cette étude forme une vue complète du champ de l'indexation multimodale de la vidéo.

Le sujet de la combinaison de l'information audio et visuelle appartient au problème général de la fusion de plusieurs sources. Les applications monomedia qui reposent sur un médium unique pour acquérir l'information des documents sont sujettes à des erreurs ; quelle que soit la qualité de l'information d'un signal, il fournit au système une vue unique de ce qui se passe. Afin d'offrir aux applications un traitement robuste, il est nécessaire de se fier à plusieurs media pour réunir l'ensemble des informations correctes. La **coopération des modalités** qui ont un fort degré de redondance et de complémentarité garantit une perception précise. Il est possible d'utiliser la redondance des media pour valider les données qu'ils fournissent. La complémentarité des media est aussi utilisée pour résoudre les ambiguïtés ou réduire les erreurs quand une perturbation environnementale affecte le système. De nombreux domaines d'analyse ont été particulièrement étudiés :

- Transcription de la parole avec utilisation des données visuelles, comme le suivi des lèvres ou des expressions faciales.
- Détection d'unités logiques pour l'indexation de journaux télévisés.
- Détection de dialogues.
- Identification de personnes.
- Détection de concept sémantique comme les tirs de fusée, la présence d'une voiture.

Jusqu'en 2000, la plupart des travaux dans ce domaine étaient un peu heuristiques et très dépendants du domaine d'application. Une tâche difficile est de développer un cadre théorique pour le traitement joint de l'audio et de la vidéo, et plus généralement, pour le traitement multimodal. Le problème principal qui est commun à tous les systèmes basés sur la fusion d'informations concerne la décision : quelle sera la décision finale quand les différents media ou sources d'information donnent des données **contradictoires** ?

1.3 Indexation de la structure

1.3.1 Structure des films de cinéma

L'indexation de la structure est une tâche essentielle du traitement des documents multimedia. Elle trouve sa justification dans le fait que, comme la table des matières d'un livre, elle fournit un accès direct aux différents composants du document. Dans le cadre du traitement des films, il semble clair que c'est le **temps** et lui seul, qui structure de manière fondamentale et déterminante tout récit cinématographique, l'espace n'étant jamais qu'un cadre de référence secondaire et annexe. C'est donc par rapport au traitement qu'elle fait subir au temps, que doit être analysée la construction d'un film.

Un livre est divisé en chapitres, contenant des sections, qui consistent en sous-sections qui comprennent des paragraphes, des images et des tableaux. De même, il est maintenant largement accepté que les films sont **hiérarchiquement structurés** en scènes, plans et images. Les documents ainsi structurés sont représentés sous la forme d'un dendogramme consistant en la structure type du document considéré : un noeud correspond à un élément structurel du document (i.e. film, scène, plan, image), un arc représente la relation d'inclusion d'un élément dans un autre (e.g. un plan est inclus dans une scène). Afin de réunir l'indexation du contenu et l'indexation de la structure, les attributs descriptifs de niveau signal ou sémantique sont associés à chaque nœud. Ces caractéristiques sous forme d'index décrivent le contenu du nœud. Les valeurs des attributs peuvent descendre ou monter le long de la hiérarchie.

1.3.2 Extraction automatique de la structure

L'extraction de la structure réalisée manuellement manque généralement de précision temporelle. Or, si cela est peu pénalisant pour les usages actuels de la documentation, un découpage approximatif peut fausser l'utilisation par d'autres applications des segments obtenus : résumé, catalogue thématique d'extraits, publication d'un corpus annoté, etc. . .L'automatisation des tâches de segmentation et de structuration devrait permettre d'une part de concentrer le travail humain sur des activités à plus forte valeur ajoutée, et d'autre part d'améliorer la précision temporelle des segments obtenus.

La segmentation automatique permet de tracer des frontières au sein de la représentation d'un objet multimedia. Elle isole des sous-objets dont le contenu est homogène selon l'information choisie pour segmenter. La problématique essentielle de la structuration est donc le choix des descripteurs utilisés pour représenter l'information fournie à l'algorithme de segmentation. La sélection de cette information parmi l'ensemble du contenu disponible (descripteurs numériques et concepts) dépend essentiellement des caractéristiques des sous-objets désirés.

1.3.3 Fusion des media

D'un point de vue scientifique, les travaux traitant de la segmentation automatique s'attachent essentiellement à une analyse monomedia (bande image ou bande son) alors que, dans de nombreux cas, une combinaison de différents media devrait permettre d'améliorer les performances. Pour rester dans le cadre de la **fusion**, nous examinerons donc les méthodes qui combinent l'information audio et visuelle pour atteindre leur but.

La fusion de l'information issue de plusieurs media pour la segmentation n'est pas une tâche triviale. Deux des problèmes rencontrés par les procédés sont les suivants :

Problème de **décision** qui est commun à tous les systèmes basés sur la fusion d'information : qu'elle sera la décision finale quand les différents media ou sources d'informations donnent des données contradictoires.

Problème de **synchronisation** qui est spécifique à l'intégration multimodale. En effet, la fréquence d'échantillonnage des descripteurs de bas niveaux dépend des media : l'élément minimal d'un signal vidéo est l'image ; quand la fréquence d'échantillonnage est de 25Hz, il est possible d'obtenir des descripteurs tous les 40ms. L'élément minimal du signal audio est la «frame» ; quand la fréquence d'échantillonnage est de 100Hz, il est possible d'obtenir des vecteurs descripteurs tous les 10ms.

Pour résoudre ces deux problèmes, différentes façons d'intégrer les descripteurs audio et visuels ont été explorées. Un premier modèle consiste à les combiner dans un seul vecteur descripteur audio-visuel avant la segmentation. La deuxième technique pratique deux segmentations indépendantes, une pour chaque modalité, puis les résultats sont fusionnés. Nous verrons par la suite que le premier modèle ne peut être appliqué dans le cadre du modèle de fusion des media développé dans cette étude, en raison de la structure particulière des films de cinéma.

Chapitre 2

État de l'art et objectifs

Ce deuxième chapitre étudie les méthodes existantes permettant l'indexation du contenu et de la structure des documents multimedia. Dans le paragraphe 2.1, nous décrivons l'information apportée par les descripteurs du traitement des films. Dans le paragraphe 2.2 nous rappelons les différents algorithmes de classification automatique du contenu. Nous présentons ensuite un état de l'art des méthodes de fusion des descripteurs numériques dans le paragraphe 2.3 et de fusion des annotations et des descripteurs numériques dans le paragraphe 2.5. Dans le paragraphe 2.6, nous identifions les techniques de segmentation employées pour la structuration des vidéos et plus particulièrement celle des films. Enfin, dans la paragraphe 2.7 nous résumons les limitations de ces méthodes en termes de performances, et nous formulons les objectifs de ce travail.

2.1 Informations fournies par les descripteurs

De façon générale les documents multimedia dont on veut extraire l'information sont stockés au sein de bases de données dans des formats permettant leur lecture (ou affichage), par un logiciel de diffusion. Ces formats ne contiennent pas d'information sur le sens du document. Afin de traiter de la sémantique, l'utilisation d'index de descripteurs, c'est-à-dire de données qui décrivent elles-mêmes le contenu des données, devient incontournable. Les descripteurs sont extrêmement variés, et l'on peut les classer suivant divers critères : leur origine, leur niveau d'abstraction, leur subjectivité, etc.

2.1.1 Origine des descripteurs

Descripteurs culturels

Les descripteurs culturels sont des données liées au **contexte** dans lequel un document a été créé. Ils sont textuels (e.g. nom de l'auteur, titre) et extractibles par recherche dans des bases de données, sur Internet ou ailleurs. Cependant, ces descripteurs ne donnent que peu d'information sur les caractéristiques physiques et sémantiques du document.

En ce qui concerne le film de cinéma, plusieurs bases de données de descripteurs culturels existent sur Internet, la plus connu étant imdb.com [imd]. Elles contiennent de nombreuses informations essentielles pour la recherche de films : titre, distribution, résumé textuel du film, photo des acteurs... Dans ce travail, nous nous intéressons exclusivement à l'extraction automatique de descripteurs. Ainsi seuls quelques descripteurs culturels seront utilisés comme information a priori sur le film : le nom des acteurs pour la reconnaissance des voix, par exemple.

Descripteurs du signal

Afin de décrire véritablement les propriétés des documents analysés indépendamment de leur contexte culturel, nous nous intéressons exclusivement aux descripteurs extraits à partir du **traitement des données**. Nous les appellerons descripteurs du signal. Ils sont issus de calculs numériques ou d'annotations manuelles effectués sur les données du document traité. Ils peuvent être numériques ou textuels.

2.1.2 Niveau sémantique des descripteurs du signal

Les descripteurs du signal se distinguent principalement par le niveau d'abstraction des propriétés qu'ils évaluent par rapport à celles du signal brut ou **niveau sémantique**.

Les descripteurs de niveau bas

Les caractéristiques de niveau bas sont des éléments d'indexation qui peuvent être extraits automatiquement par application d'algorithmes et **sans connaissance** particulière du contexte. Ils expriment des propriétés significatives directement observables ou calculables à partir du signal, comme par exemple les caractéristiques du spectre d'un signal audio, ou de l'histogramme des couleurs d'une image. Ils s'expriment sous forme numérique et sont stockés au sein de vecteurs dont les dimensions correspondent aux différentes caractéristiques du descripteur. L'intérêt principal de ces caractéristiques bas niveau est qu'elles peuvent être déterminées automatiquement à partir des documents en s'affranchissant de la sémantique du contenu. Mais, ces éléments d'indexation sont de trop bas niveau d'abstraction pour traduire le sens de l'information contenu dans les documents traités. Le courant actuel des recherches dans le domaine vise à apparier ces caractéristiques bas niveau à des informations de plus haut niveau, afin d'automatiser en partie, ou complètement, le processus d'indexation du contenu.

FIG. 2.1 – Descripteurs bas extraits à partir d'une image.

Les descripteurs de niveau moyen

Ils expriment des propriétés ayant une **signification plus évidente** ou compréhensible par un utilisateur, comme le nombre d'instruments dans un morceau de musique, la présence d'herbe dans une photo, etc... L'extraction automatique de caractéristiques de niveau moyen est une tâche extrêmement compliquée. Les derniers efforts convergent vers l'extraction de **concepts**. Les concepts sont dits qualitatifs lorsque qu'ils s'expriment par une description naturelle du langage. Dans ce cas, ils sont issus d'une grande variété d'expressions linguistiques et peuvent être un mot unique courant (montagne, ciel), une phrase nominale (offre de dernière minute), une expression contenant des conjonctions (A et B). Certaines formules plus complexes peuvent être composées de phrases prépositionnelles, de phrases à verbes et de ponctuation. Les concepts sont dits quantitatifs lorsqu'ils s'expriment par des nombres, comme le nombre de personnes dans une image ou le nombre d'instruments présents dans un extrait musical.

A un concept est associé une **taxonomie** : l'ensemble des valeurs qu'il peut prendre. Ces valeurs peuvent être continues, mais le plus souvent elle sont représentées par une variable discrète textuelle

ou numérique. La valeur prise par un concept pour un document donné est déduite par classification automatique ou annotation manuelle du document selon la taxonomie choisie. L'automatisation de cette tâche nécessite le calcul de modèles numériques complexes à partir de données extérieures au signal lui-même. Les paramètres de ces modèles sont appris sur les descripteurs de **niveau bas** de documents d'exemples annotés. Et une ou plusieurs données numériques sont associées aux valeurs des concepts ainsi extraites, un taux de confiance (estimant la validité présumée de la classification) par exemple.

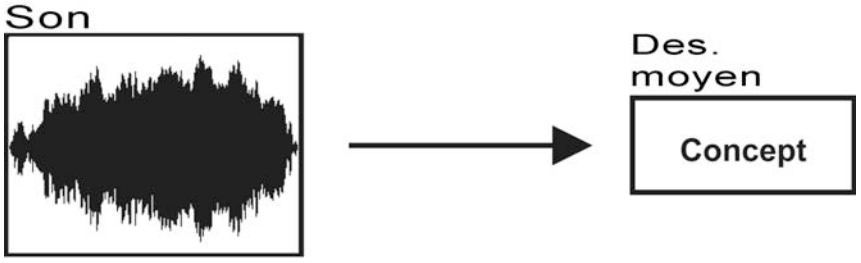

FIG. 2.2 – Descripteur moyen extrait par l'annotation manuelle d'un son.

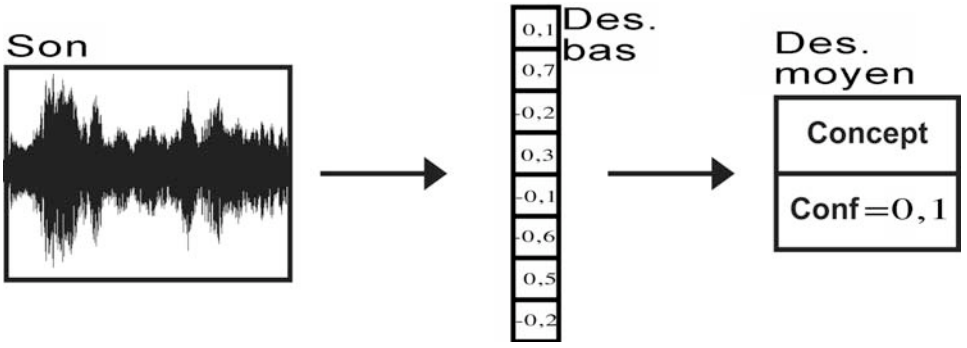

FIG. 2.3 – Descripteur moyen et taux de confiance associé extraits par classification des descripteurs bas d'un son.

Les descripteurs de niveau haut

Les descripteurs de niveau haut expriment des propriétés qui ont une **signification forte**. Ce sont des **concepts** qui peuvent être décrits en termes d'autres concepts. Leur catégorisation automatique nécessite le calcul de modèles d'apprentissage textuels et numériques à partir de descripteurs de niveaux bas et moyen issus de document d'exemple. Dans le cadre de l'analyse de films, ils expriment trois grands types de caractéristiques : le temps, le lieu, les relations entre objets (ou personnes) : actions.

Les frontières entre les différents niveaux sémantiques sont souvent floues. Pour simplifier : le niveau bas correspond aux descripteurs numériques, le niveau moyen aux concepts simples et le niveau haut aux concepts déterminés à partir d'autres concepts. Nous verrons au chapitre 3 une définition plus complète de ces niveaux.

2.1.3 Granularité spatio-temporelle des descripteurs

Dans le cadre de cette thèse, les descripteurs du signal seront extraits sur l'ensemble des données du film. Ces données sont caractérisées sur trois dimensions : deux spatiales, et le temps. La granularité spatiale, temporelle ou spatio-temporelle à laquelle est déterminé un descripteur est un critère essentiel de discrimination.

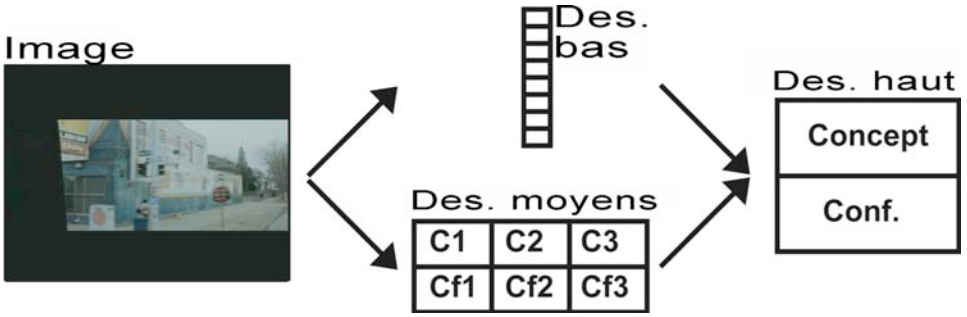

FIG. 2.4 – Descripteur haut extrait par classification des descripteurs bas et moyen d'une image.

Au niveau local

Il s'agit en général de descripteurs monomedia. Ils sont calculés sur une portion de petite taille de signal. Pour le son, sur un segment d'un échantillon à quelques dixièmes de seconde. Pour l'image, sur des voisinages de quelques pixels de l'image. Pour la vidéo, sur une zone de la vidéo durant de une image à une dizaine d'images.

Au niveau intermédiaire ou semi-local

Les descripteurs sont calculés sur une portion plus importante du signal. Les segments de son ou zones d'image peuvent être de taille fixe mais ils sont le plus souvent de taille variable, car issus de la segmentation temporelle ou spatio-temporelle du signal.

Au niveau global

Ils sont calculés sur l'intégralité du media considéré : la couleur moyenne d'une photo, le timbre global d'un morceau de musique.

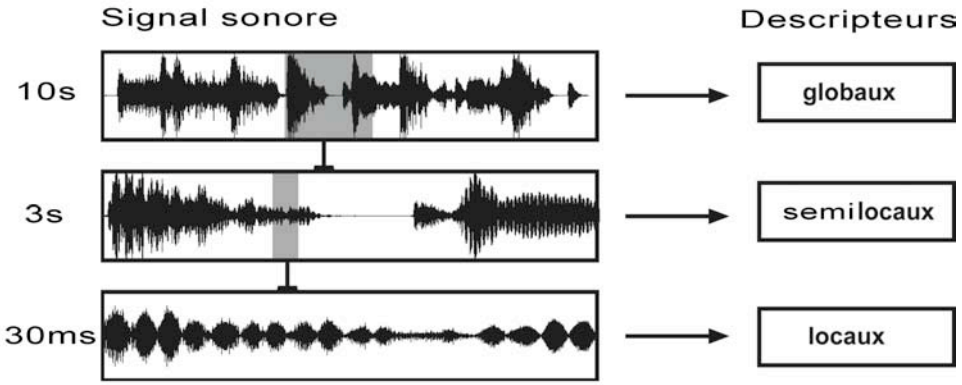

FIG. 2.5 – Descripteurs des trois niveaux de granularité d'un échantillon de musique.

Dans les trois chapitres suivants, nous présentons plusieurs descripteurs du signal de la littérature. Ces descripteurs sont caractérisés par les trois critères : le media dont il sont issus (audio, image, vidéo, multimedia), leur niveau sémantique, leur granularité temporelle.

2.1.4 Descripteurs du signal Audio

Un grand nombre de descripteurs numériques du signal audio ont été proposés afin de décrire le contenu du son. Ils proviennent de plusieurs domaines de recherche : la reconnaissance de la parole [Foo94], la classification du son [Sch97, Liu98], mais aussi de la communauté psychoacoustique [Pee00]. Les descripteurs du signal s'organisent suivant leur niveau sémantique.

Descripteurs bas audio

Au **niveau local**, les descripteurs numériques sont calculés sur un segment d'un échantillon à quelques dixièmes de seconde. On peut distinguer plusieurs familles de descripteurs liés au traitement appliqué pour les extraire :

- Descripteurs temporels. Ils sont extraits à partir de la forme d'onde ou de la courbe d'énergie (enveloppe) : taux de silence [Sch97, Bur98], corrélation croisée, énergie à court terme [Zha01], volume [Liu98], écart dynamique du volume [Liu98], modulation à 4Hz [Sch97, Liu98], pulsation [Sch97], taux de réverbération [Cou01].
- Descripteurs d'énergie. Ils décrivent l'énergie de différents types de contenus : énergie globale, énergie harmonique, énergie du bruit, énergie de bandes [Sch00a].
- Descripteurs spectraux. Ils sont calculés à partir de représentation 2D du signal (fréquence, temps), transformée de Fourier à court terme (STFT) ou transformée en ondelettes (DWT) [Sub98], centroïde spectral [Sch97, Mar98], flux spectral [Sch97], point de "roll-off" [Sch97], bande passante [Liu98], statistiques de bandes et ratios [Tza01], periodicité de bande [Liu01], inharmonicité [Mar98], rythme et tempo [Got96, Sch00a].
- Descripteurs harmoniques. Ils sont calculés à partir d'un modèle harmonique du signal : Fréquence fondamentale [Mar98, Zha01], Vibrato [Mar98], Slope [Mar98].
- Descripteurs perceptifs. Ils sont déterminés à partir d'un modèle d'écoute humaine : Loudness, Sharpness [Sch00a], "Mel-Frequency Cepstral Coefficients" ou MFCC[Foo99, Tza01]. Les MFCC se présentent sous la forme d'une représentation 2D (fréquence, temps). L'échelle des fréquences est fixée afin de correspondre à la capacité cognitive de l'oreille humaine.

Dans le cadre des applications présentées dans cette thèse, nous utilisons plusieurs descripteurs numériques locaux. Certains sont indépendants du domaine d'application (**MFCC**), mais la plupart ne le sont pas. Pour exemple : la **modulation à 4 Hertz** permet de détecter une modulation caractéristique de la parole pour la classification du son en *parole/non-parole* ; le temps de réverbération peut être utilisé pour la classification des sons d'ambiance en *intérieur/extérieur*.

Au **niveau intermédiaire et global**, les descripteurs sont issus de statistiques et d'opérations classiques sur les descripteurs : moments statistiques, maximum, minimum, médiane, dérivées, corrélations, valeur de modulations, entropie [Sch00a], paramètres de modèles gaussiens [Mar99]. On peut considérer qu'ils sont un **« résumé »** de la séquence des descripteurs locaux extraits sur le segment considéré du document.

Descripteurs audio de niveau sémantique moyen

Le niveau sémantique moyen correspond aux concepts simples extraits du contenu sémantique bas audio. Ces concepts sont déduits par catégorisation automatique de segments de son dans des classes prédéfinies. La tâche de classification nécessite le calcul de modèles numériques complexes appris à partir de segments annotés. Dans le cadre de nos applications, ces segments sont de durée fixée à 3 secondes. Il s'agit donc de descripteurs intermédiaires.

Plusieurs descripteurs moyens ont été proposés afin de décrire les caractéristiques d'un son. Ils concernent des informations variées : l'identité de la source d'émission (*parole/musique/bruit*) [Del03],

le nombre de locuteurs dans un dialogue [Mei01], la présence de silence [Ros00]. Ces concepts sont accompagnés d'un taux de confiance dans la mesure, que l'on définira dans le chapitre consacré à la classification.

La classification en *parole/musique/bruit* est un concept moyen essentiel pour le traitement du signal audio. Elle est la première étape de traitements plus complexes, notamment la reconnaissance de personnages par la voix, ou du style d'un morceau de musique. Les techniques développées dans la littérature sont capables de classer correctement approximativement 95% des extraits sonores selon cette taxonomie.

Descripteurs audio de niveau sémantique haut

Le niveau sémantique haut correspond aux concepts complexes extraits par classification des descripteurs bas et moyens de segments de sons. L'automatisation de l'indexation pour ce type de descripteurs est toujours un sujet de recherche. On peut citer notamment, **la reconnaissance du locuteur** [Hat91], l'identification d'un style de musique [Tza99] ou du nom d'un morceau [Ghi95].

2.1.5 Descripteurs du signal image

Il existe une littérature abondante sur les descripteurs du signal image. Comme pour l'audio, nous distinguerons principalement les descripteurs numériques et les concepts ou descripteurs textuels.

Descripteurs numériques Image

Un grand nombre de descripteurs numériques ont été proposés afin de décrire le contenu d'une image. Le plus souvent, les techniques d'extraction d'information liées à l'analyse d'une seule image, visent à la segmenter en régions afin d'en extraire des informations intermédiaires concernant les couleurs, les textures et les formes qu'elles contiennent.

C'est ce qui est proposé notamment par les systèmes d'interrogation de bases d'images QBIC[Ma99]. Le standard **MPEG-7** [Sal01] propose, dans sa partie dédiée à la vidéo, des schémas de description qui permettent la description des couleurs des régions d'une image (ou d'une image entière, ou d'un groupe d'images) par détermination de différents espaces de couleur, en utilisant les couleurs dominantes ou des histogrammes. MPEG-7 peut également rendre compte de la texture des régions, au niveau bas par des filtres de Gabor ou à un plus haut niveau en utilisant trois caractéristiques : la régularité, la direction et la granularité. Enfin, la forme des régions d'une image peut être décrite par une représentation des contours basée sur la courbure multi-échelles ou sur des histogrammes de formes.

Dans le cadre de cette étude, nous avons choisi d'employer les descripteurs globaux de l'image développés au CEA dans le cadre du projet d'indexeur PIRIA6 [Joi04]. L'indexeur qui utilise l'histogramme présenté par [Che03b] appelé **TextureLEP** fournit les meilleurs performances en recherche d'images par similarité. Cet histogramme est composé de deux parties : une partie texture et une partie couleur. La partie texture est fondée sur les motifs des contours locaux ("local-edge pattern" LEP). Les contours de l'image sont d'abord calculés avec un filtre de Sobel 3x3. Après seuillage, une image binaire des contours est obtenue. Ensuite, pour chaque pixel de cette image, la fenêtre 3x3 autour de ce pixel est considérée. Il y a $2^9 = 512$ configurations possibles numérotées. A chaque pixel central est associé le numéro de la configuration dans laquelle il se trouve. Il est alors possible de construire un histogramme de 512 composantes. La partie couleur, quant à elle, est un histogramme RGB : R, G et B sont quantifiés chacun en 4 valeurs, ce qui donne un histogramme de $4^3 = 64$ composantes. L'histogramme final est donc constitué de 576 composantes. Dans la suite, on appellera cet histogramme TLEP. [Mil04] a comparé TLEP avec un histogramme global de l'espace couleur HSV quantifié en 41 valeurs, et CCIV (Colour Coherence

Incoherence Vector) qui est un histogramme couleur modifié pour tenir compte de la surface des régions [GP96]. D'après ces tests, il semble que TLEP soit un bon outil pour la classification des images.

Descripteurs image de niveau sémantique moyen

Le niveau sémantique moyen correspond aux concepts simples extraits du contenu de l'image. Les descripteurs moyens pour l'image sont en général de granularité spatiale intermédiaire. Ils sont déduits par catégorisation automatique de zones d'images dans des classes prédéfinies. Ces zones sont le plus souvent issues de la segmentation spatiale d'une image et sont homogènes dans les caractéristiques liées à la classification choisie. La tâche de classification nécessite le calcul de modèles numériques complexes appris à partir de zones d'images annotées. Ces descripteurs concernent la détection de figures particulières :

- **Objets** : voiture[Sch00c], visage[Row98, Vio02, Kou03].
- **Textures** : herbe, arbre, eau, ciel, bâtiment, neige, sable. [Mil04]

Ces classifications sont de la forme concept absent/présent (0/1) et sont accompagnés d'un taux de confiance dans la mesure que l'on définira dans le chapitre consacré à la classification.

La **détection de personne** est un concept moyen essentiel pour le traitement des films. Il ouvre la voix à d'autres concepts hauts, notamment la reconnaissance de personnages. Ce type de détection n'est pas un problème trivial en raison de la variation en position, taille, éclairage et orientation des visages. Les modifications causées par l'expression faciale, une barbe, des lunettes, et les occultations compliquent encore le problème. Nous emploierons la technique de détection de visage basée sur l'algorithme Adaboost [Vio02] développée au CEA. Elles sont capables de détecter correctement approximativement 90% des visages en position frontale et verticale. Certaines techniques plus avancées ne détectent pas juste les visages ou la tête, mais aussi le corps en entier. L'algorithme applique une détection indépendante de la tête, des jambes et des bras. Après que la configuration géométrique de l'ensemble des parties détectées est validée, une classification de second niveau combine les résultats partiels des détecteurs pour classer un candidat comme personne ou non-personne.

La **détection d'objets** est une extension de la détection de personnes. Des objets spécifiques peuvent être décelés dans une image en appliquant des classifications spécialisées sur les descripteurs bas (couleur, mouvement, et forme). Un exemple de détection d'objets spécifiques basée sur l'apparence visuelle est donné en [Sch00c]. Les auteurs décrivent une technique qui détecte la présence de voitures dans une séquence vidéo. Ils utilisent le produit d'histogrammes, ou chaque histogramme représente la probabilité jointe de coefficients détecteurs et de leur position dans l'objet.

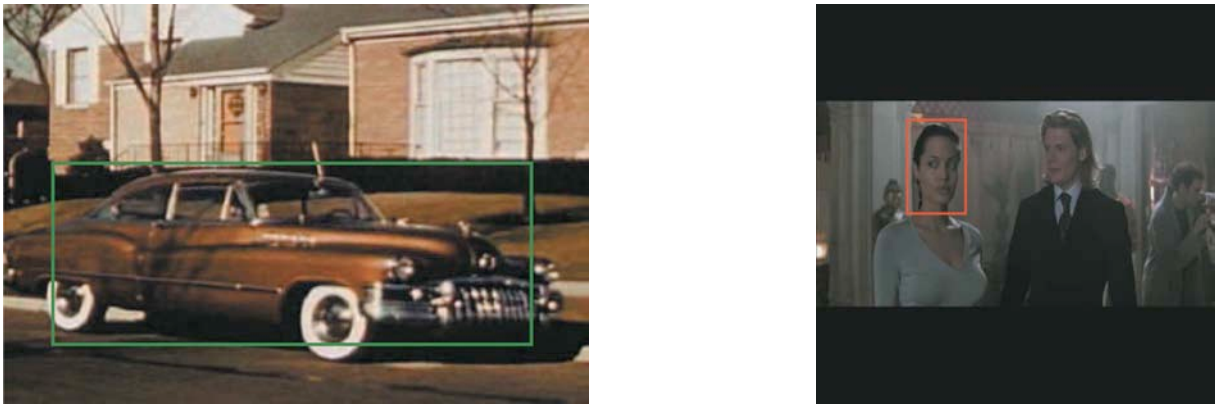

FIG. 2.6 – Identification de descripteurs moyens (voiture, visage) sur deux images.

Descripteurs image de niveau sémantique haut.

Le niveau sémantique haut correspond à des concepts plus complexes que l'on extrait d'une image. Ces concepts sont déduits par classification d'une zone (ou de l'intégralité) de l'image dans des classes prédéfinies. Plusieurs descripteurs de ce type ont été développés dans la littérature. Il s'agit principalement de la reconnaissance du lieu où une photo a été prise et de l'identification d'une personne à partir de son visage. Les concepts issus de la classification d'ambiance sont de granularité globale. Les concepts issus de la reconnaissance de personnages sont de granularité intermédiaire. Ces tâches de classification nécessitent le calcul de modèles numériques complexes appris à partir d'images de référence.

Dans le cadre de cette étude, la **reconnaissance de lieu** est employée pour caractériser des images. Cette classification est représentée par une ontologie hiérarchique de concepts comprenant plusieurs niveaux de précision définis par l'administrateur du système. Une première classification est faite pour dissocier les images d'intérieur et d'extérieur. Ensuite, les images d'extérieur sont classées en ville et paysage. De même les images d'intérieurs sont classées en *parking/magasin/maison*. Remarquons que cette taxonomie n'est pas représentative du monde réel mais correspond à la base de données dont nous disposons. Dans notre cas, le descripteur visuel TextureLEP est utilisé pour représenter l'information de niveau bas. A travers l'analyse de zones d'images, on peut aussi influencer la localisation par la présence de concepts moyens comme des textures (herbe, ciel) ou des objets particuliers (visages, voitures). Nous développerons particulièrement cette tâche par la suite.

2.1.6 Descripteurs du signal vidéo

Descripteurs numériques vidéo

Les descripteurs numériques de la vidéo correspondent généralement à des interprétations en termes de couleur, de texture et de forme. Ces informations résultent de l'analyse de chaque image ou de segments d'images de la vidéo.

Extraire les descripteurs de chacune des images d'un document vidéo rendrait le temps de traitement prohibitif. D'autre part, il faut prendre en compte le fait que deux images consécutives dans une vidéo sont assez semblables, et, si ce n'est pas le cas, cette différence est porteuse d'information au niveau structurel, puisqu'elle peut correspondre à un changement de plan. Cette observation conduit donc à favoriser un traitement de segments d'images, plutôt qu'un traitement image par image. De nombreuses approches réduisent ainsi le problème de l'extraction du contenu d'un segment d'images au traitement des descripteurs d'une seule image du segment considéré. Plusieurs algorithmes ont été développés afin de déterminer l'**image « moyenne »** ou résumé d'une séquence d'image. C'est ce qui est fait dans [Sun00b], pour segmenter la vidéo en scènes. Il est aussi possible d'extraire des descripteurs bas à partir de **statistiques** (e.g. : moyenne, écart type) des caractéristiques des images de la séquence. C'est ce qui est fait, par exemple, dans [Ass98], pour déterminer le taux de couleurs saturées dans les publicités. Dans le cadre de notre application, nous utiliserons les deux techniques : la première pour les segments de type plans, granularité temporelle locale. Elle est peu gourmande en temps de calcul, de plus nous disposons, au CEA, d'un algorithme de segmentation temporelle de la vidéo [Jos00] qui extrait l'image « moyenne » de chaque plan ; la seconde pour les segments composés de plusieurs plans, granularité temporelle intermédiaire et globale. Les descripteurs sont calculés à partir de statistiques extraites des caractéristiques des images « moyennes » des plans du segment.

D'autres approches proposent de définir et d'utiliser le mouvement d'objets visibles mais aussi les mouvements de la caméra pour l'indexation de la vidéo. Dans [Mac03] les auteurs décrivent un système de segmentation des images d'une vidéo en objets aux déplacements indépendants. La méthode commence par une segmentation basée sur la couleur de l'image. Puis les régions sont regroupées selon leurs paramètres de mouvement. Là encore, MPEG-7 intègre la notion de trajectoire soit à partir de la

donnée de points clefs et de techniques d'interpolation soit à partir des vecteurs de mouvements utilisés déjà par MPEG-1 et MPEG-2. Dans le système VideoQ [Cha98], les auteurs décrivent diverses façons d'extraire des mouvements d'objets de films vidéo et de formuler et traiter des requêtes portant sur ces mouvements. Dans le domaine des bases de données, [Li97] c'est également intéressé à la modélisation du mouvement des objets afin de permettre d'interroger rapidement un modèle orienté objet des parties de vidéos, à partir de requêtes SQL.

Dans le cadre de notre application nous n'utiliserons pas de suivi de trajectoire ou de détection de mouvement, mais ces techniques sont aisément intégrables à notre système.

Les descripteurs numériques du **son** ont été décrits au paragraphe précédent. Pour la vidéo ils sont extraits à partir de segments de la bande son. Ces segments peuvent être de taille fixe ou variable, lorsqu'ils sont issus d'une segmentation.

Descripteurs vidéo de niveau sémantique moyen.

La description sémantique de niveau moyen d'une vidéo s'appuie, comme pour l'audio et l'image, sur la notion de concept. Un concept représente une description symbolique de la vidéo ou d'une partie de la vidéo.

Dans certains domaines ciblés tels que le sport [Bab99] ou les journaux d'informations [Zha95], il est possible de réaliser une extraction automatique. Mais la plupart des systèmes demandent une annotation manuelle par l'utilisateur [Kan00]. Le système vidéo développé par IBM [Ada02] utilise un lexique varié pour représenter des concepts moyens audio (*musique/dialogue/monologue*) et image (Objets : *visage, voiture, bâtiment, pont...* , Textures : *ciel, eau, désert, neige,* ...). Cependant, aujourd'hui, l'amélioration des performances de l'indexation audio et image autorise la détermination automatique d'un grand nombre de concepts simples.

Plusieurs descripteurs de niveau moyen peuvent être extraits de façon automatique à partir d'une **image** « moyenne » d'un segment de vidéo. Ainsi les valeurs des concepts déterminées à partir de l'image la plus « représentative » sont associées au segment considéré. Les concepts les plus utilisés se rapportent aux classifications d'images présentées ultérieurement :

- Objets : *voiture, télévision, lumière artificielle, visage* [Sch00c, Vio02].
- Textures : *herbe, arbre, eau, ciel, bâtiment, sable, neige*. [Mil04]

En ce qui concerne les segments composés de plusieurs plans, granularité temporelle intermédiaire et globale, les descripteurs sont représentés par un modèle de strate (voir l'annexe A.1) appliqué au plan de la vidéo.

Pour le **son**, les concepts moyens de granularité locale sont extraits à partir de segments d'audio de taille variable (3s à l'infini). Plusieurs des classifications présentées précédemment (e.g. *parole/musique/bruit*) ont été adaptées à l'indexation automatique de vidéo.

Ces dernières années, l'étude de plusieurs descripteurs moyens **multimedia** a montré l'efficacité de la fusion de l'information provenant de plusieurs media pour la classification. Dans le cadre de la détection de personnes, le plus souvent, seules les caractéristiques visuelles sont utilisées. Cependant, la modalité auditive peut aussi fournir un indicateur de la présence d'un individu dans le plan : la classification du signal audio en parole est la première étape, avant une détection de visage. Ces techniques obtiennent un bon taux de reconnaissance d'environ 95% [Ben98]. Ce qui correspond à une augmentation de 5% par rapport aux techniques uniquement visuelles. De même, pour la détection d'objets, le son produit par un objet spécifique (ou signature) peut être détecté par l'analyse du signal audio segmenté d'une vidéo. Les segments audio classés en sons environnementaux peuvent être analysés à la recherche d'objets. Dans [Fre98] plusieurs signatures audio spécifiques sont détectées incluant les aboiements de chiens, les sonneries de téléphone et les instruments de musique. La présence d'un de ces sons fournit une forte indication sur la présence de l'objet associé.

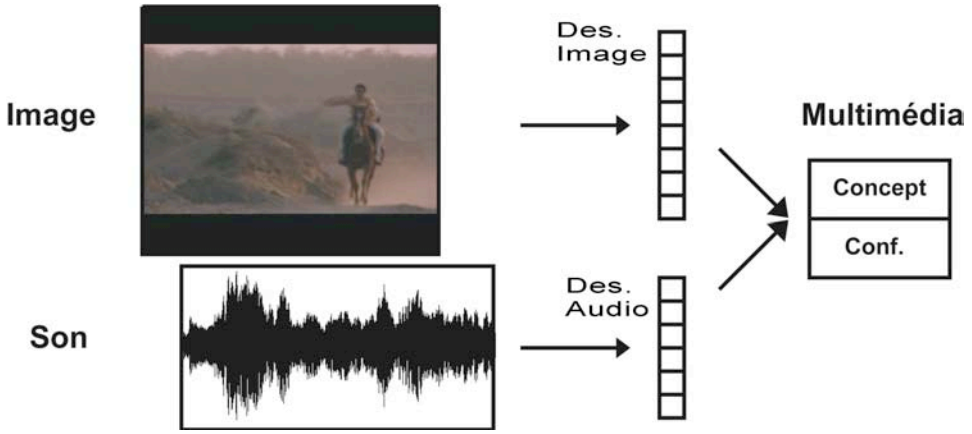

FIG. 2.7 – Concept multimedia extrait d'un segment de vidéo.

Descripteurs vidéo de niveau sémantique haut

Le niveau sémantique haut correspond à des concepts complexes que l'on extrait de la vidéo. Comme nous l'avons vu, la détermination de concepts est améliorée par la fusion de plusieurs media. Ceci est d'autant plus vrai dans le cas de concepts hauts, dont la classification est souvent ardue. Dans le cadre de notre application, les descripteurs de niveau haut et de granularité locale sont extraits (comme au niveau moyen) à partir d'une image « moyenne » du plan (pour l'image) et du segment de son synchrone. En ce qui concerne les descripteurs de niveau haut et de granularité supérieure (séquence de plans) nous utilisons un modèle de strates similaire à celui du niveau moyen.

La reconnaissance de personnes est un excellent exemple de la nécessité de l'intégration multimodale. Identifier des personnages uniquement par le signal visuel n'est pas efficace à cause de la variance forte de l'orientation, l'illumination et l'occultation des visages au sein d'une scène. De plus les détections de parole et l'identification du locuteur sont très sensibles aux bruits environnementaux. La combinaison de ces deux types d'information peut être utilisée pour améliorer les performances générales de la détection de personnes.

De même, plusieurs concepts liés à la **reconnaissance de lieu** ou localisation (*intérieur/extérieur*, *parking/magasin/maison*, *ville/paysage*) sont extraits pour les modalités visuelles et auditives. Le challenge, pour améliorer la localisation pour la vidéo, est donc de réaliser la fusion de l'information apportée par les deux media, visuel et auditif, afin que les deux collaborent à l'extraction des concepts considérés.

2.2 La classification

Classer, c'est regrouper entre eux des objets similaires, selon un critère fixé par l'utilisateur. Ce critère peut être simple ou multiple, numérique ou catégoriel, brut ou transformé : les choix possibles sont pratiquement illimités. Par construction, les objets réunis tendent à former des classes homogènes selon ce critère.

Dans le cadre de cette thèse, les critères choisis pour la classification seront sous la forme de concepts. Ainsi, deux objets appartiennent à la même classe s'ils ont la même valeur pour le concept choisi. La classification permet de déduire la valeur prise par un concept d'un nouvel objet en déterminant la classe à laquelle il appartient. Un modèle de classification comprend d'une part une taxonomie, c'est-à-dire l'ensemble des valeurs possibles pour le concept de classification et d'autre part un modèle mathématique définissant les relations entre les objets classés. Il existe de nombreux modèles de classification, et l'on peut les différencier suivant divers critères : la connaissance a priori ou non des classes de la taxo-

nomie (classification supervisée/non supervisée) ; le modèle mathématique de regroupement des objets ; le modèle de taxonomie.

Dans ce paragraphe nous définirons chacun de ces critères. Nous appelons :

Π la population d'objets et O un objet,

Δ l'ensemble des descriptions et D une description,

C un concept,

$\mathcal{M}_c(\mathrm{T}, \mathrm{X}, \Psi)$ le modèle de classification associé où

$\mathrm{T} = \{c_1, \ldots, c_k, \ldots, c_p\}$ est la taxonomie du concept de classification et c_k la $k^{ème}$ classe de la taxonomie,

$\mathrm{X} : \Pi \rightarrow \Delta$ est la fonction qui associe une description à tout élément de la population et

$\Psi : \Delta \rightarrow \mathrm{T}$ est la fonction de classement qui associe une classe à tout élément de la population.

2.2.1 Classification supervisée/non supervisée.

Il existe deux approches distinctes de la classification de données, l'approche supervisée et l'approche non-supervisée. L'approche supervisée suppose la connaissance a priori de la taxonomie du concept choisi. Alors que l'approche non-supervisée ne nécessite aucun apprentissage préalable.

Classification non-supervisée

Les algorithmes non-supervisés ou "clustering" visent à répartir des objets sur chacun desquels on a déterminé m descripteurs en un certain nombre p de classes aussi homogènes que possible. Le choix de p résulte d'un compromis à trouver entre une description qui soit à la fois suffisamment simple (p pas trop grand) et suffisamment détaillée (p pas trop petit). p est d'abord estimé par l'algorithme, puis les classes sont identifiées au sein des données. La classification non-supervisée ne comporte pas d'hypothèse sur la structure des classes. Une des qualités de ces techniques est qu'elles sont relativement objectives. Mais les classes ainsi formées ne correspondent pas forcément à des caractéristiques sémantiques précises ou que l'on recherche. De fait cette technique est justifiée ou non par son aptitude à produire des classifications qui font du sens.

La classification non-supervisée d'objets permet donc de déterminer la valeur de concept d'objet, pour des concepts à valeurs entières $\mathrm{T} = \{1..p\}$, **sans connaissance a priori sur les classes**. Cette particularité est utilisée, par exemple, dans les systèmes de segmentation en locuteur du signal audio. Elle permet de segmenter le son d'une conversation entre plusieurs personnes et d'annoter les segments ainsi extraits d'un entier (1, 2, 3) correspondant à son émetteur. Cela sans connaissance a priori du nombre de personnes présentes pendant l'enregistrement. Cette propriété sera utile pour notre système dans le cadre de la segmentation temporelle de films.

Classification supervisée

Le but premier de notre système est de déterminer des descripteurs de niveau sémantique moyen et haut. Ces descripteurs se présentent sous la forme de concepts textuels. Comme on l'a vu, le principe est d'associer aux objets le nom de la classe à laquelle ils appartiennent. Pour cela, une classification de type supervisée est pratiquée.

Les algorithmes supervisés prennent en compte une **connaissance a priori des classes** que les données sont censées contenir (par hypothèse). Cette connaissance se compose de deux parties :

- la taxonomie de classification T, i.e. le nom de toutes les classes,
- une base d'objets de référence annotée selon la taxonomie choisie.

Cette base annotée permet d'apprendre un modèle mathématique du concept. Ce modèle définit les liens entre les objets sémantiques de la taxonomie (les classes) et les objets issus du traitement du signal (descripteurs numériques). Il met en évidence les relations communes entre objets de référence d'une même classe et de classes différentes.

Classification	Concept	Taxonomie
Supervisée	intérieur	oui/non
Non-supervisée	locuteur	1/2/3

TAB. 2.1 – Exemples de modèles de classification supervisée et non-supervisée

Il existe deux types principaux de modèles des relations entre objets à classer. Ils se distinguent principalement par le choix du cadre mathématique dans lequel les descripteurs des nouveaux objets sont placés : cadre probabiliste, cadre spatial.

2.2.2 Modèles probabilistes de classification

Dans le cadre probabiliste, les descripteurs et les concepts sont modélisés par des **variables aléatoires**. Les modèles de classification projettent les classes de la taxonomie sur des probabilités d'occurrence de ces variables. Les probabilités sont évaluées à partir des distributions des descripteurs de la base d'apprentissage. Ce modèle d'apprentissage permet de déterminer la probabilité qu'un nouvel objet appartienne à l'une des classes de la taxonomie et ensuite de déterminer la valeur la plus probable du concept pour ce nouvel objet.

Fonction de classement

Soit un objet de test O. L'extraction des descripteurs numériques fournit la description D pour cet objet. Le but de la classification est de déterminer la classe $\Psi(D)$ à laquelle il appartient. Pour cela la fonction de classement Ψ, doit être définie.

Une première règle possible pour le choix de la fonction de classement pourrait être : " attribuer à chaque description la classe majoritaire ", c'est-à-dire celle pour laquelle la probabilité marginale $\mathrm{P}(c_k)$ est maximum ; c'est la règle majoritaire. Cette règle ne prend pas en compte la nature de l'objet considéré.

Une seconde règle consiste à raisonner ainsi : " si j'observe D, je choisis la classe pour laquelle cette observation est la plus probable". C'est-à-dire, celle qui maximise la probabilité qu'un objet appartenant à la classe c_k ait D pour description définie par la probabilité conditionnelle $\mathrm{P}(D|c_k)$. C'est la règle du maximum de vraisemblance. Cette règle ne prend pas en compte la probabilité d'occurrence de la classe c_k, on lui préférera donc la troisième règle.

La **règle de Bayes** (ou règle du maximum a posteriori) consiste à attribuer à une description D la classe c_k qui maximise la probabilité qu'un élément ayant D pour description, soit de classe c_k, définie par la probabilité conditionnelles $\mathrm{P}(c_k|D)$. En pratique, cette probabilité peut être estimée en utilisant la formule de Bayes :

$$\mathrm{P}(c_k|D) = \frac{\mathrm{P}(c_k)\mathrm{P}(D|c_k)}{\mathrm{P}(D)}$$

$\mathrm{P}(D)$ est une constante. Ainsi, la règle de Bayes est fondée sur le calcul des probabilités conditionnelles $\mathrm{P}(D|c_k)$ et marginales $\mathrm{P}(c_k)$ estimées à partir de la distribution des descripteurs au sein de chaque classe et du concept de classification.

Règle de Bayes

Les quantités $\mathrm{P}(D|c_k)$ et $\mathrm{P}(c_k)$ utilisées dans la règle de Bayes ne sont pas connues. Leur estimation est faite par apprentissage sur la base de données annotée selon la taxonomie de notre classification.

Supposons que Δ soit un espace de dimension m, et que nous disposions d'un ensemble d'objets de référence Σ. Soit O un objet de la population. La description $D = \{D_1, \ldots, D_j, \ldots, D_m\}$ a été déterminée par le traitement du signal de O. Nous souhaitons classer cet élément de Δ par la fonction de classement. La règle de classification de Bayes s'écrit donc :

$$\Psi_{\text{bayes}}(D) = \underset{k=1..p}{\operatorname{argmax}} \left(\mathrm{P}(D = \{D_1, \ldots, D_m\}|c_k)\ \mathrm{P}(c_k)\right)$$

La relation de dépendance qui existe entre l'ensemble de description D et le concept C peut être représentée par le réseau bayésien représenté en figure 2.8.

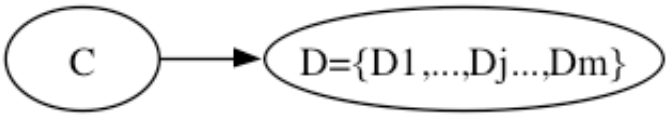

FIG. 2.8 – Représentation de la dépendance entre un concept et une description associée.

Pour rendre la méthode effective, le principe est de remplacer $\mathrm{P}(D|c_k)$ et $\mathrm{P}(c_k)$ par des estimations faites sur l'échantillon Σ.

Estimation de la probabilité $\mathrm{P}(\mathbf{c_k})$

L'estimation de la probabilité d'occurrence de la classe c_k se fait par méthode simple de comptage. Pour toute classe c_k, $\mathrm{P}(c_k)$ est estimée par la proportion d'éléments de classe c_k dans la base de référence Σ.

Estimation de la probabilité conditionnelle $\mathrm{P}(\mathbf{D}|\mathbf{c_k})$

Par contre, l'estimation de la probabilité conditionnelle $\mathrm{P}(D|c_k)$ est plus complexe, car le nombre de descriptions possibles est a priori infini, et par comptage, il faudrait un échantillon Σ de taille trop importante pour pouvoir déterminer convenablement ces quantités. Il faut donc déterminer un modèle mathématique de la probabilité $\mathrm{P}(D|c_k)$.

Certains font l'hypothèse simplificatrice suivante : les valeurs des attributs sont indépendants connaissant la classe. Cette hypothèse permet d'utiliser l'égalité suivante :

$$\mathrm{P}(D = \{D_1, \ldots, D_m\}|c_k) = \prod_{j=1}^{m} \mathrm{P}(D_j|c_k)$$

La relation de dépendance qui existe entre l'ensemble des descripteurs et le concept C peut donc être représentée par le réseau bayésien représenté en figure 2.9. Maintenant, il suffit de calculer, pour tout j et toute classe c_k, $\mathrm{P}(D_j|c_k)$, l'estimation de la probabilité qu'un élément de classe c_k ait D_j pour $j^{ème}$ attribut. Cette technique est appelée **classification naïve de Bayes**.

Dans la littérature, de nombreuses approches ont été présentées afin modéliser les probabilités conditionnelle $\mathrm{P}(D_j|c_k)$ à partir de l'échantillon Σ. La plus simple est l'estimation par histogramme continu [Che03b].

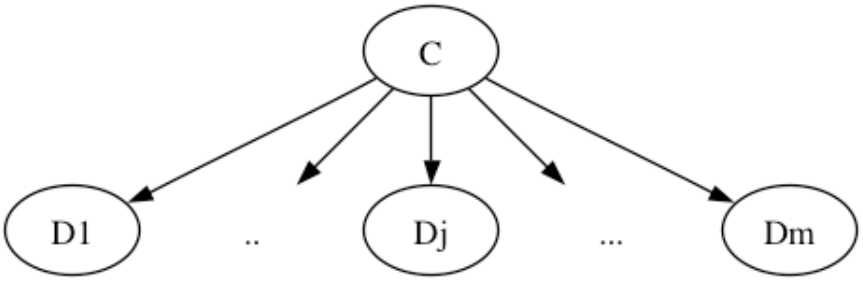

FIG. 2.9 – Représentation de la dépendance naïve entre un concept et une description associée.

Finalement, la règle de classification naïve de Bayes s'écrit donc :

$$\Psi_{\text{Nbayes}}(D) = \underset{k=1..p}{\operatorname{argmax}} \left(\mathrm{P}(c_k) \prod_{j=1}^{m} \mathrm{P}(D_j|c_k) \right)$$

Cette méthode est simple à mettre en œuvre, bien qu'elle soit basée sur une hypothèse fausse en général : les attributs sont rarement indépendants. Elle donne cependant de bons résultats dans les problèmes réels et fournit un seuil de performance pour d'autres méthodes.

Notre but, dans le cadre de la fusion des descripteurs est de prendre en compte les relations d'inter-dépendance (ou corrélation) entre descripteurs. Il est donc nécessaire d'évaluer la probabilité jointe des descripteurs numériques, sachant la classe, sans faire l'hypothèse d'indépendance.

Modeler des données multidimensionnelles en utilisant des distributions gaussiennes et des mélanges de gaussiennes est très courant [Mar98, MC99, Foo99]. Cette popularité tient au fait que les distributions peuvent être approximées par ce types de fonctions, mais surtout parce qu'il existe de nombreux résultats et techniques de calculs mathématiques fondés sur des distributions gaussiennes. Les procédés gaussiens donnent une formulation naturelle de l'apprentissage en terme de distributions conditionnelles.

Plusieurs modèles de distributions des descripteurs au sein des classes ont été envisagés. Le modèle multidimensionnel gaussien **MAP** [Wol96, Mac97] travaille en modélisant les données d'une classe par un cluster gaussien de points dans l'espace des descripteurs. Pendant la phase d'apprentissage, les paramètres (moyennes et covariances) sont estimés pour chaque classe par un algorithme EM [Bim98]. Ces paramètres permettent d'estimer la probabilité jointe globale $\mathrm{P}(D|c_k)$.

Le modèle de mélange de gaussiennes (**GMM**) [Sch97, Sla02] représente chaque classe comme l'union de plusieurs clusters gaussiens dans l'espace des descripteurs. Cette caractéristique peut s'avérer essentielle lorsque les classes présentent des structures complexes, plusieurs zones disjointes par exemple [Sla02]. Les paramètres des gaussiennes sont estimés directement à partir de l'algorithme de maximisation de l'espérance (Algorithme EM) [Bim98]. En comparaison avec le MAP, les clusters individuels ne sont pas représentés par la matrice de covariance, mais seulement par une approximation diagonale. C'est-à-dire que les clusters gaussiens ainsi obtenus ont leurs axes parallèles aux axes de l'espace des descripteurs.

Ce type de modèles de probabilité présentent des performances de classification satisfaisantes bien qu'ils ne représentent pas bien les relations qui existent entre les descripteurs. En effet, ils ne prennent pas en compte la structure topographique complexe des classes.

2.2.3 Modèles spatiaux de classification

Les modèles spatiaux permettent de mieux prendre en compte la non linéarité des structures de données. Dans le cadre spatial, les objets sont modélisés par des points dans l'espace des descripteurs. Ainsi

pour déterminer la classe d'un point O de la population, sa position dans l'espace D est étudiée par rapport aux points de référence. Il existe deux types de fonctions de classement spatial :

- Les fonctions globales : l'espace des données est découpé en plusieurs zones à chacune desquelles est associée une des classes. Ces zones sont apprises à partir de la base d'apprentissage Σ. Un objet appartient à la classe associée à la zone dans laquelle il se trouve.
- Les fonctions locales : à un objet est attribué la classe « moyenne » des points de son voisinage.

Classification globale

Soit O un objet de la population Π, soit D sa description. La classification globale nécessite l'apprentissage sur la base Σ du découpage de l'espace et de la fonction qui associe une classe à chaque zone de l'espace. Il est nécessaire de définir un modèle de découpage à la fois simple, pour éviter des calculs trop importants et complexes pour conserver la structure spatiale des classes.

De nombreux modèles de découpage ont été présentés dans la littérature. Les modélisations les plus anciennes consistent à apprendre les frontières entre classes sur la base Σ. Ce sont principalement les **«k-d tree»** et affiliés [Cha99]. L'arbre de partitionnement spatial k-d trace des frontières de décisions arbitraires ou segments de «Manhattan», dont la complexité dépend du nombre de données d'apprentissage et du nombre de limites.

Dernièrement, de nouvelles techniques de partitionnement ont été explorées dont les **Support Vector Machine** [Bur98] (voir l'annexe A.2). Les SVM sont des techniques de classification globale à deux classes. Elles permettent l'extraction de concepts binaires (0/1) comme "*parole/non-parole*" ou "*intérieur/extérieur*". Elles sont basées sur le théorème d'Hadamard et consistent à projeter les données dans un nouvel espace des descripteurs de dimension supérieure dans lequel il existe un hyperplan qui sépare les objets appartenant aux deux classes. Le taux de confiance de la classification est estimée par la distance d'un point à ce plan de séparation. Cette technique a été appliquée à de nombreux domaines de la classification. Dans le cadre du traitement de l'audio, pour la reconnaissance d'instruments de musique [Mar99], ou la classification *parole/musique* [Li00, Mor00]. Et pour l'image : la classification d'image [Mil04] ou la reconnaissance d'un visage [Goh01].

Une limitation des SVM est qu'ils ne s'appliquent pas aux classifications pour les concepts multiclasses. Pour détourner le problème certains définissent une hiérarchie dans la taxonomie du concept, afin de réaliser des classifications binaires successives. Mais souvent les taxonomies ne peuvent pas se hiérarchiser, comme par exemple dans la reconnaissance de visage. Dans [Goh01], de nombreuses méthodes sont présentées pour effectivement prendre en compte la classification multiclasse. Elles comprennent l'**OPC** (one per class). Pour chaque classe de la taxonomie, le modèle de classification du concept classe k présent ou absent est appris. Pour classer un objet, la classe attribuée est celle dont la classification montre le meilleur taux de confiance. Une autre méthode utilise un couplage par paire : PWC (pairwise coupling). Pour chaque couple de classes $\{c_{k1}, c_{k2}\}$ de la taxonomie, les classifications des concepts "O appartient à la classe c_{k1} ou c_{k2}" sont appris. La combinaison des taux de confiance fournit un taux de confiance global pour chaque classe. Celle qui montre le meilleur taux est attribuée à l'objet.

Ces techniques permettent de réaliser des classifications d'objets pour des concepts multiclasse. Elles fournissent des performances équivalentes et souvent supérieures à celles observées pour les modèles statistiques de classification (MAP et GMM).

Classification locale

Dans le cadre d'une classification locale, les classes des objets de Σ qui se trouvent dans le **voisinage** d'un point Q en question (en entrée du système) sont étudiées. On en déduit l'appartenance de cet objet à l'une des classes de la taxonomie T.

Afin de définir le terme de "voisinage", il est nécessaire d'introduire ici une distance de l'espace Δ des descripteurs. Il existe de nombreuses distances dans la littérature. En effet, la recherche d'objets (son, image…) par similarité, est un domaine vaste du traitement de l'information [Cha99]. Dans ces systèmes, les documents sont représentés par des descripteurs comme pour la classification. Et le traitement consiste à trouver le ou les documents les plus proches du document question dans l'espace des descripteurs. [Kis96] et [Zha03] définissent et comparent plusieurs distances de la littérature pour la recherche d'images. Ils montrent que la plupart d'entre elles fournissent des résultats équivalents dans ce cas. Nous choisirons donc pour cette thèse une distance simple : la distance euclidienne. Trois types de classification sont envisageables au niveau local dans un espace métrique :

1. Le « **range query** » (Q, r) est utilisé par [Cha99] pour la recherche d'image. Cet algorithme retrouve tous les éléments qui se trouvent à une distance inférieure à r de l'objet question Q. La classe de cet objet est ensuite déterminée par vote à la majorité sur les objets du voisinage. Cette technique est mal adaptée dans notre cas, car le choix de r est fortement dépendant de la distribution des descripteurs au sein des classes. Il faut donc le fixer pour chaque concept, ce qui n'est pas pratique quand il y en a beaucoup.
2. Le **plus proche voisin** (NN ou PPV) retrouve l'élément O le plus proche de l'objet Q. Et la classe de O est attribuée à Q. Cette technique est intéressante mais trop sensible au bruit.
3. La technique des **plus proches voisins** (KNN ou KPPV) montrée en figure 2.10 présente de bonnes aptitudes pour la tâche de classification du contenu : dans le cadre de l'étude du son, pour la taxonomie *musique/parole* [Sch97] ; pour la classification d'image [Mil04], notamment en *intérieur/extérieur* [Pat96]. C'est une technique très populaire. Les k plus proches voisins de l'objet Q sont isolés. La classe de cet objet est ensuite déterminée par vote à la majorité sur les k objets voisins.

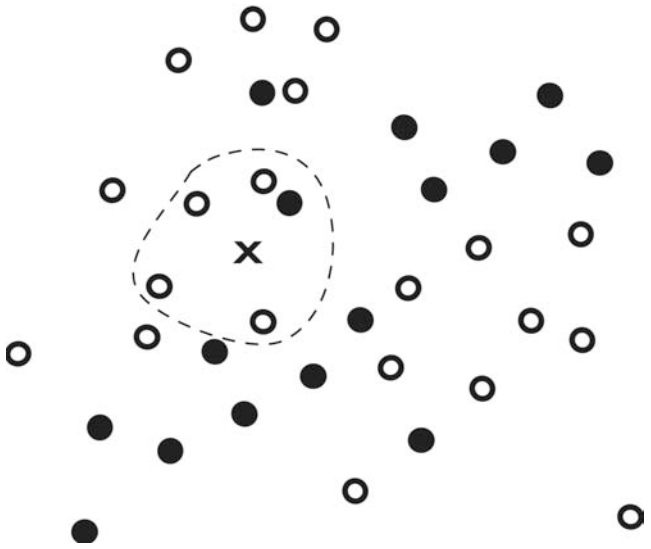

FIG. 2.10 – Modèle de classification KNN.

Remarquons que [Mil04] obtient avec les SVM de meilleurs résultats de classification d'image en *intérieur/extérieur* qu'avec des techniques locales (plus proches voisins). Par contre dans le cas de classification à plusieurs classes les plus proches voisins semblent donner de meilleurs résultats.

Dans la suite, différents modèles de classification (GMM, KNN et SVM) seront testés, la comparaison des performances obtenues nous permettra de choisir un de ces modèles pour les expériences de la thèse.

2.2.4 Taxonomie de classification

La taxonomie d'une classification est l'ensemble des valeurs possibles pour le concept associé à cette classification. Pour simplifier, une taxonomie est un ensemble de classes auxquelles les objets peuvent appartenir ou pas. On distingue deux grandes familles de taxonomies suivant la forme des classes considérées.

Taxonomie non hiérarchique

Les taxonomies non hiérarchiques ou **partitionnement** aboutissent à la décomposition de l'ensemble de tous les individus en k ensembles disjoints ou classes d'équivalence. La plupart des classifications de la littérature sont de ce type ($\mathrm{T} = \{$*intérieur, extérieur*$\}$, $\mathrm{T} = \{$*parole,musique,bruit*$\}$). Cependant, ces dernières années, d'autres types de structure de taxonomie ont été développés afin de prendre en compte les relations hiérarchiques entre classes.

Taxonomie hiérarchique

Pour un seuil discriminant ou **niveau de précision** donné, deux individus peuvent appartenir à une même classe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront donc à deux sous-classes différentes. En faisant varier le niveau précision, les n individus au départ distincts ($k = n$) se trouveront ultimement rassemblés dans un groupe unique ($k = 1$). Le dendrogramme constitue une représentation graphique de ce processus d'agrégation. Remarquons que chaque noeud du dendrogramme est un partitionnement de la classe parente.

Ce type de classification est naturelle puisque la perception humaine est fondée en partie sur une organisation hiérarchique. La plus évidente est zoologique : « Spot » est un dalmatien qui est une sorte de chien, qui est une sorte d'animal. Afin de reconnaître un objet dans le monde, nous reconnaissons d'abord un niveau intermédiaire d'abstraction de cet objet, que Rosch a appelé niveau de base [Ros76]. Après cette reconnaissance initiale, le processus procède par classifications successives vers des niveaux plus ou moins abstraits suivant la direction (haut/bas). Dans cet exemple, nous voyons que Spot est un chien avant d'identifier sa race ou son identité particulière ; le concept de classification a pour valeur $C =$ *animal/chien/dalmatien/Spot*. Cette manière de procéder à un avantage computationnel fort par rapport à la classification directe en bas de la taxonomie. En effet, il est souvent plus simple de faire des distinctions à différents niveaux de la taxonomie : à des niveaux plus abstraits, il y a moins de classes, et le processus de décision peut être simplifié. Le plus souvent, la reconnaissance se fait donc vers le bas des branches, chaque décision est relativement simple, elle ne prend en compte que quelques descripteurs du signal pour distinguer entre le petit nombre de classes à chaque nœud [McL04]. La plupart des standards de classification de produits ou de services sont basés sur des schémas hiérarchiques. Par exemple, les données bibliographiques [Gru92], les systèmes de fichier, Web open directories (Google et Yahoo) et la plupart des catalogues électroniques [War99] sont des classifications hiérarchiques. Cependant la classification par concepts supervisés n'est pas largement répandue dans de tels cas réels.

Application à l'indexation automatique du contenu

Dans le cadre d'une indexation du contenu, il semble donc approprié de choisir des taxonomies de classification de type hiérarchique. Pour une meilleure représentation, nous considérons que le concept global C de la taxonomie est un ensemble de concepts structurés hiérarchiquement appelés « concepts d'articulation ». Nous notons $C = \{C_1, \ldots, C_j, \ldots, C_m\}$. A un nœud correspond un concept auquel est associé un groupe de concepts représentés par des nœuds fils. A chaque concept correspond un modèle de classification non hiérarchique : l'ensemble des valeurs prises par ce concept ainsi que les objets annotés

et leur descripteurs du signal. Pour la description automatique du contenu par concepts, de nombreux systèmes de classification hiérarchique existent.

En ce qui concerne l'**audio**, [McL04] utilise une taxonomie de ce type pour la reconnaissance d'instruments, présentée dans la figure 2.11 (en haut). Ils supposent que cette classification se fait de façon hiérarchique : avant de reconnaître un instrument en particulier, sa famille d'instruments est reconnue. Ce principe correspond bien à notre expérience subjective. En effet, il est souvent plus facile de reconnaître si c'est un instrument à cordes qui a produit un son particulier, que de dire si l'instrument est un violon ou un alto. La classification traditionnelle des instruments est basée sur la forme, le matériau et l'histoire de l'instrument, plus que sur le son qu'ils produisent. Cependant les instruments d'une même famille ont beaucoup de propriétés acoustiques en commun. De façon générale, pour la classification automatique, il est important que la taxonomie choisie soit en relation avec les paramètres physiques des objets considérés. Ceci pour le son mais aussi pour les autres media. Par exemple, les instruments à cordes ont des propriétés de résonance complexes et de très longues attaques ; les cuivres ont une seule résonance et des attaques plus courtes. De nombreuses autres classifications de l'audio sont représentées par des modèles hiérarchiques, notamment la classification hiérarchique "*parole/musique*" présentée en figure 2.11 (à droite). Plusieurs systèmes de traitements de données audio ou vidéo extraient du son cette classification, [Zha98] notamment. Etant donnée la différence d'origine physique des classes basiques de l'audio, les auteurs montrent qu'il est préférable, pour une classification plus fine, d'adopter des approches différentes pour chaque partitionnement ; c'est-à-dire des descripteurs et des algorithmes différents.

En ce qui concerne l'image, les auteurs de [Vai99] ont mis en oeuvre une classification hiérarchique de photos : ils séparent d'abord les photos d'intérieur et les photos d'extérieur, puis classifient les photos d'extérieur en photos de ville, et de paysage. Enfin, les photos de paysage sont encore classifiées suivant qu'il s'agit plutôt d'un couché de soleil, d'une forêt ou d'une montagne. L'inconvénient de cette méthode est qu'elle ne peut pas classer correctement une photo comprenant deux fonds, par exemple une forêt et une montagne. Au CEA, C.Millet [Mil04] a développé une classification hiérarchique de ce type pour des régions d'images photographiques, présentée dans la figure 2.12.

Plusieurs études ont montré l'avantage en temps de calcul et en pertinence des résultats des modèles hiérarchiques [Ala98, McL04]. Remarquons que les classifications hiérarchiques effectuées du haut vers le bas ont un défaut majeur. Si le système commet une erreur à un nœud du système lors de la classification, les partitionnements de rangs inférieurs sont ensuite inadaptés et aboutissent à une classification fausse à ces niveaux. Cependant, les auteurs de [Dum00] ont montré sur des classifications de pages Web par un algorithme naïf de Bayes que pour un petit nombre de descripteurs par classification (une dizaine), la classification hiérarchique donne de meilleurs résultats qu'un seul partitionnement. Par contre, au delà, ce qui est souvent le cas, il n'y a pas d'amélioration des performances.

Dans notre système, nous définirons plusieurs modèles de taxonomie hiérarchique. Une pour l'audio, inspirée de la classification *parole/musique* de [Zha98], et l'autre pour la classification d'ambiance inspirée par [Ala98]. De plus, nous verrons que le modèle probabiliste présenté dans cette étude permet d'estimer l'ensemble des concepts de la taxonomie en même temps, et ainsi de résoudre le problème cité au-dessus.

Taxonomies complexes

Il existe d'autres types de taxonomie plus complexes. Les classifications hiérarchiques sont des schémas en arbre ; ainsi ils représentent un modèle très large de données puisque de nombreux schémas peuvent facilement être convertis sur ce modèle. Par contre, ce type de représentation ne peut pas présenter les modèles les plus généraux employés dans les cas réels. Il sont incapables de partager des sous-structures. De plus, ils ont des contraintes référentielles et il est impossible de réduire de telles propriétés en une structure en arbre.

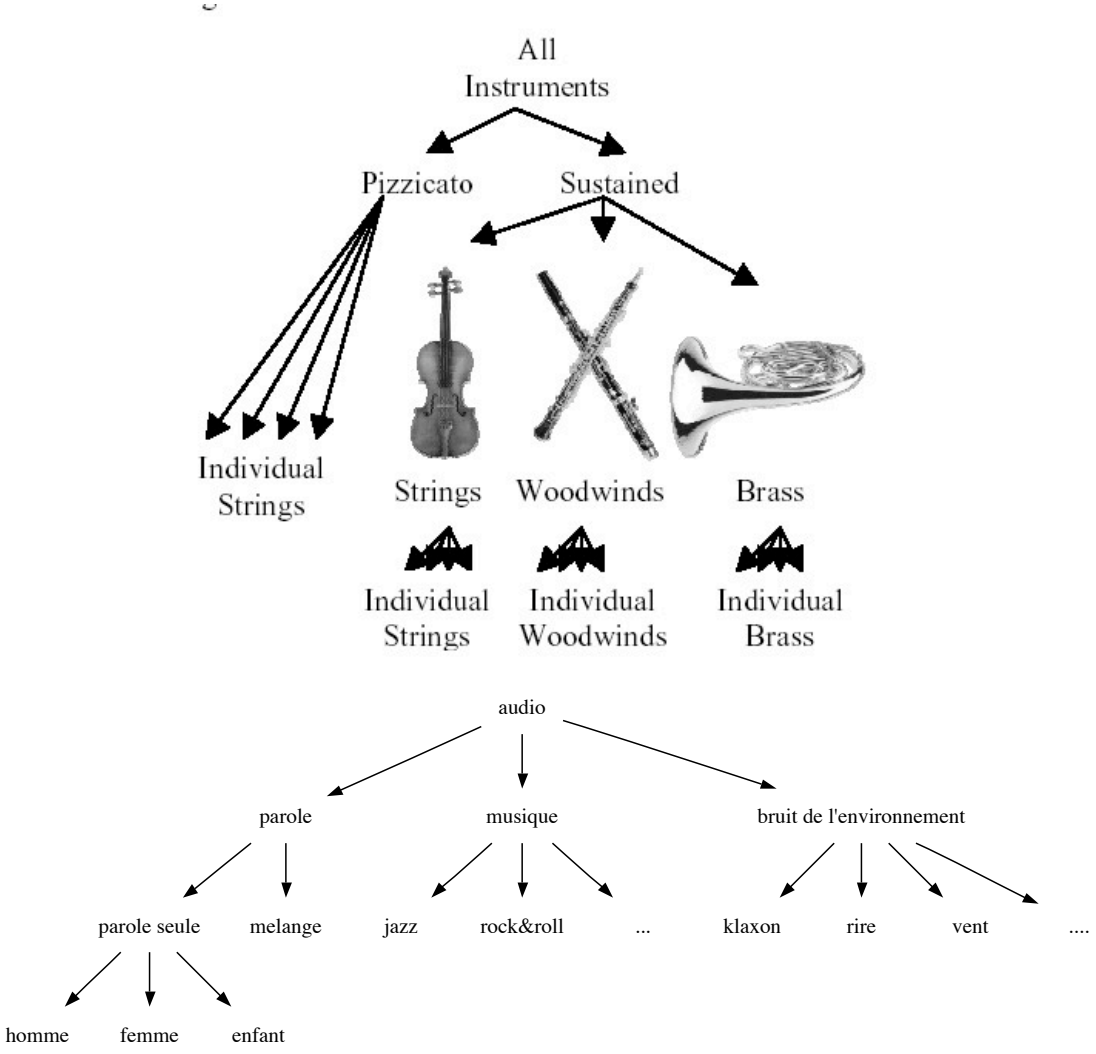

FIG. 2.11 – Modèles de classification hiérarchique de sons et d'instruments.

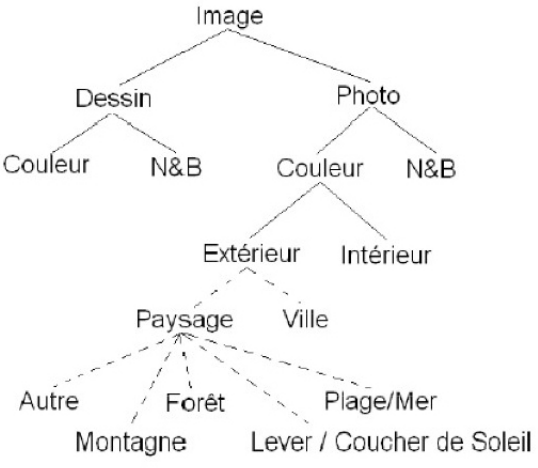

FIG. 2.12 – Modèle de classification hiérarchique d'images.

2.3 Fusion de descripteurs numériques

La classification automatique de **descripteurs bas** du signal estime la valeur de descripteurs moyens et hauts à partir de l'information provenant de caractéristiques physiques différentes. De plus, les descripteurs bas employés aujourd'hui comprennent plusieurs dimensions. Ainsi, l'ensemble des descriptions employées pour extraire un concept comprennent souvent un nombre important de caractéristiques. Les modèles de classifications présentés au paragraphe précédent considèrent les relations entre les informations provenant des différentes dimensions. Ils combinent plusieurs descripteurs et construisent des modèles de classification sur l'ensemble de ces informations. Nous pouvons donc considérer que ces modèles réalisent déjà une **fusion de données**.

Dans les faits, l'emploi de plusieurs descripteurs augmente la dimension de l'espace des descriptions, ce qui le rend plus clairsemé ("sparse"). Le problème de classification devient complexe et les risques de sur-apprentissage des données augmente. Cependant, cette approche est efficace pour les concepts dont un nombre suffisant d'exemples est disponible pour l'apprentissage.

La technique de fusion la plus simple consiste à concaténer les descriptions afin de former un descripteur unique de grande dimension. Ce modèle est appelé **"early" fusion**. Au sein de la recherche, de nombreuses études ont montré la nécessité d'opérer des modifications de ce descripteur afin d'améliorer les performances de classification (résultats et temps de calcul). Nous retiendrons trois tâches essentielles :

1. la standardisation des distributions,
2. la modification de l'espace des descripteurs,
3. la diminution du nombre de dimensions.

Les modèles de fusion sont déterminés par analyse et apprentissage de bases de données annotées. Les modèles de classification de concepts sont ensuite appris et appliqués sur les descripteurs bas transformés. Soit $\Sigma = \{O_i\}_{i=1..n}$ la base d'objets d'apprentissage et O_i le $i^{ème}$ objet. Nous appelons Δ l'espace des descriptions, $D(O_i) = \{D_j(O_i)\}_{j=1..m}$ la description du $i^{ème}$ objet et $D_j(O_i)$ la $j^{ème}$ dimension de la description du $i^{ème}$ objet. Nous notons $\Upsilon : \Delta \rightarrow \Delta'$ la fonction de transformation qui associe à une description D sa description transformée D'.

2.3.1 Standardisation de la distribution

Les descripteurs numériques employés pour décrire le contenu d'un objet proviennent de modèles d'extraction extrêmement variés. Ils concernent plusieurs caractéristiques physiques et peuvent décrire différents media. Il est donc nécessaire de normaliser les distributions de ces descripteurs afin de ne pas handicaper le système de classification.

La première étape de la normalisation consiste à centrer chaque dimension des descripteurs. Une nouvelle description est définie :

$$D'_j(O_i) = D_j(O_i) - \frac{1}{n} \sum_{ii=1..n} D_j(O_{ii})$$

où la moyenne à été soustraite à chaque composante. Ce traitement permet de réduire le bruit statique de convolution des descripteurs.

Il est ensuite nécessaire de fixer la dispersion sur les composantes de la population d'objets. Une nouvelle description est définie par :

$$D''_j(O_i) = \frac{D'_j(O_i)}{\sqrt{\frac{1}{n} \sum_{ii=1..n} D'_j(O_{ii})^2}}$$

où chaque composante a été divisée par son écart-type estimé sur la base d'apprentissage. Ainsi, les dimensions deviennent indépendantes de l'échelle de mesure et la classification n'est pas perturbée par un descripteur dont la variance serait trop importante.

Nous appelons description "standardisée" la nouvelle description $D'' = \Upsilon^{\text{stand}}(D)$ dont chaque composante est de moyenne nulle et d'écart-type égal à 1. Pour simplification des notations, elle sera notée D par la suite, et nous considérerons que cette transformation a été appliquée.

Récemment d'autres approches de normalisation des descripteurs ont été approfondies notamment dans le domaine de la reconnaissance du locuteur. Le principe étant de normaliser chaque composante de D afin de rendre sa distribution normale. Cependant cette opération semble entraîner une perte d'information trop importante, notamment quand elle est associée à un algorithme de classification dont le modèle de données n'est pas gaussien (e.g. plus proche voisin).

2.3.2 Transformation de l'espace des descripteurs

Afin d'améliorer la tâche de classification, il est souvent nécessaire de modifier l'espace des descripteurs pour mieux considérer les informations qu'il contient. De façon générale, ceci est fait par projection de l'espace des descripteurs vers un nouvel espace numérique de même dimension ou de dimension inférieure. Dans cette section nous resterons dans le cadre de transformations sans changement de la dimension. La modification du nombre de dimensions de l'espace est traité dans la section suivante.

La transformation de l'espace des descripteurs a été développée afin de réaliser deux tâches principales :

- rendre les descripteurs moins dépendants,
- augmenter la séparabilité entre classes.

Analyse en composante principale

Les tâches de compréhension et de classification obtiennent de meilleures performances lorsque les vecteurs d'observations sont **décorrelés** [Whi03], afin, d'une part de prévenir le sur-apprentissage et d'autre part d'« ancrer » les observations au sein de clusters simples. Les données répondent ainsi au critère de covariance diagonale des modèles gaussiens de classification (GMM, SVM). Une technique classique pour la décorrélation des descripteurs est l'**analyse en composante principale** (ACP) [Liu99].

Une ACP consiste à transformer les descripteurs standardisées D de Δ en de nouveaux descripteurs notés D' au moyen de combinaisons linéaires. Ce que l'on écrit matriciellement par $D' = D\mathbf{U}$ où $\mathbf{U}$ est la matrice de la fonction de transformation Υ. La corrélation des composantes D'_j de la nouvelle description est minimisée par le choix de $\mathbf{U}$, tel qu'elle soit la matrice des vecteurs propres de la matrice de corrélation des descriptions d'apprentissage. Les vecteurs de $\mathbf{U}$ forment une base orthonormée de projection. La somme des valeurs propres vaut m : une ACP transforme les descripteurs corrélés en descripteurs décorrélés tout en conservant la variance totale.

Plusieurs méthodes apparentées à l'ACP ont été créés afin de mieux prendre en compte certaines caractéristiques de l'information. La NMF réalise une décomposition similaire à l'ACP mais contraint sa fonction de transformation à être positive. Ce qui convient à l'idée que l'on additionne l'information de chaque dimension pour obtenir une décision finale. Dans ce cas les descripteurs transformés ne sont plus décorrélés.

Analyse en composantes indépendantes

Certains algorithmes de classification (comme les algorithmes naïfs de Bayes) fournissent de meilleures performances lorsque les dimensions des descripteurs sont indépendantes. Comme l'ACP, l'analyse en composantes indépendantes (ICA) [Hyv99] réalise une projection des données vers un nouvel espace.

Informellement, on dit que deux variables sont indépendantes si la réalisation de l'une n'apporte aucune information sur la réalisation de l'autre. La corrélation est un cas particulier de dépendance dans lequel la relation entre les deux variables est strictement monotone. Alors que la PCA ne considère que les moments du 2^{nd} ordre et décorrèle les dimensions, l'ICA prend en compte les statistiques d'ordre supérieur et rend **indépendantes** (au sens statistique) les dimensions des observations. Il existe plusieurs critères d'indépendance appelés fonctions de contraste : vraissemblance, entropie, information mutuelle etc.... La plupart d'entre elles sont très proches. L'optimisation de cette fonction permet d'extraire $\mathbf{U}$ par des techniques de minimisation classiques. L'ICA a beaucoup été employée ces dernières années pour la reconnaissance de visage en vidéo, citons [Liu99] pour exemple. Dans [Cas01], les auteurs réalisent une ICA qui permet d'obtenir une décomposition du signal audio en dimensions ayant une signification sémantique. Les dimensions résultantes correspondent au sens perçu de l'audio.

Cependant, l'ICA, comme la PCA ne prend pas en compte l'information fournie par l'annotation des objets de référence. Or il semble important, dans le cadre d'une modification de l'espace des descripteurs pour la classification, de maximiser la discrimination entre les classes du concept choisi.

Analyse discriminante linéaire

Afin de prendre en compte une organisation des données tel que le concept d'appartenance à une classe. La LDA [Net00, Pee02b, Loo02] cherche à maximiser la variance interclasse sans augmenter la variance totale des descripteurs. Cette technique a été développée par Fisher [Fis36] pour les concepts à 2 classes et étendu par Rao [Rao48] pour le cas multi-classe.

La LDA cherche une combinaison linéaire des variables qui **maximise la discrimination entre classes**. Ce critère est exprimé par la maximisation du critère de Chernoff noté :

$$J_{\text{cher}}(\mathbf{U}) = \text{tr}((\mathbf{U}\mathbf{W}\mathbf{U}^t)^{-1}(\mathbf{U}\mathbf{B}\mathbf{U}^t))$$

où $\mathbf{W}$ et $\mathbf{B}$ sont les matrices d'inertie intraclasse et interclasse.

De manière générale la LDA augmente fortement les performances de classification [Pee02b]. Cela est dû au fait qu'elle structure les données afin de rendre les différentes classes disjointes. Cependant cette transformation associée à un modèle de classification qui opère son propre traitement des données (e.g. les SVM) peut entraîner un problème de sur-apprentissage et une diminution des performances. Remarquons de plus, que la LDA suppose que les données ne soient pas hétérogènes (i.e. données dans lesquelles les classes n'ont pas les mêmes matrices de covariance).

Nous appelons description "transformée" la nouvelle description $D' = \Upsilon^{\text{trans}}(D)$ résultat de la transformation de l'espace des descripteurs sans diminution de la dimension. Pour simplifier les notations, elle sera notée D par la suite et nous considérons que cette transformation a été appliquée.

Dans le cadre de notre système nous emploierons deux techniques de projection afin de modifier l'espace des descripteurs :

- l'ACP afin de décorréler les dimensions du vecteur descripteur ;
- la LDA afin de prendre en compte la taxonomie de classification.

Nous montrerons leur utilité dans le cadre spécifique du film. De plus, nous déterminerons la position de chacune d'entre elles au sein du schéma de traitement des données.

2.3.3 Modification du nombre de descripteurs

L'utilisation d'un grand nombre de descripteurs pour la classification ou la segmentation peut apporter de nombreux désavantages : une taille trop importante des données stockées [Mar98] ; pour une classification donnée, certaines caractéristiques peuvent détériorer la reconnaissance si elles sont sans pertinence pour la classification choisie [Pee02b] ; la plupart des algorithmes de classification donnent

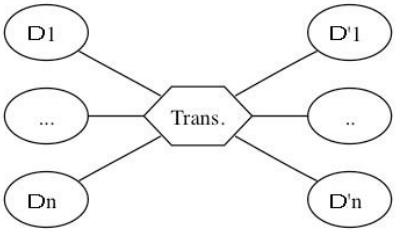

FIG. 2.13 – Transformation des descripteurs.

de meilleurs résultats si le nombre de dimensions est réduit (e.g. les SVM) ; enfin, s'il y a moins de descripteurs à calculer, le temps de calcul est limité. Il semble donc indispensable de réduire le nombre de dimensions du vecteur descripteur en essayant de conserver les caractéristiques du signal permettant la discrimination des classes.

La réduction du nombre de dimensions est le sujet de nombreuses recherches ; citons [Sco92, Rau91]. Des approches variées existent : dont la sélection des descripteurs [Del03], la "projection pursuit"[Int91], l'"independance grouping"[Bag97] et les methodes de sous-espaces. Toutes ces méthodes contiennent des approximations variées.

Pour un objet O, l'extraction des descripteurs numériques permet d'obtenir une description D de dimension m. On cherche ici à diminuer la dimension de l'espace des descriptions tout en gardant un maximum de l'information contenue dans les données. Le choix du nombre m' de facteurs retenus, fruit d'un compromis plus ou moins aisé, constitue une décision souveraine de la part du chercheur, qu'il lui faut assumer et pouvoir défendre. Retenir, $m' = m$ composantes équivaut à garder toute l'information initiale, et donc ne permet pas de simplifier la structure des liaisons entres variables. Inversement, ne garder qu'un petit nombre de dimensions peut revenir à n'expliquer qu'un pourcentage trop faible de l'information totale, et à résumer de façon excessive la complexité de la structure des liaisons entres variables ; à moins que quelques dimensions seulement suffisent à expliquer une proportion importante de l'information totale. On a là affaire à un dilemme typique entre les objectifs opposés de "conservation de l'information" et de "simplification de l'information". Quel que soit le compromis adopté, une chose est sûre, si l'on décide de ne conserver qu'un certain nombre m' de dimensions, il s'agira des dimensions qui contiennent **le maximum d'information**. Les différences des techniques de réduction des descripteurs résident dans la définition de l'information utile à conserver. Après avoir éclairé ce concept, les dimensions sont classées par apprentissage sur la base, selon la quantité d'information qu'elles contiennent. Puisque le pouvoir d'explication décroît du fait de leur ordonnancement par valeurs décroissantes de l'information, il est de moins en moins intéressant de les conserver pour le traitement.

La réduction du nombre de dimensions peut être décomposé en deux tâches distinctes, l'élimination de descripteurs et la projection vers un sous-espace.

Méthodes des sous-espaces

En ce qui concerne les méthodes des sous-espaces, l'hypothèse est que l'information est confinée dans un **espace linéaire de dimension réduite**.

Ces techniques consistent donc à projeter les données vers un nouvel espace par transformation ($D' = \Upsilon^{\text{trans}}(D)$). Les dimensions D'_j sont ordonnées selon la quantité d'information qu'elles contiennent. Pour les algorithmes PCA et LDA l'idée est que l'information contenue dans la $j^{ème}$ dimension transformée D'_j est proportionnelle à la valeur propre associée λ_j. Pour l'ICA une quantité similaire est associée à chaque dimension. Les m' dimensions de la description transformée D' qui contiennent le plus d'information sur les données sont conservées.

La recherche s'est empressée de proposer un certain nombre de critères permettant de faire apparaître objectif le choix de m'. Trois critères se dégagent, ils sont purement empiriques et sans fondement statistique, mais néanmoins très simples d'utilisation et donc largement répandus :

1. S'arrêter dès que la proportion d'information totale atteint un seuil fixé, par exemple 90% (critère de Joliffe).
2. Ne garder que les valeurs propres $\{\lambda_j\}_{j=1..m'}$ supérieures à leur moyenne (critère de "Kaiser"). Si l'on travaille avec la matrice des corrélations, cela revient à exclure les valeurs propres inférieures à 1.
3. Faire un "scree graph" (scree = éboulis) consistant à représenter les valeurs propres en fonction de j. Souvent, un "coude" apparaît, marquant un changement de régime dans la décroissance des λ_j. Alors, seules les valeurs propres apparaissant avant l'apparition du coude sont conservées ("critère" de "Cattell").

Ces techniques de réduction de la dimension par projection PCA et LDA rencontrent plusieurs difficultés liées au fait qu'elles présupposent que les variances sont des mesures adaptées de liaison entre variables. Pour cela, il faut que ces liaisons soient à-peu-près **linéaires**, ce dont on peut s'assurer en produisant les diagrammes de dispersion pour tous les couples de variables, qui ne doivent pas manifester de courbure marquée. Dans le cas contraire, il serait judicieux de tenter de linéariser ces relations (au moyen de transformations de variables adéquates) avant de procéder à une projection ; ou mieux, d'appliquer une réduction de l'espace qui considère les liaisons entre variables d'ordre supérieur, comme l'ICA.

D'autres méthodes de transformation préservent les relations importantes et potentiellement complexes entre les dimensions. Elles se caractérisent par la définition de ces relations et donc de l'information à conserver lors de la réduction : la décompostion en valeurs singulières (SVD) [Dee90, Sla03] permet de réduire la dimension en conservant les distances entre objets de la base de données. Par rapport à la PCA, elle dispose d'une meilleure capacité à considérer les non linéarités entre variables[Dee90]. Les techniques comme la SVD, qui ne modifie pas les distances dans le nouvel espace des données, sont particulièrement adaptées aux classifications spatiales locales (range query et plus proches voisins) qui utilisent cette information. Sur la même idée, plusieurs techniques fondées sur la préservation des voisinages locaux ont été développées, notamment la LLE et l'ISOMAP [Vla02]. Ces algorithmes réduisent le nombre de dimensions en conservant la topologie locale non linéaire des données. L'ISOMAP préserve en plus la distance géodésique entre chaque paire d'objets. Mais par rapport au LLE, il crée des matrices non creuses, dont la structure fait perdre du temps et de l'espace. Cependant ces techniques ne considèrent pas l'information liée au concept de classification contrairement aux techniques LDA.

Dans les faits, lorsqu'un objet Q est posé en question, toutes ces méthodes nécessitent l'extraction de l'intégralité des descripteurs bas de Q pour procéder à la classification. Nous l'utiliserons dans notre système pour diminuer la taille des index, et améliorer la classification.

Elimination de descripteurs

En ce qui concerne l'élimination de descripteurs, l'approximation est qu'une grande partie de l'information concernant la classification est contenue dans un petit nombre de descripteurs. Cette méthode consiste donc à déterminer les m'' descripteurs les plus porteurs d'information et à éliminer les autres. Elle se fait sans transformation de l'espace des données. Plusieurs algorithmes permettent de classer les dimensions selon la quantité d'information qu'elles contiennent.

Un **algorithme de transformation** permet d'extraire la « première » dimension de l'ensemble des points. C'est celle qui contient le maximum d'information en terme de variance (PCA, ICA) ou de discrimination des classes (LDA). Comme les variances des descripteurs sont normalisées, chaque composante

du premier vecteur de **U** (la matrice de transformation) donne le poids de chaque dimension du descripteur initial et donc son importance pour la dimension « première ». Une méthode consiste à choisir le nombre m'' de descripteurs qui montre la pondération la plus forte. Pour déterminer m'', les critères cités précédemment sont employés en remplaçant les valeurs propres par des poids.

Cependant, cette technique ne prend pas en compte la variance sur les autres axes discriminants de **U**. Ainsi, lorsque une décroissance lente de l'information est observée le long des dimensions transformées, une grande partie de l'information n'est pas prise en compte pour le classement. La Step-wise PCA (ou LDA...) affine cette technique en pratiquant des PCA (ou LDA) en cascade. A chaque itération, la meilleure dimension est sélectionnée (ou la dernière est éliminée).

L'**information mutuelle** $\mathrm{MI}(D_j, C)$ mesure la quantité d'information apportée en moyenne par une réalisation de D_j sur les probabilités de réalisation du concept de classification C. La quantité MI est donc une mesure de l'information utile apportée par un descripteur. Elle est estimée par apprentissage de la base d'exemples :

$$\mathrm{MI}(D_j, C) = \sum_{i,k} \mathrm{P}(D_j(O_i), C = c_k) \ \log\left(\frac{\mathrm{P}(D_j(O_i), C = c_k)}{\mathrm{P}(D_j(O_i))\mathrm{P}(C = c_k)}\right).$$

Les probabilités jointes et marginales sont déterminées sur la base d'exemples de la même façon que pour les modèles de classification (classification, algorithme EM et comptage). La méthode d'élimination consiste à choisir le nombre m'' de descripteurs qui montrent l'information mutuelle la plus forte.

Cependant cette technique ne considère pas les relations de dépendances entre variables. La Step-wise MI affine cette technique en pratiquant des MI en cascade. A chaque itération les informations mutuelles $\mathrm{MI}(\{D_1, \ldots, D_{j-1}, D_{j+1}, \ldots, D_{m'}\}, C)$ de l'ensemble des descripteurs sélectionnés, sauf un, et du concept, sont estimées pour toutes les dimensions j. Le meilleur ensemble de dimensions, celui dont la MI est la plus forte, est sélectionné.

Dans [Del03], nous avons comparé plusieurs algorithmes d'élimination de descripteurs pour une application de classification d'extrait audio en *parole/musique/mélange*. Les résultats observés montrent que les algorithmes basées sur la MI obtiennent de meilleures performances que ceux fondés sur une transformation (PCA, LDA). [Pee02b] arrive aux même conclusions pour une classification de sons d'instruments.

Applications de la réduction de dimensions

Dans les faits, l'utilité de la réduction du nombre de dimensions semble discutable. En effet, à moins de disposer d'une quantité conséquente de descripteurs numériques (> 1000), ce qui n'est pas notre cas, la perte d'information est souvent trop importante, ce qui diminue les performances de classification. Nous essaierons de le vérifier dans nos expériences ultérieures.

Par contre, pour d'autres desseins, comme la représentation d'objets dans un espace 2D, la transformation de l'espace par réduction de la dimension est indispensable. Il est possible d'appliquer les méthodes précédemment étudiées. Cependant plusieurs techniques ont été développées spécialement pour l'affichage graphique. Notamment les MDS qui s'apparentent aux SVD et les décompositions SOM largement utilisées aujourd'hui pour visualiser des archives sonores [Rau01, Kur99], par exemple.

Nous appelons description "réduite" la nouvelle description $D' = \Upsilon^{\mathrm{red}}(D)$, résultat de la réduction du nombre de dimensions de l'espace des descripteurs. Pour simplification des notations, elle sera notée D^{red} par la suite.

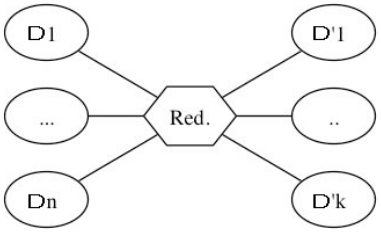

FIG. 2.14 – Réduction du nombre de dimensions de description.

2.4 Fusion des classifications

Récemment, la combinaison parallèle de classifications ou **"late" fusion** a été proposée comme une voie de recherche permettant d'améliorer les performances (taux de reconnaissance, fiabilité,...) des systèmes de classification. Cette fusion exploite la complémentarité qui peut exister entre les classifications.

De façon idéale, la fusion des descripteurs numériques devrait fonctionner pour tous les concepts, puisque l'ensemble des informations est disponible pour les classifications. Cependant, en pratique, le nombre limité d'exemples d'apprentissage augmente les risques de sur-apprentissage. Il est donc parfois nécessaire d'employer une stratégie alternative. Si les descripteurs sont suffisamment décorrélés, il est possible de les traiter indépendamment. Dans de telles situations, il est envisageable de modéliser un concept pour chaque descripteur ou groupe de descripteurs, de façon indépendante, et de fusionner la décision des classifications par la suite. Un modèle du même concept est créé pour chaque groupe de descripteurs, ce qui résulte en de multiples classifications auxquelles sont associés différents résultats. Ces résultats peuvent prendre différentes formes : la classe majoritaire pour chaque classification, une liste ordonnée de classes, une liste accompagnée de taux de confiance. La valeur finale du concept est ainsi estimée par la fusion de ces résultats.

Dans ce paragraphe nous étudions plusieurs exemples de fusion des classifications et nous présentons certaines règles de fusions. De plus, nous explorons les techniques de normalisation des scores pour la fusion qui permettent d'augmenter la performance globale de la classification.

2.4.1 Règles de fusion

À ce propos, plusieurs méthodes ont été développées et intensivement utilisées dans des problèmes de reconnaissance différents. Nous pouvons citer par exemple la reconnaissance du manuscrit [Xu92], la reconnaissance de chiffres manuscrits [Hua95, Kit98], la reconnaissance de visages [Ach96] et la vérification de signatures [Baj97]. Elles se distinguent principalement par leur capacité d'apprentissage et le type de sortie des classifications qu'elles combinent.

Les méthodes de type I combinent les classifications dont chacune attribuent une classe unique au concept à déterminer. Parmi ces méthodes, on trouve le vote majoritaire [Kit98], la théorie de Bayes [Xu92], la théorie de Dempster-Shafer [Xu92] et la méthode BKS (Behaviour Knowledge Space) [Hua95]. Les méthodes de type II combinent des listes ordonnées de classes. Le Borda count appartient à cette catégorie [Erp00]. Les méthodes de type III utilisent en plus une « **confiance** » associée à chacune des classes de la taxonomie. Elles incluent les réseaux de neurones [Hua95] et les règles fixes comme le maximum, la somme, la moyenne et le produit [Kit98]. Le choix d'une méthode de combinaison reste cependant l'un des problèmes les plus difficiles rencontrés lors de la conception d'un système à plusieurs classifications. En effet, les travaux qui traitent ce problème testent généralement un ensemble de méthodes de combinaison pour retenir la plus adéquate en fonction de certains critères de performance. Cependant, les résultats obtenus restent étroitement dépendants des applications traitées et par conséquent, difficiles

à généraliser en dehors d'un contexte applicatif donné. D'autre part, il existe certains travaux intéressants qui traitent le problème de l'évaluation des méthodes de combinaison avec des données réelles en utilisant des bases de données différentes[Erp00]. Toutefois l'utilisation de données réelles ne permet pas d'avoir suffisamment de variabilité dans les performances des classifications à combiner pour analyser de façon fiable le comportement de ces méthodes de combinaison [Kit98].

Dans le cadre de notre étude, nous utiliserons une fusion de type III, car elle prend en compte la totalité de l'information fournie par les différentes classifications. De plus, la plupart des algorithmes de classification calculent une probabilité ou un taux de confiance dans la valeur donnée au concept. Les auteurs de [Lee97] montrent que pour un grand nombre de classes, il est préférable d'utiliser une méthode de type II. Cependant, ils emploient une méthode de type III sans modifier la distribution des taux de confiance, ce qui donne trop de poids à certaines classifications fausses. Une des solutions est de normaliser les distributions de ces scores.

Dans la suite, le problème de la fusion des scores sera envisagé de façon numérique. Soit un modèle de classification $\mathcal{M}(\mathrm{T}, \mathrm{X}, \Psi)$ et un objet O ayant D pour description. La valeur prise par le concept de classification pour un objet de test est évaluée par plusieurs modèles de classifications notés $\mathcal{M}^j(\mathrm{T}, \mathrm{X_j}, \Psi_j)_{j=1..m}$. Le **concept "intermédiaire"** associé à la $j^{ème}$ classification est appelé C_j. Nous définissons D_j, la description de l'objet O utilisée pour déterminer le concept intermédiaire C_j. Enfin, l'appartenance à une classe est traduite par un score : une probabilité, un taux de confiance, une distance.... Pour la classification j, les scores attribués sont stockés au sein du vecteur des scores du concept C_j noté $S_j(O) = (s_j^k(O))$, où $s_j^k(O)$ est le score attribué à la $j^{ème}$ classification de l'objet dans la classe c_k.

2.4.2 Normalisation des scores de classification

Les scores numériques utilisés pour décrire la confiance dans l'appartenance d'un objet à une classe proviennent de modèles d'extraction extrêmement variés. Ils concernent des caractéristiques physiques différentes et peuvent provenir de plusieurs media. Il est donc nécessaire de normaliser les distributions afin de ne pas handicaper l'algorithme de classification. Les étapes de la normalisation des scores sont principalement liées aux caractéristiques des probabilités. En effet, dans le cadre de la création d'un système probabiliste de description du contenu, il est important pour nous de modifier ces scores afin qu'ils aient la forme de probabilités.

Nous noterons $P^k(O)$ la pseudo-probabilité qu'un objet O ayant D pour description appartiennent à la $k^{ème}$ classe de la taxonomie $\mathrm{T_j}$; $P^k(O)$ est une approximation de la probabilité conditionnelle $\mathrm{P}(D|C = c_k)$. De même $P_j^k(O)$ est la pseudo-probabilité qu'un objet O ayant D_j pour description appartienne à la $k^{ème}$ classe de la $j^{ème}$ taxonomie intermédiaire $\mathrm{T_j}$, ainsi $P_j^k \sim \mathrm{P}(D_j|C_j = c_k)$.

Les scores sont définis sur des intervalles différents suivant l'algorithme de classification : $[0, 1]$ pour les modèles probabilistes (e.g. naïf de Bayes), $\mathbb{R}^+$ pour les modèles locaux (e.g. KNN), $\mathbb{R}$ pour les modèles globaux (e.g. SVM). Il est donc nécessaire de les modifier, afin de les projeter sur le même intervalle. Ce processus est un point critique du système d'intégration des classifications puisqu'il doit se faire sans réduire trop le pouvoir de discrimination de chacun des scores. Nous choisissons de projeter ces scores sur l'intervalle borné $[0, 1]$, afin de rester dans un cadre probabiliste. Pour ce faire, la technique la plus simple est d'appliquer une transformation linéaire[Lee97]. Si nous notons $\Sigma = \{O_i\}_{i=1..n}$, la base d'apprentissage et Q un objet question. Pour la $j^{ème}$ classification le nouveau score de la $k^{ème}$ classe s'écrit :

$$s_j'^k(Q) = \frac{\max_i(s_j^k(O_i)) - s_j^k(Q)}{\max_i(s_j^k(O_i)) - \min_i(s_j^k(O_i))}$$

Afin d'augmenter la discrimination des classes, dans le cadre d'une classification binaire par SVM, [Jou97] et [Ben98] projettent le taux de confiance en appliquant la fonction sigmoïde de Genoud [Gen96].

$$\text{genoud} : \mathbb{R} \rightarrow]0,1[\ , \ x \mapsto \frac{1}{1+\exp(-x)}$$

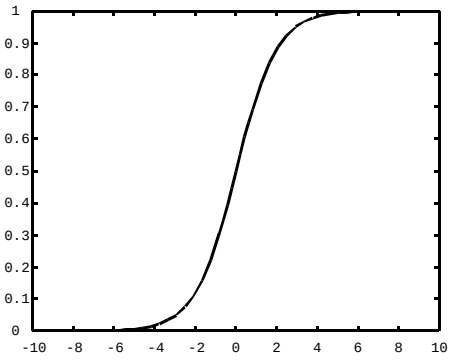

FIG. 2.15 – Fonction sigmoïde de Genoud sur l'intervalle $[-10, 10]$.

La forme de cette fonction (figure 2.15) augmente la séparabilité des points dont le score est proche de la valeur seuil 0 et diminue celle des points dont la classification est plus évidente. Dans ce cas le nouveau score s'écrit :

$$s'^k_j(Q) = \text{genoud}\left(s^k_j(Q)\right)$$

[Jou97] applique cette même fonction à des scores à valeurs dans $\mathbb{R}^+$ issus d'une classification par plus proches voisins. Dans ce cas le nouveau score s'écrit :

$$s'^k_j(Q) = \text{genoud}\left(s^k_j(Q) - se\right) \text{ avec } se = \frac{1}{n}\sum_{i=1}^{n} s^k_j(O_i)$$

se étant un seuil de décision fixé par l'utilisateur. Dans ce cas, il est estimé par la moyenne des exemples annotés afin de centrer le terme de la fonction autour de 0.

Dans notre cas, les scores sont principalement issus de modèles de classification par support vector machine. Ce sont des réels centrés sur 0 ; l'objet question appartient à la classe k pour le $j^{ème}$ modèle si $s^k_j(Q)$ est positif. C'est pourquoi nous appliquerons la normalisation :

$$s'^k_j(Q) = \text{genoud}\left(\frac{s^k_j(Q)}{\alpha\sqrt{\frac{1}{n}\sum\limits_{ii=1..n} s^k_j(O_{ii})^2}}\right),$$

où chaque score a été divisé par un terme composé d'une constante α et de l'écart-type des scores de la base d'apprentissage, ceci pour fixer la dispersion autour de la valeur seuil 0 et ne pas favoriser un modèle ou une classe dont la variance est trop importante. Remarquons que, plus α est petit, plus la séparation des points autour de zéro est augmentée par la normalisation. Si cette constante est choisie trop faible, alors la discrimination des points éloignés du seuil est fortement diminuée, ce qui risque d'entraîner une perte d'information importante et de corrompre la fusion des classifications, lorsque celle-ci est ambiguë (plusieurs classifications contradictoires). Au contraire, si α est trop importante, alors la transformation de Genoud perd son pouvoir de discrimination, car elle est linéaire au voisinage de 0. D'expérience nous fixerons $\alpha = 1/3$, ce qui nous semble constituer un bon compromis.

Enfin, la définition choisie pour les taxonomies de concepts entraîne la densité des valeurs possibles pour un concept ce qui entraîne : $\sum_{k=1..p}(s_j^k(Q)) = 1$. Ceci est assuré par la normalisation simple :

$$s_j'^k(Q) = \frac{s_j^k(Q)}{\sum_{k=1}^{p} s_j^k(Q)}.$$

Pour la $j^{ème}$ classification intermédiaire, la normalisation des scores d'un objet question permet d'obtenir un vecteur score intermédiaire, noté $S_j(Q) = \{s_j^k(Q)\}_{k=1..p}$. La dimension k correspond à la pseudo-probabilité que l'objet appartienne à la classe c_k pour le modèle de classification j : $s_j^k(Q) = P_j^k(Q)$.

2.4.3 Fusions simples des scores

La technique de fusion des vecteurs scores envisage de déterminer les pseudo-probabilités $P^k(Q)$ que l'objet en question appartienne à la classe c_k, sachant sa description pour le modèle global de classification, ceci pour tout k. La méthode consiste à combiner l'ensemble des vecteurs **scores intermédiaires** afin d'obtenir un vecteur **score final** pour le concept choisi. Nous notons $S = \{s^k = P^k(Q)\}_{k=1..p}$ ce vecteur dont la dimension k est la pseudo-probabilité globale que l'objet appartienne à la $k^{ème}$ classe. La fonction de fusion notée $\mathcal{F}_c$ est définie par :

$$\mathcal{F}_c \left(\begin{array}{l} \{S_j(Q))\}_{j=1..m} \mapsto S(Q) \\ [0,1]^m \rightarrow [0,1] \end{array} \right)$$

Pour réaliser la combinaison des classifications, plusieurs fonctions ont été développées dans la littérature. [Zha98] présente les trois techniques simples les plus couramment utilisées :

1. Minimum : $s^k(Q) = \min_{j=1..m}(s_j^k(Q)) \quad \forall k \in \{1,\ldots,p\}$
2. Maximum : $s^k(Q) = \max_{j=1..m}(s_j^k(Q)) \quad \forall k \in \{1,\ldots,p\}$
3. Moyenne : $s^k(Q) = \frac{1}{m}\sum_{j=1..m} s_j^k(Q) \quad \forall k \in \{1,\ldots,p\}$

Il montre que la moyenne obtient de meilleurs résultats pour la classification de visages, car elle prend en compte toutes les classifications et est donc moins sensible au bruit. [Lee97] applique la moyenne arithmétique, qui obtient des résultats équivalents à ceux de la moyenne géométrique. Cependant ces fonctions attribuent le même poids à chaque concept intermédiaire, ce qui est gênant lorsque certains ne sont pas discriminants pour la taxonomie choisie. C'est pourquoi il a été introduit une somme pondérée [Bru95] utilisée dans de nombreuses applications :

$$s^k(Q) = \sum_{j=1..m} \omega_j s_j^k(Q) \quad \forall k \in \{1,\ldots,p\}\,.$$

Notamment pour la reconnaissance du locuteur où les poids ω_j sont appris par méthode dichotomique afin de minimiser le nombre d'erreurs. De même, on peut définir une multiplication pondérale :

$$s^k(Q) = \prod_{j=1..m} s_j^k(Q)^{\omega_j} \quad \forall k \in \{1,\ldots,p\}$$

Elle est appliquée par [Net00] qui apprend les poids par une méthode simple « amoeba search » pour une application audio-visuelle de reconnaissance de la parole. Les auteurs montrent l'équivalence de la moyenne et du produit pondéré pour les tâches de classification. Cependant, ces techniques définissent des relations polynomiques entre les scores de classification.

2.4.4 Fusion par classification des scores

Afin de mieux considérer les dépendances complexes et sûrement non-polynomiques entre ensembles de descripteurs, une autre méthode a été développée pour la détermination de concepts binaires de taxonomie $\mathrm{T} = \{0, 1\}$. Dans ce cas, le résultat de la classification du concept intermédiaire j peut être caractérisé par une seule variable. En effet, pour un objet question Q, la probabilité que le concept soit présent et la probabilité qu'il soit absent sont liées par la relation linéaire : $P_j^1(Q) = 1 - P_j^0(Q)$. La probabilité du concept présent, $s_j^1(Q) = P_j^1(Q)$, est choisie pour représenter le score de classification. Ainsi, le vecteur **score de présence** est défini comme $S^1(Q) = \{s_j^1(Q) = P_j^1(Q)\}_{j=1..m}$ où s_j^1 est la pseudo-probabilité estimée par le $j^{ème}$ modèle que le concept C soit présent dans l'objet.

Ce vecteur placé dans l'espace des scores permet de pratiquer une nouvelle classification pour le concept global choisi. Ce nouveau modèle est appris sur les scores de classification des objets de la base d'apprentissage Σ. Nous l'appelons $\mathcal{M}_c(\mathrm{T}, S^1, \Psi)$ où $\mathrm{T} = \{c_1, \ldots, c_k, \ldots, c_p\}$ est la taxonomie, $S^1 : \Pi \rightarrow [0, 1]$ est la fonction qui associe un vecteur score de présence à tout élément de la population et Ψ est la fonction de classement qui associe une classe à tout élément de la population.

Différents algorithmes de classification ont été employés pour ce type de fusion, notamment les réseaux bayésiens [Aqa02], les réseaux de neurones [Ben02], les plus proches voisins [Gri04] et les SVM [Ben98]. [Ben02, Gri04] montrent, pour la combinaison de différentes classifications de textes, que les SVM fournissent les meilleurs performances. De même, dans le cadre d'applications de fusion de media, [Ben98] et [Aqa02] ont montré l'efficacité de la late fusion pour le traitement de données multimedia. Les auteurs définissent un nouveau concept intermédiaire pour chaque modalité. Ils remarquent que les SVM obtiennent de meilleurs résultats de classification que les modèles naïfs de Bayes, car ils prennent mieux en compte les relations de dépendances entre concepts intermédiaires.

Nous pensons que cette technique de fusion est aisément applicable au cas multi-classe. En effet, nous savons que les classifications multi-classe produisent un ensemble de m vecteurs scores S_j. Il est aisé de pratiquer la concaténation de ces vecteurs afin d'obtenir un vecteur score global $\mathbf{S}$. Ce vecteur est ensuite classé par un modèle de classification multi-classe du concept global de départ.

Ainsi, la fusion des classifications ou « late » fusion permet d'utiliser l'information de plusieurs modèles d'un concept afin de déterminer sa valeur. Dans le prochain chapitre nous verrons comment fusionner plusieurs concepts moyens différents pour influencer la classification d'un concept de niveau sémantique haut.

2.5 Fusion des concepts et des descripteurs numériques

Récemment, la description par des concepts a été proposée comme une voie de recherche permettant d'améliorer les performances de classification d'autres concepts. La fusion exploite la complémentarité de l'information qui peut exister entre concepts.

2.5.1 Fusion des concepts

La première application de la fusion des concepts est l'extraction de descripteurs de niveau sémantique haut à l'aide d'annotations textuelles appelées **concepts « atomiques »** (figure 2.16). Il est ainsi possible de faire des inférences sur des concepts complexes fondées sur les relations entre des concepts de niveau sémantique inférieur détectés précédemment. Les concepts atomiques sont des descripteurs de niveau moyen binaires $\mathrm{T} = \{0, 1\}$ indiquant la présence d'**objets** (voitures, hélicoptère, visage, lunette), d'**événements** (explosion, homme qui marche) etc.... La fusion de ces concepts permet de déterminer des descripteurs de niveau haut souvent multimedia comme le lieu (extérieur, plage) [Gri04], ou des

concepts complexes : la détection de parole et la détection d'un visage dans une vidéo doit mener à l'inférence « personnage qui parle », par exemple.

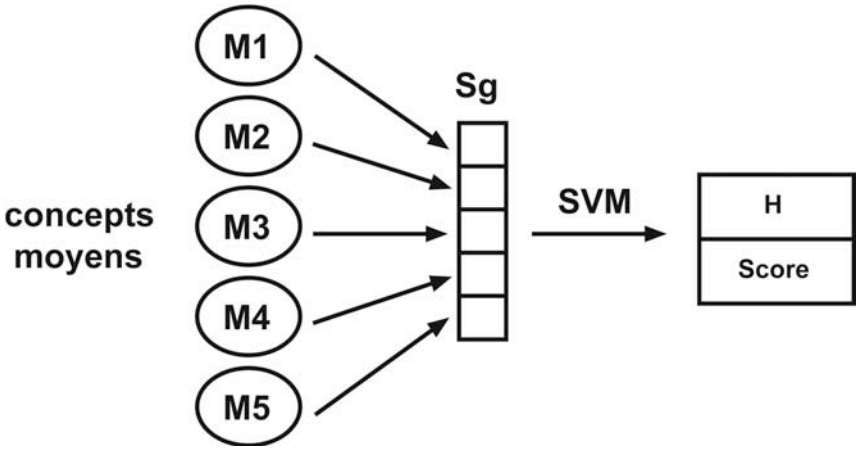

FIG. 2.16 – Classification d'un concept haut par des concepts moyens.

Soit $M = M_1, \ldots, M_p$ l'ensemble des concepts atomiques utilisés pour l'extraction du concept haut H. Nous définissons le vecteur score de présence des concepts atomiques de H noté $S^1(Q) = \{s_j^1(Q) = P_j^1(Q)\}_{j=1..m}$, où $s_j^1(Q)$ est la pseudo-probabilité que le $j^{ème}$ concept atomique soit vrai pour l'objet Q. Ce vecteur placé dans l'espace des scores permet de pratiquer la classification pour le concept haut choisi. Nous appelons $\mathcal{M}_m(\mathrm{T}, S^1, \Psi)$ ce modèle.

Jusqu'à présent, seul les **réseaux bayésiens** [Luo01] ou des variantes [Nap00] ont été employés pour la fusion des concepts. Remarquons que comme pour la late fusion, il semble possible d'appliquer d'autres modèles de classification, les SVM ou les KNN, afin de mieux prendre en compte la dépendance entre concepts atomiques. Mais aucun de ces algorithmes n'a été testé à ce jour.

La fusion des concepts exploite les interrelations complexes d'information entre concepts atomiques. Cependant un grand nombre de caractéristiques physiques sont trop complexes pour être décrites par des concepts : la couleur d'une image par exemple. C'est pourquoi nous devons aussi envisager la fusion de concepts et de descripteurs numériques du signal.

2.5.2 Fusion des concepts et des descripteurs numériques

La fusion de concepts et de descripteurs numériques permet d'utiliser l'ensemble des informations simples pouvant décrire un objet afin de déterminer la valeur de concepts de niveau sémantique supérieur. Plusieurs modèles d'intégration des concepts et des descripteurs numériques ont été développés dans la littérature.

Réseaux bayésiens

Le plus simple est un modèle génératif qui utilise le formalisme des réseaux bayésiens (RB) proposé par [Luo01]. Dans le cadre de la classification de photographie en *intérieur/extérieur*, les auteurs fusionnent certains concepts granulaires (herbe, ciel…) avec des descripteurs numériques bas (couleur, texture) de l'image. Ils obtiennent une amélioration sensible des résultats de classification par rapport à la classification par descripteurs numériques seuls.

Les auteurs réalisent une **late fusion** hiérarchique des concepts atomiques et des descripteurs bas en considérant les relations de dépendances décrites par le réseau représentées par figure 2.17. Soit le concept H de modèle de classification $\mathcal{M}_{BM}(\mathrm{T}, \mathrm{X}, \Psi)$. A chaque descripteur bas ou moyen est associé

un concept intermédiaire H_j, puis pour chacun des deux types de descripteurs, un nouveau concept intermédiaire est créé, H_B pour les descripteurs bas et H_M pour les moyens.

Cette technique permet d'extraire des concepts de niveau haut à partir de concepts moyens et de descripteurs numériques. Mais ce réseau bayésien suppose l'indépendance des descripteurs entre eux.

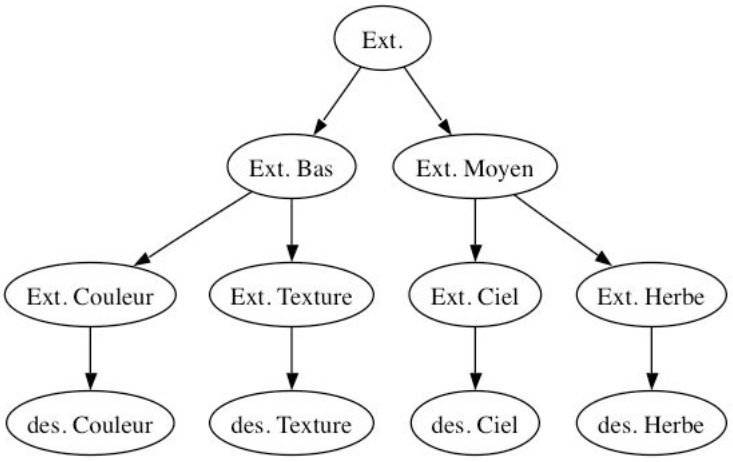

FIG. 2.17 – Représentation graphique de la dépendance du modèle RB haut.

Multinet

C'est pourquoi, [Nap00] envisage la fusion de concepts et des descripteurs bas de façon différente. En effet, les auteurs placent tous les concepts au même niveau sémantique. Ils peuvent appartenir à n'importe laquelle des catégories de base : objets(voitures, homme, hélicoptère), lieu (extérieur, plage), ou des événements (explosion, homme qui marche), et ils sont reliés par une relation de dépendance conditionnelle avec les descripteurs bas qui leurs sont associés. Intuitivement, il est clair que la présence de certains concepts suggère une forte possibilité de détecter d'autres concepts. De manière similaire, certains concepts sont moins à même d'apparaître en la présence d'autres concepts. La détection de ciel et/ou de mer augmente la probabilité du concept "plage", et réduit la probabilité de "intérieur". Pour intégrer tous les concepts et modeler leurs interactions, les auteurs proposent un réseau de concepts appelé multinet. Une figure conceptuelle du multinet est présentée dans la figure 2.18, les signes positifs montrent les interactions positives, et vice versa.

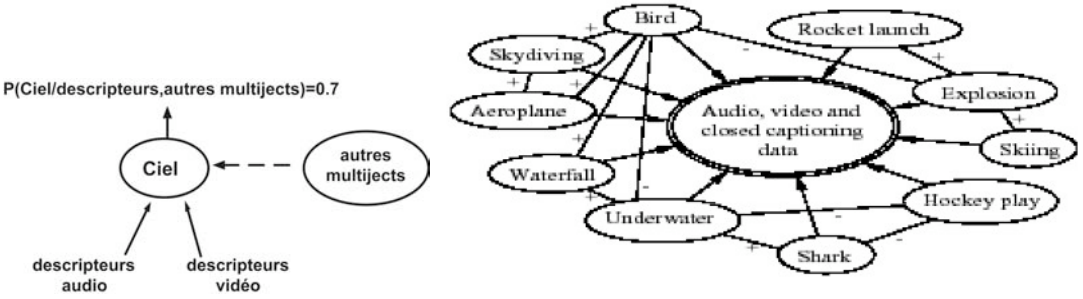

FIG. 2.18 – Modèle multinet d'indexation du contenu.

Le désavantage du multinet, comme pour les réseaux de neurones, est qu'il doit envisager la relation entre tous les concepts du système de traitement. De plus, il oblige à un calcul en deux étapes, ce qui alourdit encore le temps de calcul. Dans notre cas, le nombre important de concepts et de descripteurs

rend les calculs prohibitifs. Une solution est donc de créer un nouveau modèle de fusion permettant de prendre en compte les divers types de relations de dépendance.

2.6 Structure et segmentation

La segmentation est une étape importante de l'extraction d'information, elle permet de mettre en évidence la structure interne d'un objet. Elle trace des frontières au sein de la représentation, ce qui isole des sous-objets dont le contenu sera analysé. La segmentation peut être spatiale, temporelle ou spatio-temporelle, suivant le media auquel elle est appliquée et l'objet qu'elle doit isoler. Comme les descripteurs, la structure peut être extraite à différents niveaux de granularité (locale, intermédiaire) et de sémantiques (bas, moyen, haut). Il est important de définir le type d'objets que l'on cherche à découper. En effet, pour simplifier, la segmentation délimite des sous-objets au sein desquels les caractéristiques varient peu. Ainsi le choix des caractéristiques sémantiques à mettre en valeur conditionne le choix des descripteurs utilisés pour segmenter.

Dans ce chapitre, nous définissons les tâches de segmentation pour les media utilisés dans notre système. Puis nous étudions particulièrement la structure du film et les techniques de segmentation associées.

2.6.1 Support de la segmentation

Segmentation d'image

L'image est un support à deux dimensions spatiales. La segmentation permet donc d'isoler des régions à deux dimensions à l'intérieur desquelles les caractéristiques varient peu. On peut distinguer deux types de segmentation d'images suivant le niveau sémantique des descripteurs employés.

[Luc01] présente une étude d'ensemble des techniques de segmentation d'images **bas niveau**. Deux grandes familles d'algorithmes se distinguent, elles sont caractérisées par le traitement appliqué :

1. Segmentation sur un histogramme [Liu94] : cette technique considère que la couleur de la surface d'un objet est constante. L'image est projetée sur un histogramme de descripteurs et les différents objets apparaîtront comme des pics dans cet histogramme. La dispersion des points au sein de chaque pic est due aux effets d'ombrage, au bruit ou au dispositif d'acquisition. Ainsi le problème de segmentation d'une image revient à détecter des pics dans un histogramme. Problème : les régions ainsi isolées ne sont pas compactes dans l'espace.
2. Segmentation sur les régions : cette technique considère la compacité spatiale des régions à découper en plus de l'homogénéité des couleurs. Cette capacité est assurée soit en divisant l'image en petite régions homogènes de taille fixe, puis en les regroupant ("split and merge" [Jai95]), soit en agrandissant progressivement une région de l'image ("region growing" [Ros82]) par agglomération de régions à proximité de la région initiale selon un critère d'homogénéité. Cette dernière technique est principalement utilisée pour détecter une seule région de l'image.
3. Segmentation par détection des contours : cette technique isole des régions homogènes et compactes par détection des contours des régions de l'image. Ceci est réalisé soit en analysant un gradient de couleur sur l'image [Car94], soit en déformant un contour initial par minimisation d'un critère d'énergie ("snakes" [Sap97]).

Aux niveaux supérieurs, l'information utilisée pour segmenter se présente sous la forme de concepts. Cette segmentation est précédée d'un découpage de niveau bas, puis les régions sont annotées par classification du concept choisi. La classification peut être supervisée ou non. Dans le cas supervisé, les valeurs prises par le concept permettent de détecter la présence de certains objets caractéristiques.

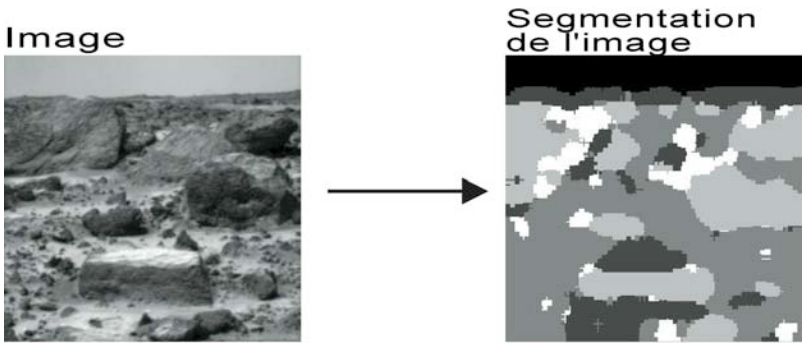

FIG. 2.19 – Structuration d'une image par segmentation de niveau bas.

Les concepts supervisés utilisés sont assez variés : des couleurs (rouge, bleu), des textures (herbes, eau)[Mil04], des formes, des visages[Vio02], des objets[Sch00c]. . ., et dans le cas non-supervisé, la classification regroupe les pixels de l'image (ou des petites régions) en grandes régions homogènes[Pap94]. Les concepts moyens ou hauts non supervisés peuvent être déterminés par un ensemble de descripteurs bas, moyens et hauts variés.

FIG. 2.20 – Structuration d'une image par segmentation de niveau moyen.

Segmentation de l'audio

Le son est un support à une dimension : le temps. La segmentation permet donc d'isoler des segments de sons à l'intérieur desquels les caractéristiques varient peu. Comme pour l'image, deux types de segmentation du son se distinguent suivant le niveau sémantique des descripteurs employés.

Dans sa thèse, S. Rossignol [Ros00] présente une étude des techniques de segmentation **bas niveau** du son. La segmentation bas niveau se fait à l'aide de techniques de seuillage et écrêtage sur les caractéristiques du son. Elle permet l'extraction de phones, de notes [Coi94] ou même de lignes mélodiques [Ghi95, McN96]. Dans le cadre de l'extraction du son pour le démixage [Sch99], il est parfois possible de segmenter le spectrogramme 2D d'un son, comme une image, afin d'isoler les caractéristiques temps fréquences d'un segment particulier. Dans notre système nous n'utiliserons pas de segmentation de ce type pour l'audio. Pour mettre en valeur le contenu sémantique du son, nous préférerons découper selon des caractéristiques de niveau sémantique supérieur.

Pour la segmentation du son, il est souvent plus indiqué de découper selon les valeurs discrètes prises par un **concept**. Le principe est de classer des segments consécutifs de sons dans des classes prédéfinies ou non, à l'avance. Ceci afin d'isoler les segments au sein desquels le concept choisi ne varie pas. Pour les classifications supervisées, il s'agit le plus souvent de concepts simples : segmentation sur le

volume (*silence/non silence*), ou sur le type de son (*parole/musique*) [Ajm02]. En ce qui concerne les cas non-supervisés plusieurs applications ont été développées notamment pour la segmentation en locuteurs [Mei01], ou la segmentation de la structure de morceaux de musique [Pee02a]. Cette classification peut se faire avec ou sans notion de contiguïté temporelle.

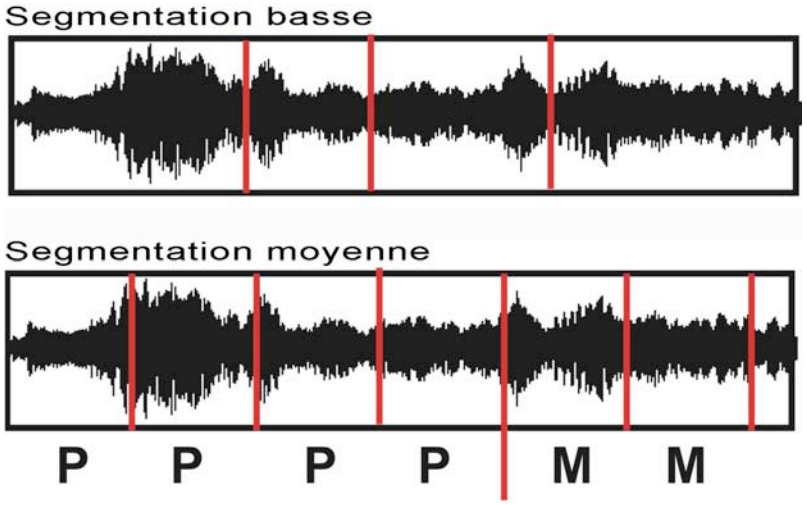

FIG. 2.21 – Segmentation de niveau bas et moyen (*Parole/Musique*) d'un extrait de son.

Segmentation de la vidéo

La vidéo est un support à trois dimensions : deux d'espace et une de temps. La segmentation permet donc d'isoler plusieurs types d'objets. A un instant donné correspond une image dont il est possible d'extraire des zones à deux dimensions. Pour une séquence d'images, la segmentation délimite des régions à trois dimensions au sein du flux vidéo. La plupart du temps, ces segmentations sont réalisées à l'aide de modèles de découpage d'images fixes adaptés pour la vidéo. Ceci est fait en passant de modèles 2D à 3D (« region growing » par exemple) ; ou en utilisant des modèles de suivis (suivi d'objet ou de visage). Dans le cadre de cette thèse nous n'utiliserons pas de segmentation spatiale pour la vidéo, nous nous contenterons de modèles **temporels (1D)**. Comme pour l'image et le son, deux types de segmentation peuvent être distingués suivant le niveau sémantique des descripteurs employés.

Au **niveau bas**, de nombreuses segmentations par seuillage ou écrêtage d'un signal caractéristique de la vidéo ont été présentés dans la littérature. Il s'agit principalement de techniques de détection des frontières de plans, ou de scènes. [Lie01] présente une étude globale de ces algorithmes. Couramment , les caractéristiques choisies décrivent la couleur ou l'intensité lumineuse des images de la vidéo.

FIG. 2.22 – Segmentation de niveau bas des plans d'une vidéo.

Pour la vidéo, il existe très peu de segmentations sur les descripteurs de **niveau supérieur**. Quelques segmentations utilisent des concepts supervisés : la segmentation de programmes TV (publicité, film,

émission de divertissement etc…) [Liu98], la segmentation sur la présence de personnages. Mais, la plupart du temps, il s'agit de concepts dont les valeurs ne sont pas connues à l'avance (classification non supervisée) avec ou sans notion de contiguïté temporelle [Coo02]. Ces algorithmes sont employés afin de découper un film en plans et en scènes.

FIG. 2.23 – Segmentation des scènes d'un film par un concept haut non-supervisé.

2.6.2 Technique de segmentation bas niveau

La segmentation par écrêtage ou seuillage est souvent employée comme première segmentation pour des systèmes plus complexes, comme la segmentation en phone pour la reconnaissance de parole ou la segmentation en plans pour le traitement d'un film. Nous étudierons ici uniquement la segmentation temporelle (1D).

Segmentation par écrêtage et seuillage

La première étape consiste à extraire un grand nombre de caractéristiques ou descripteurs de niveau bas. Ensuite, une fenêtre de taille N et un point de référence sont définis. [Sun00a] montre que ce type de segmentation donne les meilleurs résultats lorsque ce point est au centre de la fenêtre considérée. Une fonction de similarité est définie entre les parties de la fenêtre avant et après le point central. La fenêtre glisse le long des caractéristiques de l'objet avec un pas constant qui correspond au taux de discrétisation de la segmentation. Et une fonction dite d' « observation » est estimée par calcul de similarités entre les deux segments de la fenêtre.

Deux cas de figure sont envisageables pour le calcul de cette fonction. Pour le monomedia, il est préférable de définir une fonction d'observation à une dimension. Par exemple, [Tza99] utilise la distance de Mahanobis entre les descripteurs des deux segments adjacents ou sa dérivée (première ou seconde). Afin de détecter les points de transition, les points de crêtes sont ensuite isolés par seuillage ou par analyse de la dérivée. [Coi94] présente plusieurs techniques pour réaliser cette tâche. Mais, afin de mieux considérer les descripteurs individuellement, il peut être intéressant de définir une fonction d'observation de plus grande dimension, une par media ou même une par descripteur. [Ros00] choisit cette option pour la segmentation de l'audio : l'auteur réalise une étude globale des fonctions de décision pour fusionner les points de coupures détectés sur chaque dimension. Il existe de nombreuses techniques pour regrouper des fonctions de décisions, que l'on abordera dans le cas de la segmentation multimedia.

En pratique, le choix des fonctions d'observation est assez large. Il faut juste garder à l'idée qu'elles sont d'autant plus appropriées qu'elles présentent des pics grands et fins, quand des transitions surviennent et que leurs moyennes et variances restent petites pendant les zones stables.

Segmentation par le critère BIC

Pour exemple, nous décrivons une technique de segmentation par seuillage utilisée par [Iye00] pour la détection du changement de locuteur : la segmentation par le critère BIC [Gus01, Che98, Che03a].

La technique BIC évalue les dissimilarités des descripteurs entre les deux segments adjacents. Les distributions de chaque segment et du segment formé par les deux segments réunis sont modélisées par un modèle de distribution statistique, en général un **modèle gaussien**. Le principe est de comparer la probabilité des deux hypothèses : les deux segments sont issus du même modèle, ou de deux modèles différents. La valeur BIC d'une hypothèse est définie par la probabilité d'obtenir le segment sachant cette hypothèse.

Soit les deux hypothèses : Hyp_1, on suppose que les données ont une distribution uni-gaussienne, Hyp_2 on suppose que les données des segments à gauche et à droite du point central ont des distributions mono-gaussiennes différentes. La valeur $DBIC = BIC(Hyp_2) - BIC(Hyp_2)$ est, ensuite, étudiée. Un point de coupure est placé au maximum de la courbe situé entre deux zéros consécutifs.

Un des avantages du BIC est qu'il n'y a pas de valeur de seuil. [Wac00] montre une amélioration significative de la segmentation par rapport aux techniques métriques similaires. C'est pourquoi [Iye00] l'utilise pour la segmentation de l'audio et de la vidéo de films.

Les techniques de segmentation bas niveau obtiennent de bonnes performances lorsque la sémantique de l'objet à découper est assez simple (e.g. un plan). Mais, de manière générale, il est difficile de trouver des fonctions caractéristiques qui font du sens. C'est pourquoi il est souvent plus approprié de segmenter les objets sur des concepts discrets, afin d'extraire une sémantique bien définie.

2.6.3 Segmentation par classification du contenu

Le principe de la segmentation par classification du contenu est assez simple. Il s'agit de regrouper entre eux des segments temporels d'objets par classification. Lorsque la classe attribuée change, un point de coupure est déterminé. Cela permet d'isoler des segments d'objets au sein desquels la valeur du concept choisi pour la classification reste la même.

Tout d'abord, il est nécessaire de définir les segments à regrouper. Pour cela, il existe deux méthodes distinctes. La première consiste à découper l'objet en segments successifs de taille fixe : par exemple de $3s$ pour le son. La deuxième, suppose une segmentation initiale de niveau inférieur qui a isolé des segments de tailles variables au sein desquelles les caractéristiques liées à cette première segmentation varient peu. C'est le cas pour la segmentation du film en scènes qui est précédée d'une segmentation de niveau bas en plans. Cette étape correspond à une sorte de sous-échantillonnage du signal, qui permet d'isoler la structure basse de l'objet considéré.

La seconde étape consiste à associer à chaque segment ainsi délimité la valeur prise par le concept choisi pour la segmentation. Les choix possibles pour ce concept sont illimités. Il peut être connu ou non a priori : classification supervisée ou non supervisée. Ce peut être un descripteur de niveau sémantique moyen ou haut suivant le niveau d'abstraction de la segmentation.

Classification non supervisée

La méthode la plus simple pour segmenter selon un concept est de pratiquer une classification non supervisée des segments successifs.

De nombreuses approches existent pour estimer, par **apprentissage**, le modèle de concept le plus approprié ($\mathcal{M}_d(\mathrm{T}, \mathrm{X}, \Psi)$) à la description d'un ensemble de points. Le principe est de maximiser la séparation entre les classes de la taxonomie. Ceci est réalisé par un algorithme de regroupement. [Foo03] propose un regroupement par « k-moyenne », mais il existe de nombreuses techniques plus performantes qui

permettent notamment une évaluation robuste du nombre de classes : Adaptive Robust Competition[Sau02]. Dans ce cadre, il est essentiel de bien choisir les descripteurs utilisés pour représenter les objets. Ce choix conditionne, en effet, la détermination d'un concept qui fait sens pour la segmentation désirée. De même, le choix de la distance entre objets est très important pour réaliser un bon regroupement. Comme on l'a vu, de nombreuses distances ont été développées dans la littérature (e.g. euclidienne, Mahanobis). [Jia00] propose d'y inclure un terme de temps afin de contraindre la durée des segments ainsi isolés, mais nous verrons par la suite qu'il existe des techniques plus simples pour cela.

Généralement, ce type de segmentation est utilisé lorsqu'il est impossible d'isoler l'ensemble des valeurs prises par un concept. Notamment dans le cadre de la segmentation d'une conversation en locuteurs où le nombre de locuteurs et le modèle de leurs voix n'est pas connu a priori (concept moyen). De même pour la segmentation en scènes [Jia00] d'une vidéo où le nombre de scènes comme leur modèle est inconnu (concept haut).

Classification supervisée

Cependant, dans le cadre de la segmentation, il est souvent plus simple et plus robuste de choisir un concept et de segmenter l'objet selon ce concept. La classification supervisée associe à un segment de l'objet considéré la valeur du concept choisi pour la segmentation. L'ensemble des valeurs prises par ce concept est connu a priori. Cette technique permet de tracer des frontières autour des segments au sein desquels le concept supervisé ne varie pas.

L'avantage de la segmentation supervisée par rapport à la segmentation non supervisée est qu'elle produit des segments dont la sémantique est connue de l'utilisateur. Ce qui autorise ensuite des **traitements adaptés au contenu** de chaque segment. Cette technique de segmentation a été utilisée dans le cadre de nombreuses applications audio : la segmentation en *parole/musique/bruit* [Wac96, Wan00] est souvent appliquée comme première segmentation du son avant d'autres traitements. Pour cela plusieurs algorithmes de classification supervisée sont applicables, notamment les plus proches voisins [Bag01] ou les SVM [Dee90].

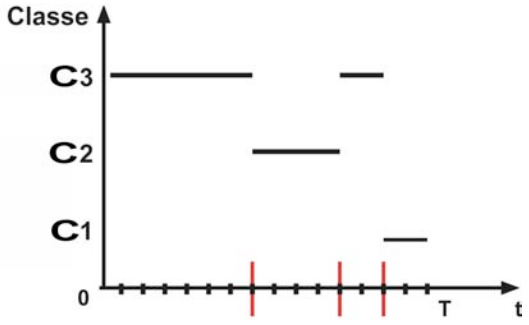

FIG. 2.24 – Segmentation d'un objet par classification d'un concept supervisé à 3 classes.

Globalement les techniques de segmentation par classification (supervisée ou non) apportent de bons résultats. Cependant, elles ne prennent pas en compte la temporalité de l'objet considéré. En effet, il semble approprié d'utiliser les propriétés liées à la succession temporelle des segments.

2.6.4 Prise en compte du temporel

La vidéo et la musique ont une nature spécifique ; ils ne sont pas juste un ensemble d'éléments mais une succession temporelle spécifique d'éléments. Les algorithmes simples de classification ne prennent pas en compte cette spécificité. Plusieurs raffinements de la classification considèrent la temporalité

de l'objet considéré. Le lissage des descripteurs par filtrage moyen assure une certaine continuité des caractéristiques du signal. De même, dans le cas non-supervisé ("clustering"), l'inclusion d'un terme temporel à la distance entre segments favorise la réunion des éléments proches dans le temps. Cependant, nous trouvons plus approprié de formuler cette contrainte en utilisant l'approche des modèles de Markov cachés (MMC) décrite dans l'annexe A.3. Les MMC permettent de déterminer la séquence de classes la plus probable à partir des probabilités d'appartenance des éléments aux classes et des probabilités de transition entre classes. Pour les MMC les valeurs prises par le concept choisi sont appelées états. [Dug96] et [Blu04] présentent un état de l'art des techniques de classification de séquences par MMC.

Les modèles de Markov cachés

Les modèles de Markov cachés sont un outil statistique très puissant pour modéliser des séquences qui peuvent être caractérisées par une **séquence d'états**. Les **MMCs** ont été introduits par Baum et ses collaborateurs dans les années 1960-70. D'abord utilisés en reconnaissance de la parole à partir des années 80 ([Rab78], [Rab93], [Pol97]), ils ont ensuite été appliqués à la reconnaissance de textes manuscrits [Bik99] et à la bioinformatique.

Le principe de ces modèles est de déterminer la séquence la plus probable d'états qui correspond à la suite de segments. Comme pour la classification, ou la segmentation, il existe deux hypothèses distinctes pour ce faire.

Cas supervisé

La segmentation par MMC supervisée a été employée dans de nombreux systèmes de traitement de l'audio et de la vidéo : [Gir99] pour la segmentation de vidéo en 6 classes ; [Iur01] présente un modèle de structures de journaux télévisés à 12 états (fig 2.25) ; [Soh99] segmente l'audio en 2 classes *parole/non parole* ; [Bor98] réalise la segmentation de vidéos en plans. Ce dernier, utilise un modèle MMC à plusieurs états : (*normal, fade, dissolve, cut*). Lorsque l'algorithme détecte un effet de transition dans un segment, il place une coupure au milieu de celui-ci. Ce qui donne de meilleurs résultats que les modèles par seuillage pour la même tâche.

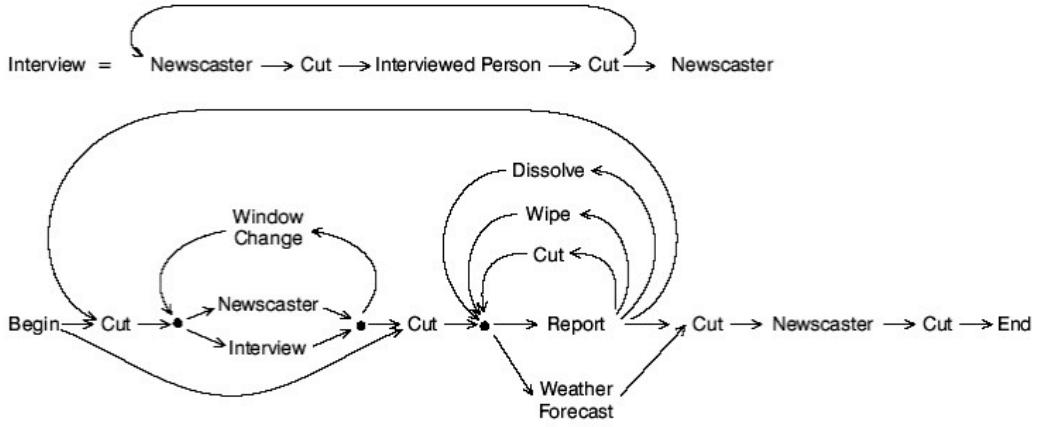

FIG. 2.25 – Segmentation par MMC

Cependant, lorsque les structures isolées ne peuvent pas être décrites par des concepts supervisés, c'est le cas pour la segmentation d'un film par exemple, où le modèle des scènes ne peut pas être connu à l'avance, il est nécessaire d'apprendre le modèle MMC.

Cas non supervisé

On est ici dans le cas où la taxonomie de la segmentation n'est pas connue a priori. De même pour les matrices de transition et d'émission. Comme il a été vu, pour la classification, le problème consiste à trouver la taxonomie et son modèle de classification qui correspondent le mieux à la structure des données.

Ce type de segmentation non supervisée par MMC a été employé pour de nombreuses applications où l'information contenue dans les objets à segmenter n'est pas connue a priori. Elle permet sans connaissance de détecter les structures du contenu d'un objet. Notamment pour la segmentation d'un dialogue en locuteurs [Mei01], ou pour l'extraction de la structure et le résumé d'un morceau de musique [Bre98]. En pratique, cette segmentation par rapport à une classification simple permet de lisser la variation du concept choisi, comme le montre la figure A.3. En effet, pour les MMC la probabilité de rester à un état décroît de façon exponentielle dans le temps [Gop98]. Cela a pour effet de diminuer la sur-segmentation en présence de segments bruités et de favoriser la segmentation lorsqu'un segment est trop long.

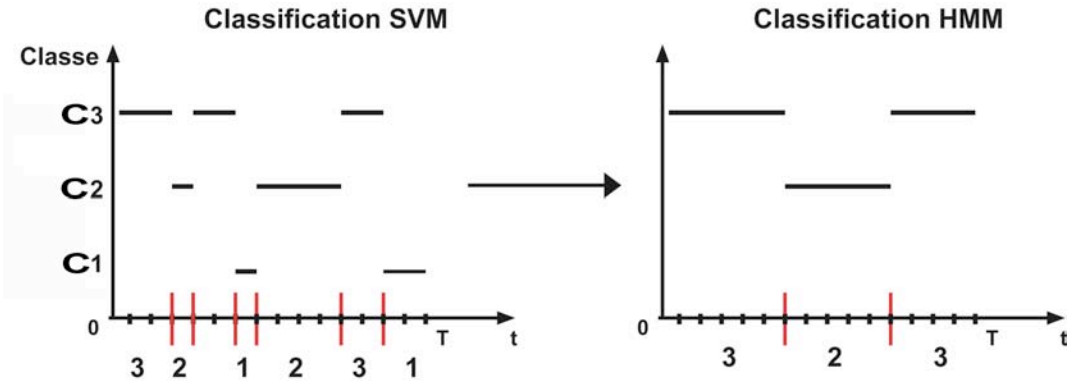

FIG. 2.26 – Comparaison des modèles de segmentation SVM et MMC.

2.6.5 Segmentation de films

Etablir la structure cinématographique de la vidéo trouve sa justification dans le fait que, comme la table des matières d'un livre, elle fournit un accès direct aux différents composants d'une vidéo. Il est maintenant largement accepté que les films sont **hiérarchiquement structurés en scènes, plans et images**. Une telle structure reflète sinon le processus de création de la vidéo, les différentes étapes du montage. Segmenter un film pose trois questions, celle de la délimitation des séquences, celle de leur succession, celle de leur structure interne.

L'appréciation des structures narratives constitue une étape cruciale de l'interprétation du film. En proposer une segmentation, c'est déjà dépasser le stade de la description simple. La plupart des **systèmes de segmentation de films** proposent un modèle de structure hiérarchique comportant deux niveaux : le plan et la scène. Les plans sont définis comme des séquences continues d'images prises sans arrêter la caméra. Les scènes sont définies comme des suites de plans contigus qui sont sémantiquement reliés, bien que certains systèmes [Li02] n'imposent pas cette condition.

La relation entre deux segments successifs est de deux types : temporelle (rapport de chronologie ou de simultanéité), et causale (rapport de cause à effet : le premier provoque l'effet exprimé dans le second). L'analyse de ces relations apporte plusieurs informations sur la structure d'un film : une scène est caractérisée par une unité de temps et de lieu. La succession caractéristique de visages différents indique un ensemble de plans de dialogue.

Plusieurs **effets cinématographiques** peuvent indiquer le changement de séquence. Le raccord par coupure franche ou « cut » entre deux moments de film n'est pas exceptionnel. D'autres éléments du

film peuvent marquer la séparation : fondu au noir, fondu enchaîné, effet de flou, utilisation de volets, de rideau, fermeture d'iris, blancs, intertitres. Il est aussi possible de trouver comme signe de ponctuation certaines caractéristiques sonores : l'utilisation de la voix-off, d'un leitmotiv musical. Dans les faits, ce type de coupure permet de détecter les frontières entre plans. En ce qui concerne les segmentations de granularité temporelle plus importante, la ponctuation peut influencer la décision, mais le plus souvent les plans sont regroupés en séquences plus longues par similarité de leur descriptions.

Le Plan

L'apparition du concept de plan est liée à l'invention de la caméra. Il est défini comme la **séquence continue** la plus longue issue d'une prise. Une prise étant ce que la camera a capturé pendant une course ininterrompue. La détection des plans est la plus basique des tâches de segmentation vidéo, puisqu'elle est intrinsèquement liée à la façon dont le film a été produit. Ainsi, la segmentation en plans est souvent la première étape de l'extraction de structure vidéo. De façon générale elle est réalisée par un modèle de niveau sémantique bas.

Après le travail séminal de Nagasaka et Tanaka en 1991 [Nag91], de nombreuses études ont été menées sur la détection automatique de plans dans les vidéos [Zha93, Qué99]. L'enjeu est d'essayer de détecter les différents types de transitions de plans qui peuvent survenir dans une vidéo. On trouve dans TREC 2001 [Sme01] la comparaison de différentes approches de segmentation temporelle. Si la détection des coupes franches entre deux plans est assez aisée, la détection des limites entre deux plans liés par des effets spéciaux de la caméra, comme un fondu enchaîné par exemple, est nettement plus difficile. Dans [Sar97], les raccords «cut» de vidéo sont déterminés par écrêtage de la différence d'énergie entre deux segments successifs. Les ponctuations plus longues sont détectées de la même façon avec des segments plus longs. Dans [Hua98], les auteurs utilisent les variations de couleurs et de mouvements au sein de l'image. La corrélation de phase calculée entre deux images est utilisée pour détecter les coupures dans le mouvement. Pour les changements de couleur, les histogrammes de couleur de chaque paire d'images successives sont comparés. Les résultats de la détection des coupures de couleur et de mouvement sont comparés. Une fonction de décision est appliquée pour fusionner les points de coupures ainsi détectés.

La segmentation en plans n'est plus aujourd'hui un sujet de recherche ; en effet, de nombreux algorithmes donnent de très bons résultats. Notamment l'algorithme développé au CEA par P. Josserand [Jos00] ; qui offre un pourcentage appréciable de segmentations justes.

La Scène

Les scènes (aussi appelées unités d'histoire) sont un concept bien plus vieux que le cinéma, puisque issu du théâtre. Traditionnellement, dans le cadre du film, une scène est une séquence de plans qui est temporellement et spatialement cohérente dans le vrai monde. La segmentation automatique d'une vidéo en scènes est une opération complexe. Ceci est dû à deux raisons : premièrement, c'est une tâche très subjective qui dépend du conditionnement culturel humain, l'entraînement professionnel et l'intuition ; deuxièmement, elle se fixe sur les actions du vrai monde et les configurations temporelles et spatiales d'objets et de personnes. Elle exige donc la capacité d'extraire le sens physique des images et, on l'a vu, il s'agit d'une tâche extrêmement difficile pour un ordinateur.

Pour rester dans le cadre de la **fusion des media**, nous examinerons ici les méthodes de détection de frontières de scènes qui combinent l'audio et l'information visuelle pour atteindre leur but. Il devrait être noté cependant que la majorité des algorithmes de détection de frontière de scènes compte uniquement sur l'information vidéo.

Comme pour la segmentation d'une vidéo en plans, la segmentation en scènes est exactement définie et caractérisée par le **contenu** choisi pour segmenter. Le choix de cette information est dictée par les caractéristiques classiques des scènes :

– Une scène se définit par une unité de lieu et de temps, au sein de la séquence des plans ; on observe donc une certaine constance dans la luminosité, les couleurs, ou les lignes de l'image ;
– les coupures de scènes pour l'audio et l'image coïncident, ainsi la fin (ou le début) d'une scène correspond à une coupure de clip (pour l'audio), et une coupure de plan (pour l'image) ;
– certaines ponctuations filmiques complexes indiquent une coupure de scène (e.g. fade, voix off).

Le travail introduit dans [Sun00b] utilise un modèle de mémoire fini pour segmenter de manière indépendante l'audio et les données vidéo en scènes. Deux fenêtres d'ambiguïté sont utilisées pour fusionner les scènes audio et vidéo. Les deux algorithmes de segmentation sont de niveau sémantique bas (écrêtage). L'algorithme audio de segmentation utilise les corrélations des enveloppes de caractéristiques audio. L'algorithme de segmentation image utilise les corrélations entre images "moyennes" de plans. Les frontières de scènes dans les deux cas sont déterminées par la détection des minimums locaux de corrélation. Les auteurs unifient les segments résultants en utilisant un algorithme des plus proches voisins. Le résultat est ensuite affiné en utilisant la distribution de l'alignement temporel appris à partir de vidéos annotées.

Dans [Sar97], l'audio est distingué par un concept à quatre classes choisies au préalable : **silence, parole, musique et bruit**. Cette information est ultérieurement combinée à la valeur de probabilité de détection d'une coupure visuelle, qui a servi à segmenter la vidéo en plans. L'information qui concerne le son et les silences est utilisée pour indiquer un changement de scène possible. Une coupure de scène est détectée si les caractéristiques du locuteur ou de la musique changent (ou si un silence est détecté) près d'un endroit où une coupure de plans a été signalée. Il s'agit là, d'un modèle de niveau sémantique hybride : bas pour l'image et moyen pour le son.

Les auteurs dans [Yos01] proposent une méthode pour la détection de frontière de scènes en exploitant les caractéristiques audio et vidéo. L'extraction du **bruit de fond** et sa classification, et l'extraction de caractéristiques visuelles réunissent une séquence de plans au sein d'une seule scène. Des frontières de scènes candidates sont extraites par détection d'effets visuels fréquemment utilisés dans les changements de scène tels que les fondus. D'autres frontières sont aussi extraites des caractéristiques audio en détectant « les coupures audio » basées sur la classification de *bruit/musique de fond*. Les coupures audio localisées à proximité (au sens temporel) d'un changement de l'image sont analysées afin de déterminer un changement de scène.

Dans [Jia00], les segments avec et sans parole sont détectés. Les segments sans parole sont classés en musique et son écologique. Cette classification est basée sur la périodicité audio et d'autres caractéristiques. Les coupures audio sont détectées en utilisant des segments d'une seconde. La position de ces coupures est comparée aux frontières des plans vidéo. Toutes les frontières de plan situées dans un intervalle de moins d'une seconde d'une coupure audio sont choisis comme candidats de scène. Ensuite, de façon séquentielle, un algorithme de corrélation de couleurs est utilisé pour rassembler les plans, ce qui évite la sous-segmentation. Ce modèle est de niveau sémantique moyen, la classification est non-supervisée pour l'image et supervisée pour le son.

Dans [Kyp04] les segments audio sont projetés sur certains vecteurs propres afin d'isoler les changements causés par les **variations du bruit de fond**. Les auteurs utilisent la distance dans cet espace entre les segments audio et plusieurs segments de référence choisis. Les indications de changement de scène de la piste audio sont identifiées en traitant ce vecteur de distance. L'information vidéo est utilisée pour aligner les indications de changement de scène audio avec les changements de plans avoisinants. De plus, les effets vidéo sont identifiés et utilisés de façon indépendante afin de détecter les changements de scène.

Pour résumer, ces systèmes fonctionnent en 4 étapes distinctes :

1. segmentation de la vidéo en plan,

2. segmentation du son en segments homogènes,
3. une coupure candidate est détectée lorsque les détections coïncident,
4. élimination de certains points de coupure.

Notons qu'à part quelques recherches [Sun00c, Sun00b] la plupart des auteurs n'utilisent pas les **similarités** qui existent au sein d'une scène audiovisuelle entre les plans. Or l'unité de lieu et d'action au sein d'une scène entraîne des similitudes entre les descriptions des plans de cette scène. Nous appliquerons donc le modèle MMC non-supervisé, décrit dans 2.6.4, pour la segmentation des films. De plus, la hiérarchie habituelle plan-scène, ne caractérise pas entièrement l'ensemble des structures de narration présentes dans les films de cinéma. Nous introduirons deux niveaux supplémentaires. Les **actes** ou groupes de scènes sont des séquences de plusieurs scènes, il s'agit des grands blocs narratifs du film. Les **groupes de plans** sont des séquences de plans à l'intérieur de scènes. Par exemple une scène de dialogue débute par quelques plans d'introduction du lieu et des personnages, auquel succède un groupe de plans composé d'une alternance de plans sur les personnages. La segmentation hiérarchique de notre système devra donc prendre en compte les niveaux narratifs courants : scènes et plans, mais aussi les niveaux introduits ici : actes et groupes de plans.

2.7 Résumé des méthodes et objectifs

Résumons maintenant les buts de ce travail au vu des tâches à accomplir et des résultats des algorithmes existants.

2.7.1 Utilisation des modèles de contenu

Notre première constatation sur l'état de l'art est que la performance de l'indexation du contenu dépend de la quantité et de la qualité de l'information présente au sein des modèles de traitement de données.

Le tableau 2.27 qui présente les performances des modèles de fusion et de classification appellent plusieurs remarques. Premièrement, en ce qui concerne les descripteurs bas, les performances obtenues par les modèles de traitement de données sont bien définies, mais la validité de ces techniques reste à montrer dans le cadre du modèle de contenu choisi. Deuxièmement, en ce qui concerne les descripteurs moyens : d'une part, la normalisation des scores obtenus par les concepts est indispensable afin de ne pas favoriser certains descripteurs ; d'autre part, le modèle de classification naïf de Bayes semble moins performant que les modèles prenant en compte les dépendances entre concepts, notamment les « Support Vector Machine » (SVM). Troisièmement, en ce qui concerne la fusion des descripteurs bas/moyen et audio/vidéo : le modèle multinet semble trop complexe et coûteux pour le système que nous développons. En revanche, les fusions «Early» et «Late» montrent des performances différentes : le choix du modèle dépendra des caractéristiques de l'application.

Nous proposons donc comme but essentiel de ce travail l'étude de l'apport des modèles de traitement de données spécifiques, appris et structurés, que nous appellerons modèles de contenu pour l'indexation. Tout d'abord, nous identifierons un ensemble de descripteurs et de concepts (audio, vidéo et multimedia) pertinents pour la caractérisation du contenu des films. Puis nous développerons le modèle de contenu correspondant dont les caractéristiques permettront d'explorer l'ensemble des problématiques de fusion et de classification.

Des études semblables ont déjà été réalisées pour l'indexation monomedia de photographies ou d'extraits musicaux et pour l'indexation multimedia de pages du web (image/texte) ou l'identification de personnes (image/son). Les méthodes de fusion de descripteurs (bas/moyen et multimedia) pour la classification se sont montrées performantes. Mais les résultats obtenus sont souvent incomplets, la fusion

« early » (concaténation) des descripteurs bas et moyen n'a pas été testée. De même, les performances obtenues pour la fusion de media concernent certaines applications particulières. Nous chercherons donc à confirmer et étendre ces résultats en construisant un modèle hiérarchique de contenu adapté au film. Nous effectuerons des tests de classification de chacun des concepts de ce modèle. Enfin, nous étudierons en particulier les performances de classification de plusieurs concepts audio (voix/musique, identification d'une voix), image et multimedia (identification du lieu).

2.7.2 Utilisation des modèles de structure

Notre deuxième constatation sur l'état de l'art de l'indexation de la structure est que les performances dépendent du modèle de segmentation et surtout du choix du modèle de contenu utilisé pour segmenter.

Le tableau 2.28, qui montre les performances des méthodes existantes selon les descripteurs utilisés, mène à deux remarques.

Premièrement, les modèles d'écrêtage semblent moins performants que les modèles de classification, eux-mêmes moins performants que les modèles MMC.

Deuxièmement, les résultats obtenus par les modèles de segmentation des plans et des clips sont bons. En ce qui concerne la segmentation des scènes multimedia, les performances sont limitées par les problèmes d'écrêtage et de fusion des décisions contradictoires.

Nous proposons donc, comme second but de ce travail, l'étude des performances de structuration de films obtenues par un modèle de segmentation prenant en compte plusieurs caractéristiques : segmentation MMC, fusion des media, descripteurs moyens.

Des modèles semblables ont déjà été testés sur des films de cinéma ou des vidéos. Les systèmes développés se sont révélés performants. Mais ces modèles n'utilisent pas l'information apportée par les concepts images. Une attention particulière sera donc portée afin d'identifier l'apport des différents groupes de descripteurs à la segmentation et ceci pour chaque niveau de la structure. De plus, seul deux niveaux de granularité (plans/clips et scènes) sont considérés pour modéliser la structure.

Nous n'étudierons pas les modèles de fusion des points de coupures pour la segmentation multimedia. Nous choisirons un modèle simple fondé sur les plus proches voisins qui sera décrit par la suite. Cependant, les performances de segmentation pourraient être améliorées par l'emploi de techniques plus complexes (voir notamment [Coi94]).

Descripteurs	Tâche	Modèles	Section	Performances
BAS	Normalisation	Moyenne nulle	1.3.1	Réduit le bruit statique de convolution
		Ecart type à 1	1.3.1	Egalise le poids des descripteurs Légère perte d'information.
	Elimination	PCA …	1.3.3	Sélectionne les descripteurs les moins corrélés. Ce ne sont pas forcément les plus pertinents.
		LDA IM	1.3.3	Sélectionne les descripteurs les plus pertinents pour la classification Eliminer des descripteurs entraîne une perte d'information.
	Transformation	PCA, ICA	1.3.2	Rend les descripteurs indépendants Ne prend pas en compte la structure de classes
		LDA	1.3.2	Augmente la séparabilité entre classes
	Classification	Naïf de Bayes	1.2.2	Bonnes performances Ne prend pas en compte la dépendance des descripteurs
		SVM, KNN, GMM	1.2.3 1.2.2	Bonnes performances. Un plus pour les SVM
MOYENS	Normalisation	Pseudo probabilités	1.4.2	Egalise les scores de classification. Perte d'information
	Classification	Naïf de Bayes	1.2.2	Bonnes performances Ne prends pas en compte la dépendance entre concepts.
		SVM, KNN, GMM	1.2.3	Bonnes performances. Un plus pour les SVM.
HAUTS	Fusion des descripteurs - moyen et bas - image et son	Early fusion	1.2	Prend en compte les dépendances entre descripteurs. Coûteuse en temps de calcul.
		Late fusion	1.4.4	Apprentissage simplifié. Ne prend pas en compte la dépendance entre descripteurs de groupes différents
		multinet	1.5.2	Prend en compte toutes les dépendances Très coûteux en temps de calcul.

FIG. 2.27 – Classement des méthodes d'indexation du contenu selon les modèles utilisés.

Contenu	Niveau	Modèles	Section	Performances
Descripteurs bas image	Plans	Écrêtage	1.7.2 1.7.6.a	Bonne performances. Sujet au bruit : e.g. gros plans. Problèmes lors des effets de transitions visuels.
		Clustering	1.7.3	Bonnes performances, problème lors des effets de transitions.
	Scènes	Écrêtage	1.7.2 1.7.6.b	Segmentation limitée par le choix d'un seuil fixe pour l'intégralité du film.
		Clustering	1.7.3.b 1.7.6.b	Performances moyennes. Est fortement sujet au bruit.
		HMM non-supervisé	1.7.4.b	Bonnes performances. Effet de lissage par rapport au clustering.
Descripteurs bas audio	Clips	Classification	1.7.3.a	Bonnes performances. Mais est sujet au bruit.
		HMM supervisé	1.7.4.a	Presque parfait
	Scènes	Écrêtage	1.7.2	Problèmes liés au choix du seuil de coupure. Ne prends pas en compte certains effets de transition audio : e.g. silences.
Descripteurs bas image et audio	Scènes	Écrêtage et fusion des points de coupure image et son	1.7.2 1.7.6.c	Segmentation moyenne, car limitée par les problèmes liés aux techniques d'écrêtage.
Descripteurs bas image et descripteurs moyens audio	Scènes	Ecrêtage des descripteurs images et fusion avec les limites des clips	1.7.2 1.7.3.a 1.7.6.c	Bonnes segmentations, mais écrêtages et problèmes de fusion des points de coupures lorsque les deux médias sont en désaccord.

FIG. 2.28 – Classement des méthodes d'indexation de la stucture selon les modèles utilisés.

Chapitre 3

Choix d'un modèle descriptif

Un des but de cette thèse est de mettre en évidence la nécessité et l'efficacité de la fusion des données dans le cadre d'applications liées à l'analyse de films de cinéma. Elles sont de plusieurs types : classification, détection de structure, résumé, recherche ; et toutes peuvent donner lieu à des fusions de descripteurs.

Dans ce chapitre, nous allons définir un cadre matériel permettant de modéliser les structures spatio-temporelles des films et le contenu de ces structures aux paragraphes 3.1 et 3.2. Ce cadre permet d'appliquer un modèle de représentation commun appelé index pour les différentes applications de l'analyse de film. Nous montrons enfin dans le paragraphe 3.3 comment le contenu du film influence la détection de la structure.

3.1 La structure temporelle du film

Dans le complexe espace-temps (ou continuité espace/durée) qui charpente l'univers cinématographique, il est clair maintenant que c'est le temps et lui seul, qui structure de manière fondamentale et déterminante tout récit cinématographique, l'espace n'étant jamais qu'un cadre de référence secondaire et annexe. C'est donc par rapport au traitement qu'elle fait subir au **temps** que doit être analysée la construction d'un film.

L'analyse des structures du récit filmique s'appuie sur le **structuralisme** : théorie du milieu du 20ème siècle qui considère que les individus sont marqués par des structures inconscientes déterminées par leur culture, leur langue et par la société où ils vivent et que ces structures se retrouvent à différents niveaux de la vie dans l'organisation sociale. Claude Lévy-Strauss travaille sur des récits mythiques [LS55] et tend à mettre en évidence qu'il existe des points communs dans la structure de différentes productions signifiantes (structures des mythes, structures du langage, structures sociales...).

L'image et le son constituent les éléments de base du langage cinématographique. Ils sont la matière première filmique et déjà cependant une réalité particulièrement complexe. Leur genèse est en effet marquée par une ambivalence profonde : elle est le produit de l'activité automatique d'appareils techniques capables de reproduire exactement et objectivement la réalité qui lui est présentée, mais en même temps cette activité est dirigée dans le sens précis voulu par le réalisateur. C'est pourquoi, dans la suite nous considèrerons deux types de structure : la **structure matérielle** liée au processus de création technique du film et la **structure narrative** liée à la nécessité d'organiser les images et le son dans certaines conditions d'ordre et de durée afin de révéler « l'histoire » racontée par le film.

3.1.1 Structure matérielle du film

Mises de côté les considérations de sujet, d'académisme et d'esthétique, un film est la diffusion simultanée d'images et de son.

Signal image

Le signal image d'un film est une succession d'**images fixes** (figure 3.1) rectangulaires dont la proportion (16/9) est dictée par les caractéristiques du champ de vision humain plus large latéralement que verticalement.

FIG. 3.1 – Image tirée d'un film

On a longtemps cru que le phénomène de persistance rétinienne permettait d'expliquer pourquoi l'on ressent la succession d'images fixes d'un film comme des scènes en mouvement. Cette explication de l'illusion du mouvement au cinéma a cependant été rejetée par les psychologues pour plusieurs raisons que l'on n'explicitera pas ici (voir [Cer]). L'illusion du mouvement au cinéma serait donc produite par un autre phénomène qu'on appelle l'effet Phi. Celui-ci se manifeste dès que deux images légèrement décalées dans l'espace sont présentées rapidement l'une à la suite de l'autre. Notre cerveau y voit alors automatiquement un mouvement, résultat du travail d'intégration des champs récepteurs des cellules rétiniennes et des différentes aires corticales visuelles impliquées dans la détection et l'orientation du mouvement. Quant à la persistance rétinienne, elle s'est plutôt vue attribuer un rôle de réduction de l'effet de scintillement de l'image cinématographique causé par l'ouverture et la fermeture de l'obturateur du projecteur 48 fois par seconde. Mais même cette fonction a été remise en question.

Les images sont capturées par la caméra, elle est donc « l'agent actif d'enregistrement de la réalité matérielle et de création de la réalité filmique ». Les limitations de celle-ci (nécessité d'interrompre la course de la bobine) et les besoins de la narration (un film n'est pas en temps réel) entraînent l'apparition du deuxième niveau de la structure visuelle : **le plan** (figure 3.2).

FIG. 3.2 – Plan de film

Techniquement parlant d'abord, un plan, est du point de vue du tournage, le fragment de pellicule impressionnée entre le moment ou le moteur de la caméra est mis en route et celui où il est stoppé ; du point de vue du monteur, le morceau de film entre deux coups de ciseaux puis entre deux «collures» ; du point de vue du spectateur enfin, le morceau de film entre deux raccords.

Une définition psychologique et esthétique du plan est beaucoup plus délicate mais disons que «le plan est une totalité dynamique en devenir qui contient en elle sa négation et son dépassement dialectique», c'est-à-dire qu'en incluant un manque, un appel, une tension esthétique ou dramatique, il suscite

le plan suivant qui l'accomplira en l'intégrant visuellement et psychologiquement.

De manière générale un film comporte plusieurs plans, sauf certains cas pathologiques. Ainsi dans «La Corde», Hitchcock à poussé la simplification du montage à un degré indépassable puisque le film ne comporte qu'un seul plan par bobine et même, du point de vue du spectateur, un seul plan pour tout le film, les raccords de bobines étant pratiquement invisibles puisqu'ils ont lieu sur un fond obscur (le dos d'un personnage, un coffre, un mur). Dans les années 50, un film « normal » comporte environ de 500 à 700 plans. Aujourd'hui, quand un film contient moins de 1000 plans, il est considéré comme plutôt lent.

Signal sonore

On sait que le cinéma devint sonore puis parlant, un peu par hasard, en 1926 parce qu'une société de production américaine, la Warner, se trouvait au bord de la faillite et qu'elle tenta, comme une solution désespérée, cet essai devant lequel les autres firmes reculaient par crainte d'un échec commercial. Le public fit aussitôt un accueil enthousiaste à ce nouveau genre de films, tandis qu'un grand nombre de personnalités du cinéma parmi les plus grandes (Chaplin par exemple) exprimaient leur scepticisme ou leur hostilité. Il est aujourd'hui parfaitement admis que la venue du « parlant » a profondément bouleversé l'esthétique du cinéma.

La définition de la structure matérielle du son d'un film est un peu plus complexe que pour l'image. La bande son se caractérise par sa continuité alors que la bande image du film est une suite de fragments. Elle rétablit donc en quelque sorte la continuité à la fois au niveau de la simple perception et à celui de la sensation esthétique. La bande sonore est en effet, par nature et par nécessité, bien moins fragmentée que l'image : elle est en général relativement indépendante du montage visuel et beaucoup plus conforme au « réalisme » en ce qui concerne l'environnement sonore. Par ailleurs, le rôle de la musique est primordial comme facteur de **continuité** sonore et dramatique.

Cependant de nombreuses études ont montré l'efficacité et la justesse d'un modèle de structuration matérielle de la bande son en « frames » et « clips », de la même manière que la bande image est composée d'images et de plans.

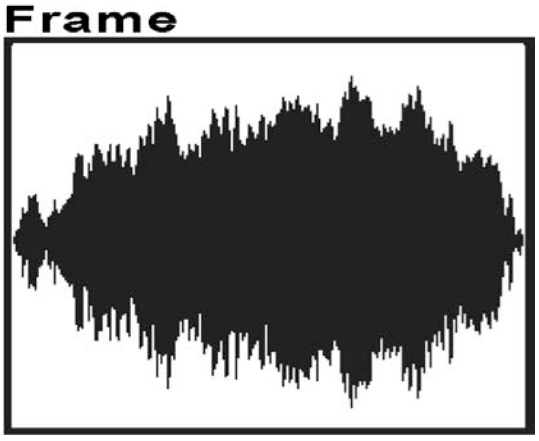

FIG. 3.3 – Frame audio

Une **frame** (figure 3.3) de son (comme une image) correspond au signal audio perçu par un spectateur pendant un intervalle de temps fixe et de taille réduite. Le choix de la taille est un compromis entre la nécessité d'une description précise du signal et le besoin d'extraire du sens de ces frames. Ainsi, la durée globalement admise pour celles-ci est de 3s. De façon générale, le son peut être considéré comme homogène sur cet intervalle, c'est-à-dire que ses caractéristiques physiques et sémantiques varient peu. Dans le cas contraire, cette variation brusque a une signification précise et se retrouve par l'analyse du contenu des frames situées avant et après la frame considérée.

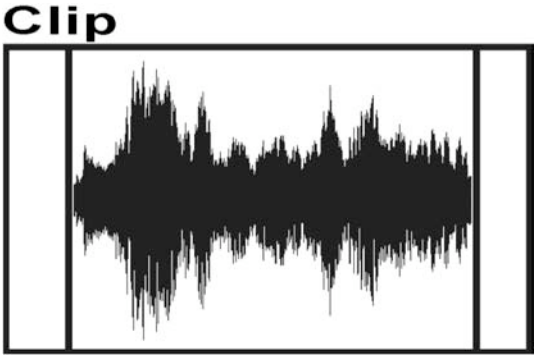

FIG. 3.4 – Clip audio

En ce qui concerne le niveau supérieur de la structure matérielle pour le signal audio, la définition devient assez complexe. En effet, la continuité du son empêche une définition du type plan pour l'image. Cependant, plusieurs auteurs définissent le niveau de structure « **clip** » (figure 3.4), notamment pour le traitement de vidéos personnelles ou de documentaires. Un clip est un ensemble continu de frames audio homogènes selon certaines catégories spécifiques du son.

3.1.2 Premier niveau de la segmentation de l'audio : les clips

Il est logique de répartir les phénomènes sonores en trois grandes catégories, la première comprenant tous les bruits quels qu'ils soient, la seconde étant réservée à la musique non déterminée par un élément de l'action, et la troisième à la parole.

Parole

L'utilisation normale de la parole permet de supprimer cette plaie du cinéma muet que sont les intertitres. Elle libère dans une certaine mesure l'image de son **rôle explicatif** et lui permet ainsi de se consacrer à son rôle expressif, en rendant inutile la représentation visuelle de choses qui peuvent être dites ou, mieux, évoquées. Enfin la voix-off ouvre au cinéma le riche domaine de la psychologie en profondeur en rendant possible l'extériorisation des pensées les plus intimes (monologue intérieur). Ainsi l'utilisation d'une transcription des paroles (obtenues des sous-titres ou d'un traitement automatique) pourrait permettre d'indexer automatiquement des concepts simples (objets, lieux) ou plus abstraits (sentiments).

Musique

A un niveau élémentaire, c'est-à-dire en **accompagnement** d'effets, de scènes ou de séquences plus longues, la musique peut être appelée à jouer plusieurs rôles :

- **Rythmique**. Remplacement ou sublimation d'un bruit réel, mise en relief d'un mouvement ou d'un rythme visuel. Il y a contrepoint musique-image sur le plan du mouvement et du rythme, correspondance métrique exacte entre le rythme visuel et sonore. Le rôle que joue la musique ici lui est particulièrement approprié, mais il est assez limité et en tout état de cause assez peu fécond.
- **Dramatique** . La musique, en créant l'ambiance, intervient comme contrepoint psychologique en vue de fournir au spectateur un élément utile à la compréhension de la « tonalité humaine » du film. Cette conception de son rôle est évidemment la plus répandue.
- **Lyrique**. La musique peut contribuer à renforcer puissamment l'importance et la densité dramatique d'un moment ou d'un acte en lui donnant une dimension lyrique qu'elle est spécifiquement propre à engendrer.

Bruits environnementaux

En ce qui concerne les bruits, nous pouvons distinguer :

- **Les bruits naturels**. Tous les phénomènes sonores perçus dans la nature vierge (bruits du vent, du tonnerre, de la pluie, des vagues, de l'eau courante, cris des animaux, chants des oiseaux, etc. . .)
- **Les bruits humains**. Dans lesquels il faut différencier les bruits mécaniques (machines autos, locomotives, avions ; bruits de rue, d'usines, de gares, de ports) ; les paroles-bruit : c'est le fond sonore humain, le son des paroles fait partie intégrante de l'atmosphère authentique d'un film : il lui donne cette « collaboration musicale » dont parlait René Clair ; enfin la musique-bruit : celle qui est produite par un poste de radio : elle est le plus souvent un simple fond sonore mais peut prendre une valeur de symbole.

Les bruits tel qu'il sont décrits ici, sont utilisés de façon « réaliste », c'est-à-dire conformes à la réalité : on n'entendra que les sons produits par des êtres ou des choses apparaissant sur l'écran ou connus pour se trouver à proximité, sans qu'il y ait aucune volonté d'expression particulière dans la juxtaposition image/son. Pourtant le son, si réaliste qu'il soit, est rarement utilisé de façon brute : « Au début du film sonore, on enregistrait à peu près tous les sons que pouvait capter le microphone. On remarqua bientôt que la reproduction directe de la réalité donnait une impression aussi peu réelle que possible et que les sons devaient être « choisis » au même titre que les images »[Cla51]. Souvent en effet nous regardons quelque chose et nous prêtons l'oreille à un son venant d'ailleurs ; ou bien nous sommes trop absorbés pour percevoir les sons qui arrivent à nos oreilles : pour ces raisons, un continuel synchronisme, loin d'être réaliste, produira un effet anti-naturel.

Ainsi la plupart des sons d'un film (dialogue, bruitage et musique) sont enregistrés en studio ou sur le plateau de tournage de manière à isoler chaque **source** individuellement. Ils sont ensuite réunis par montage sur la bande audio du film. Ceci nous fait penser qu'un modèle de structure par émetteur (ou source) est tout à fait approprié pour le cinéma. Ce modèle pourrait être représenté par des strates, comme nous les avons définies précédemment, chaque strate correspondant à un émetteur en particulier. Cependant, il nous est impossible d'envisager ce modèle puisque nous ne disposons pas de moyen pour séparer les sources. En effet, contrairement à l'image, les objets de base sont superposées par addition dans le temps (musique-parole par exemple). C'est pourquoi, il est nécessaire de choisir un modèle adapté à nos possibilités. Afin d'extraire les **clips**, une segmentation fondée sur la classification des états

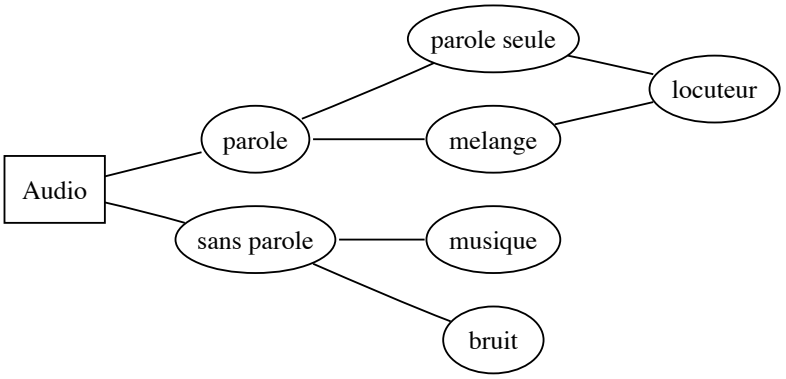

FIG. 3.5 – Modèle hiérarchique d'états pour la segmentation MMC des clips audio

des frames audio sera introduite. Celle-ci est réalisée par la combinaison d'algorithmes MMC supervisés et non supervisés. Le modèle de classification hiérarchique global est présenté dans la figure 3.5. Les classifications Parole/Sans parole, Parole seule/mélange et musique/bruit sont supervisées ; alors que la segmentation en locuteur est une classification non-supervisée apprise sur un segment de parole. Cette

segmentation permet d'extraire les éléments atomiques de la structure audio : les clips. Les segments ainsi obtenus sont homogènes pour les catégories descriptives du son.

3.1.3 Structure narrative du film

L'organisation des plans et des clips d'un film dans certaines conditions d'ordre et de durée est appelée montage. Il correspond à la progression dramatique du film. Au-delà de la création du mouvement (apparence de la vie), de la création du rythme (rapport de longueur des plans et des clips), il est le support de l'organisation narrative du récit : il crée l'histoire. C'est le rôle le plus important du montage, du moins lorsqu'il remplit un but expressif et pas seulement descriptif : il consiste alors « à rapprocher des éléments divers pris dans la masse du réel et à faire surgir un sens nouveau de leur confrontation ». "L'attitude consciemment créatrice à l'égard du phénomène à représenter commence donc au moment où la coexistence indépendante des phénomènes se défait, et qu'à sa place s'institue une corrélation causale de ses éléments, dictée par l'attitude à l'égard du phénomène, attitude dictée par l'univers mental de l'auteur" [Eis58].

Il est largement accepté aujourd'hui que les films (comme les livres) organisent le déroulement de la narration selon une **structure hiérarchique** : à un niveau donné, un élément structurel inclut d'autres éléments du niveau inférieur. Pour des raisons de simplicité, nous ne prenons pas en compte, ici, les éventuelles relations « circulaires » qui peuvent exister au sein de certains films.

Dans le cadre de l'analyse de films, de nombreux modèles de structures de ce type ont été présentés. Christian Metz dans [Met68] fait une proposition de segmentation du film. Il effectue le relevé d'un certain nombre d'agencements et propose de classer ces agencements sous la forme d'un tableau. Linguiste, Metz nomme son tableau : la grande syntagmatique de la bande image. En linguistique, le syntagme désigne un groupe de morphèmes ou de mots qui se suivent avec un sens. Le syntagme désigne aussi ce groupe formant une unité dans une organisation hiérarchisée de la phrase (syntagme verbal, nominal). Il propose une liste de huit grands types de segments autonomes (Défaut : il manque le son). Les syntagmes sont les segments autonomes formant une unité de sens. Le problème est que ce type de modèle est difficilement analysable par un ordinateur en raison de sa complexité. Dans le cadre du traitement automatique de la structure, il est donc nécessaire de définir une organisation plus simple.

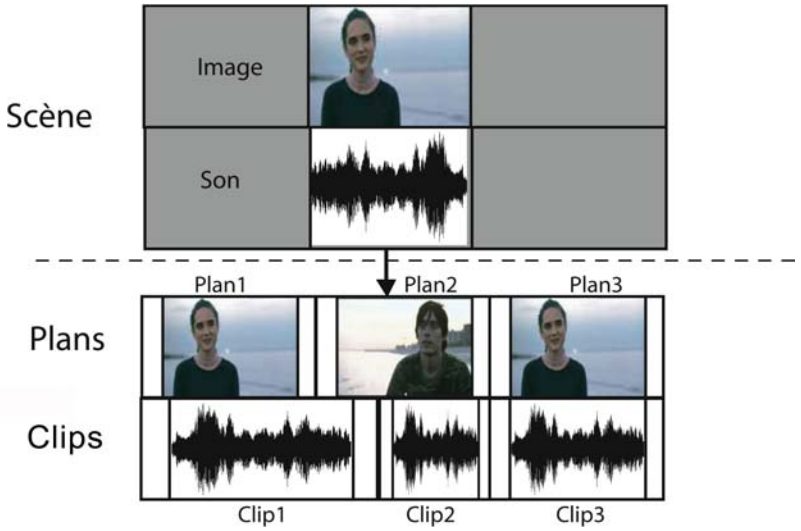

FIG. 3.6 – Structure hiérarchique de film à 2 niveaux

Les premières études ont été menées sur des structures hiérarchiques à deux niveaux plans (clips) et scènes, en considérant la synchronisation des scènes auditives et visuelles. Ce type de modèle est

présenté dans la figure 3.6. Nous introduisons deux niveaux supplémentaires : groupe de plans et groupe de scènes, ceci afin de mieux caractériser les structures complexes des films.

Scène et séquences

La scène est une suite de plans déterminée plus particulièrement par l'unité de lieu et de temps : on parlera par exemple, de la scène de la viande pourrie dans « Potemkine » (remarquer l'analogie avec une scène ou un tableau dans une pièce de théâtre) ; par contre la séquence est une notion spécifiquement cinématographique : c'est une suite de plans caractérisée par l'unité d'action (e.g. la séquence de la fusillade sur les escaliers, dans le même film) et l'unité organique, c'est-à-dire la structure propre qui lui est donnée dans le montage. Remarquons qu'aujourd'hui, peu font la distinction entre ces termes. Dans la suite nous confondrons les deux et nous appellerons ces types de structures cinématographiques : « scènes ».

Groupes de scènes

Une scène se définit donc spécifiquement par l'organisation rythmique du matériel filmé, tandis que les groupes de scènes sont les éléments qui composent un drame, de la même manière qu'une pièce de théâtre est composée de scènes et d'actes. Il s'agit des grands blocs narratifs constitués de plusieurs scènes du film.

Le nombre de groupes de scènes (ou d'actes) dans un film est assez variable, mais, en général, ils sont au nombre de trois. Comme Aristote le pensait, pour raconter une histoire, il est besoin de trois choses : un début, un milieu et une fin. La plupart des films modernes obéissent à cette règle. Le premier groupe de scènes contient les scènes d'exposition du sujet, des personnages, et du genre du film. Le deuxième est celui de la confrontation : ce que veut le personnage principal. Son « but », le fait s'affronter aux «épreuves» de l'histoire. C'est l'acte où se trouve l'apogée de l'intrigue, le «climax » ; c'est aussi le plus long. Le troisième acte est celui du dénouement, toutes les énigmes sont résolues, les perturbations de la vie du personnage principal s'apaisent. Un nouvel équilibre se construit sur de nouvelles bases. C'est un ordre nouveau, qui ne lui plait pas forcément, mais désormais sera sa vie. D'ordinaire c'est l'acte le plus bref, car l'effet recherché est l'accélération : la tension est à son comble.

Groupes de plans et groupes de clips

Une scène visuelle peut être structurée en plusieurs groupes de plans ayant chacun leur sens narratif propre. Nous prenons pour exemple une scène de dialogue (au téléphone) du film American Pie présenté en figure 3.7. Celle-ci débute un groupe de plan contenant trois plans d'introduction des lieux et des personnages de la scène, auxquels succède un groupe de plans composé d'une succession de plans fixes sur les deux personnages. Remarquons que, dans ce cas, le grossissement de l'échelle du plan confère une valeur psychologique et dramatique à la conversation. D'autres types de scènes peuvent comporter ces structures. Souvent, trois groupes de plans se distinguent dans une scène : un début, présentation du lieu et des personnages ; un milieu, l'action ; une fin, le dénouement. De la même manière une scène auditive peut être structurée en plusieurs groupes de clips ayant chacun leur sens narratif propre. En général, au sein d'une scène, les limites des groupes de plans et groupes de clips ne coïncident pas ; elles présentent souvent un léger décalage ; dans ce cas la structure narrative reste la même pour l'image et le son ; parfois, à un groupe de plans correspondent plusieurs groupes de clips, et inversement ; mais, les coupures peuvent aussi être totalement disjointes. Pour notre exemple, le dialogue audio, c'est à dire l'alternance des voix des personnages, commence dès le premier plan, ainsi cette scène ne comprend qu'un seul groupe de clips. Au contraire, une scène de poursuite débutera par le bruit de la ville et peut-être un court dialogue, c'est le premier groupe de clips ; lorsque l'action démarre, la musique apparaît, ce qui a pour effet d'accentuer la tension, deuxième groupe de clips.

FIG. 3.7 – Segmentation d'une scène en groupes de plans

Dans la suite, nous montrerons que ces structures (groupe de scènes, scènes, groupes de plans) sont détectables par segmentation et cohérentes avec les films classiques du cinéma. Pour ce faire, nous ferons le lien entre l'organisation de la structure et le contenu des éléments structurels.

3.1.4 Schéma hiérarchique du film

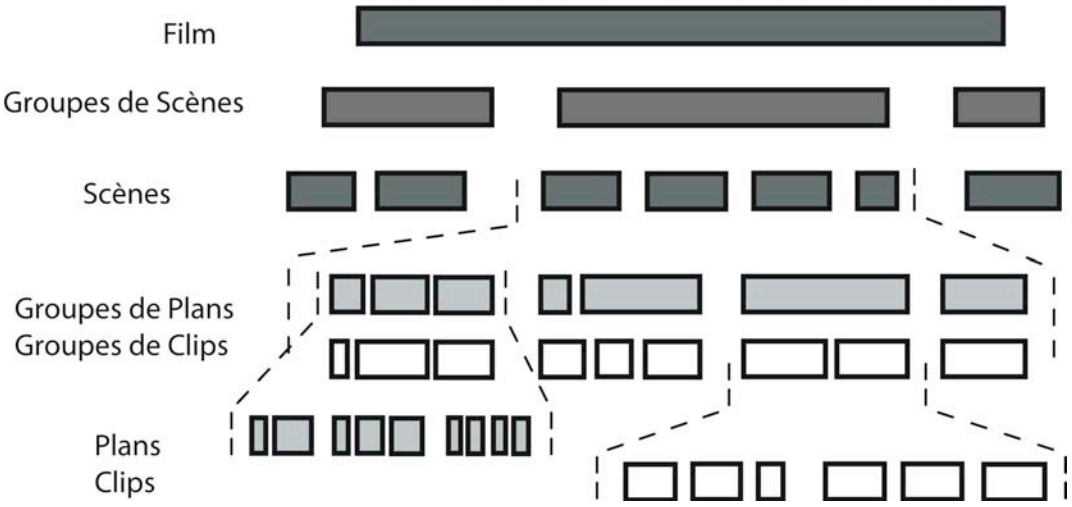

FIG. 3.8 – Modèle de structure temporelle des films

3.2 Modèle de contenu

Dans le cadre de l'analyse du contenu des structures, l'indexation se heurte à l'interprétation des images et du son. Les recherches dans ce domaine sont basées sur le modèle développé par **Panofsky**. Il décrit trois niveaux de signification :

1. Le premier, le niveau pré-iconographique, se limite aux éléments présents dans une image (objets, personnes, évènements), que l'on peut identifier de façon concrète ; on peut donc parler de premier degré, mais aussi de dénotation ou d' «**ofness**».
2. Vient ensuite le niveau iconographique (temps, lieu, action), qui exige de l'indexeur qu'il interprète les images qu'il voit (ou les sons) pour déterminer sur quoi elles portent ; on parle alors de connotation ou d' «**aboutness**».
3. Enfin, le dernier niveau, dit iconologique, s'attache à la valeur symbolique que peut renfermer une image ; on entre ici dans le domaine de l'abstrait. Or détecter dans une image ou un son la

représentation d'un concept immatériel exige une culture permettant d'établir la relation entre une image (un son) et un concept sémantiquement complexe. Cette culture n'est pas partagée par tous et les différences culturelles multiplient les valeurs symboliques attribuées à une même image, d'autant plus qu'un créateur n'a pas toujours conscience de la portée symbolique de son oeuvre. Également, si un film peut contenir une valeur symbolique évidente pour certains, surtout dans le cas des films d'auteur, c'est dans son intégralité qu'elle est perçue. Dans le cas d'un simple plan détaché de son contexte, cela relèverait plutôt de la mystification. Aussi le côté symbolique ne sera pas traité davantage.

L'analyse automatique des films amène une dernière catégorie d'information qui n'appartient à aucune des catégories définies par Panofsky. Ce niveau comporte les éléments d'indexation qui sont extraits **automatiquement** par application d'algorithmes et sans connaissance particulière du contexte. Ils expriment des propriétés significatives directement observables ou calculables à partir du signal.

3.2.1 Le contenu de niveau bas

Ces éléments d'indexation ne traduisent pas le sens de l'information contenu dans les documents traités. C'est pourquoi ce niveau est nommé « niveau sémantique bas ». En général, il contient des éléments monomedia. Pour l'image, il s'agit des caractéristiques de couleur, luminosité, texture, forme, positions.... Pour le son, un grand nombre de descriptions audio ont été proposées. Elles proviennent de plusieurs domaines de la recherche et concernent des caractéristiques physiques variées (volume, modulations, spectres...).

Pour un utilisateur non averti ces informations n'ont aucun sens. Cependant, ces données sont indispensables à la détermination automatique du contenu de niveau supérieur : elles sont la **«signature» numérique** des concepts sémantiques. C'est pourquoi elles sont nécessaires à tous les systèmes de recherche, résumé et segmentation des films.

3.2.2 Le contenu pré-iconographique : « ofness »

Le deuxième niveau de description du contenu se limite aux éléments présents dans un film que l'on peut identifier de façon concrète. Il se divise essentiellement en trois catégories :

- **Objets**. Il s'agit des éléments du décor. Le concept de décor comprend aussi bien les paysages naturels (ciel, arbre...), que les constructions humaines (bâtiment, voiture...). Ces éléments peuvent avoir une représentation dans l'image (bâtiment), dans le son (klaxon) ou dans les deux (voiture).
- **Personnes**. Cette catégorie concerne tous les humains (ou aliens...) présents dans le film, elle comprend les acteurs et les figurants. La présence d'une personne peut se traduire au niveau de l'image (visage), du son (voix-off ou personnage se trouvant à proximité de l'écran), et des deux (personnage parlant à l'écran).
- **Evénements** . Tout ce qui arrive, ce qui survient. (manifestation, sonnerie de téléphone, tir de roquette, musique...). Cette catégorie contient des éléments visuels, auditifs, et multimedia.

Ces concepts sont représentés par des modèles de source et sont caractérisés par leur présence (1) ou leur absence (0).

3.2.3 Le contenu iconographique : « aboutness »

Le troisième niveau concerne les caractéristiques qui nécessitent une interprétation et donc la connaissance de certaines « conventions sociales » de description. Il se divise en trois catégories :

- **Temps**. Information concernant le moment où l'action se passe dans le monde réel (e.g.15h30) et dans le film (e.g. 3ème scène). En ce qui concerne le vrai monde cette caractéristique est assez complexe à déterminer de façon automatique, sauf pour quelques concepts simples comme

jour/nuit pour une image d'extérieur. En ce qui concerne la temporalité du film, si la segmentation donne de bons résultats, cette information peut s'avérer utile à l'analyse.
- **Lieu**. L'ensemble des informations se rapportant à la dénomination de l'endroit où se déroule l'action. Il s'agit de concepts qui se caractérisent par leur niveau de précision : généralement, un lieu est décrit de façon hiérarchique (e.g. *intérieur/magasin/rayon des films*).
- **Action**. Opération d'un agent, physique, chimique, mécanique, immatériel ; plus simplement tout ce que l'on fait. Cela concerne l'ensemble des relations matérielles qui peuvent exister entre objets : Jean et Pierre dialoguent, un verre se casse au sol…. Il s'agit bien évidemment de concepts visuels, auditifs et multimedia.

3.3 Réunion du contenu et de la structure

Les plans et les clips sont considérés comme l'unité de base pour l'indexation des films. Le contenu d'un plan ou d'un clip correspond à l'ensemble des informations de niveau bas, « ofness » et « aboutness » présent au sein de cet élément. Notre hypothèse est que les éléments de la structure narrative (visuelle et auditive) se caractérisent par la continuité temporelle du contenu des plans et des clips qu'ils contiennent.

3.3.1 Structure et continuité du contenu

L'organisation temporelle des plans et des clips, réalisée par le montage, constitue le fondement le plus spécifique du langage cinématographique. La succession des plans (et des clips) traduit la progression dramatique du film. Elle est fondée sur le « regard » ou la « pensée » des personnages ou du spectateur, l'un et l'autre sont assimilables en vertu de l'identification perceptive du spectateur avec le personnage (phénomène fondamental du cinéma). Cette identification, et donc la « réalité » du film, ne se réalisent que si l'enchaînement des éléments filmiques maintient la tension psychologique. Ainsi le «dynamisme mental» est le facteur essentiel de liaison entre les plans. Ce qui met en évidence la notion complémentaire, celle du «**dynamisme perceptif**».

La liaison entre les plans, qu'elle soit fondée sur le dynamisme mental ou perceptif, repose sur le fait que chaque plan doit préparer le suivant en contenant un élément qui demande une réponse (ou un accomplissement) et que le plan suivant satisfera. Cette contrainte, se traduit dans les faits par des **principes de montage** que les réalisateurs respectent, sous peine de « perdre » le spectateur.

Entre deux plans successifs, il doit y avoir d'abord **continuité** du **contenu matériel**, c'est-à-dire présence dans l'un et l'autre d'un élément identique qui permettra l'identification rapide du plan et de sa situation : par exemple on montrera d'abord une vue générale de New York avec les Tours Jumelles puis un individu au pied de celles-ci. Il est indispensable également d'assurer une continuité du **contenu dynamique** : si on montre d'abord un personnage marchant à droite, il faudra éviter de le faire avancer en sens inverse dans le plan suivant, sous peine de faire croire au spectateur qu'il revient en arrière ; de même pour les mouvements de caméra. Eventuellement, il devra y avoir aussi continuité du **contenu structural**, c'est-à-dire composition identique ou semblable, de façon à assurer une liaison visuelle : la même rue filmée d'abord d'une fenêtre du côté droit, puis d'une fenêtre du côté gauche, paraîtra différente. Ce principe commande particulièrement la règle dite « règle des 180 degrés » dans le champ/contre-champ, la caméra ne doit pas franchir le plan défini par les deux personnages sous peine de donner l'impression au spectateur qu'il saute alternativement d'un côté à l'autre de l'écran. Il y a également un problème de **continuité de taille**, c'est-à-dire de mise en rapport des plans selon leur grosseur : il est bon de passer progressivement du plan général au gros plan et vice versa, sinon le spectateur risque, par suite de l'absence de références spatiales communes aux deux plans, de ne pas comprendre de quoi il s'agit. Par exemple, on ne montrera pas une foule puis un gros plan du visage d'une femme dans cette foule sans ménager une transition. Des problèmes se posent aussi en ce qui concerne les **rapports de temps**

entre les plans : mais il semble qu'ici la liberté du réalisateur soit quasi complète étant donné le caractère d'ellipse permanente du langage cinématographique. Enfin il faut assurer une certaine **continuité de longueur** : on évitera de juxtaposer des plans de longueurs très différentes (sauf effets voulus) sous peine de donner l'impression d'un style haché et fâcheusement désordonné.

Ainsi le montage est justifié dans son principe par le fait que le cinéma est art, c'est-à-dire choix et mise en ordre, comme toute œuvre de création. Le metteur en scène choisit des éléments visuels et auditifs dont la continuité constituera l'histoire, et donc le film. Une **rupture** dans ces continuités, parce qu'elle entraîne une rupture de la tension psychologique, marque un « tournant » de la narration, ce qui peut indiquer un changement d'élément structurel : groupe de plans, scène ou groupe de scènes. Ainsi à un niveau donné de la structure, un élément est caractérisé par la continuité temporelle de son contenu et la rupture de cette continuité a ses limites.

3.3.2 Segmentation du contenu

La segmentation est une étape importante de l'extraction d'information ; elle permet de mettre en évidence la structure interne du film. L'hypothèse de base est que les éléments de la structure visuelle et auditive de niveau supérieur à un (groupes de plans et clips, scènes, groupe de scènes) se caractérisent par la continuité temporelle du contenu des plans et des clips qu'ils englobent.

Les premiers modèles de segmentation des films sont donc fondés sur la détection des points de ruptures de la continuité. Celles-ci se traduisent par une variation forte de la **similarité** visuelle et (ou) auditive entre les segments de film situés avant et après le point de rupture. Ainsi pour détecter les limites des segments, un algorithme de segmentation par écrêtage est appliqué à la séquence des descripteurs du contenu. Souvent, des effets de transitions visuelles ou auditives caractéristiques indiquent la présence de points de rupture. Il s'agit des effets visuels de transition et de quelques effets sonores. La fusion de la segmentation par écrêtage et de la détection des transitions amène de bonnes performances de segmentation des scènes.

Cependant, nous ne disposons pas d'algorithmes de détection de transition et ceci n'étant pas l'objet de notre recherche, nous n'avons pas développé ce type de techniques. De plus, la segmentation par écrêtage ne semblent pas adaptée à l'extraction des structures de type groupe de plan et groupe de scènes. Dans les deux cas, les points de rupture sont assez complexes à isoler. En ce qui concerne les groupes de plans par exemple : la figure 3.7 montre une scène de dialogue du film « American Pie » comprenant deux groupes de plans. A l'œil nu, nous constatons que les points où la variation du contenu visuel est la plus forte (e.g. entre le plan 1 et 2) ne correspondent pas au point de rupture réel. Nous avons aussi souligné les problèmes liés au choix du seuil d'écrêtage. Pour que cet algorithme de segmentation soit efficace, il faudrait apprendre un seuil pour chaque scène afin de la découper en groupes de plans. Mais cela ne limiterait pas les problèmes de sur-segmentation.

Nous faisons donc l'hypothèse que la continuité temporelle du contenu d'un élément de structure se traduit par l'**homogénéité du contenu** des plans (ou clips) qu'il contient ; ce qui revient à supposer que la variation du contenu au sein d'un élément est moindre comparée à celle de l'élément qui l'englobe.

Pour notre exemple de scène de dialogue, cette hypothèse semble tout à fait justifiée. Le deuxième groupe de plans se distingue particulièrement par l'homogénéité de ses plans : un personnage, intérieur, plan proche, luminosité basse... La segmentation d'un élément de structure donné consiste donc à regrouper les plans homogènes de cette scène en groupes de plans. Cette opération est réalisée par un algorithme de **classification MMC non-supervisé** : le nombre d'états et le modèle de contenu de ces états n'est pas connu a priori. Plusieurs études ont montré l'efficacité de cette technique pour la segmentation de morceaux de musique [Pee02a], par exemple.

Le principal problème de cette technique réside dans le **choix des descripteurs** pertinents pour représenter le contenu des plans et des clips. La question se pose donc de savoir quel contenu présent dans

la séquence doit être pris en considération pour segmenter. Le critère d'exhaustivité fait face ici à celui de la sélectivité. Le problème récurrent de la qualité versus la quantité refait surface encore une fois, et nous devons faire appel à notre jugement et à notre discernement. Devant ce dilemme, on préfère presque invariablement opter pour la qualité. La difficulté consiste à déterminer les descripteurs qui font qu'une segmentation est de qualité. Il n'existe pas de réponse définitive sur ce point, tout dépend du film à traiter et de ses caractéristiques. Cependant, nous l'avons vu, certaines règles de montages peuvent nous mettre sur la piste. Une scène se caractérise souvent par l'homogénéité du lieu, par exemple.

Afin de mettre en valeur le contenu pertinent pour la segmentation à chaque niveau de la structure, nous testerons plusieurs groupes caractéristiques de descripteurs : bas, moyens, hauts et multimedia.

Chapitre 4

Cadre probabiliste pour les films

Ce quatrième chapitre décrit un cadre assez général de construction et d'utilisation de modèles probabilistes pour l'extraction du contenu multimédia des films. Nous commençons par introduire dans le paragraphe 4.1 le modèle d'objets multimedia fondé sur un modèle génératif des descripteurs du contenu. Nous exprimons alors les distributions des descripteurs grâce à la loi de Bayes afin d'intégrer le modèle au sein d'un réseau bayésien naïf. Notre hypothèse est que ce réseau ne considère pas suffisamment les relations de dépendances entre descripteurs. Nous proposons donc un nouveau modèle du contenu pour la classification multimédia. Dans le paragraphe 4.2 nous présentons l'application de ce modèle d'objets à la classification de concepts moyens. Nous expliquons dans le paragraphe 4.3 comment utiliser ces modèles pour résoudre les tâches de classification de concepts hauts. Nous détaillons l'application de ces modèles au cas de fusion de descripteurs (multimedia et bas/moyen) pour la classification.

4.1 Modèle probabiliste du contenu

4.1.1 Modèle génératif de contenu

Sur un élément atomique de la structure, clip (audio) ou plan (image), les variables observées entretiennent des relations de **dépendances** fortes. Construire un modèle probabiliste d'un élément de film consiste à modéliser ces variables par une distribution de probabilité générale capable de représenter tous les cas de figures possibles. La distribution doit alors représenter les diverses descriptions des objets multimedia (contenu bas, moyen et haut).

Définitions

Soit Q un objet question, nous définissons les ensembles de variables aléatoires :

B l'ensemble des descripteurs de niveaux bas notés $B = \{B_1, \ldots, B_j, \ldots, B_{m_B}\}$, où B_j est le $j^{ème}$ descripteur bas de l'objet.

M l'ensemble des descripteurs de niveaux moyens notés $M = \{M_1, \ldots, M_j, \ldots, M_{m_M}\}$, où M_j est le $j^{ème}$ descripteur moyen de Q.

H l'ensemble des descripteurs de niveaux hauts notés $H = \{H_1, \ldots, H_j, \ldots, H_{m_H}\}$, où H_j est le $j^{ème}$ descripteur haut de Q.

C l'ensemble des concepts de Q : $C = \{M, H\} = \{C_1, \ldots, C_j, \ldots, C_{m_C}\}$.

Les variables de B sont à valeurs dans $\mathbb{R}$, elles sont dites "observées" puisqu'elles sont obtenues par traitement du signal sans connaissance a priori. Les variables de C sont des variables discrètes à valeurs dans un ensemble fixé (la taxonomie) et seront évaluées par traitement du modèle de contenu. La

distribution de **probabilité jointe générale** s'écrit :

$$P^{\text{gen}} = \mathrm{P}(B, C)$$

Cadre bayésien pour la classification

La classification générale de l'objet Q consiste à lui attribuer les valeurs des concepts qui maximisent la probabilité d'observer ces concepts C sachant les variables observées B. Cette règle d'estimation est la règle du **maximum a posteriori** et s'écrit :

$$\widehat{C} = \arg\max_{C}(\mathrm{P}(C|B)),$$

Par définition, nous avons :

$$\mathrm{P}(C|B) = \frac{\mathrm{P}(B, C)}{\mathrm{P}(B)}$$

et ainsi,

$$\widehat{C} = \arg\max_{C}(\mathrm{P}(B, C))$$

Ainsi, la classification de l'objet question peut-être estimée par **maximisation** de la probabilité jointe générale P^{gen}. Dans notre cas le nombre de concepts étant assez réduit, une vingtaine au maximum, cette probabilité sera évaluée pour toutes les valeurs possibles de l'ensemble des concepts.

L'expression de la probabilité jointe générale des variables aléatoires se montre complexe ; une méthode simplificatrice consiste à restreindre les structures modélisables afin de l'exprimer par plusieurs termes indépendants. L'idée à la base de cette méthode est de spécifier un certain nombre de dépendances entre variables aléatoires, à l'aide de connaissances a priori du phénomène modélisé. Cela permet de réduire la complexité de l'inférence et de l'apprentissage par rapport à un modèle où toutes les dépendances statistiques sont prises en compte.

Relations de dépendance entre descripteurs

L'**indépendance** entre deux descripteurs D_1 et D_2 signifie physiquement que l'observation de D_2 ne donne aucune information sur la probabilité d'obtenir D_1. Elle se traduit par la simplification de la probabilité conditionnelle $\mathrm{P}(D_1|D_2) = \mathrm{P}(D_1)$. Cette définition est symétrique puisque $\mathrm{P}(D_2|D_1) = \mathrm{P}(D_2)$. Elle peut également s'écrire sous la forme $\mathrm{P}(D_1, D_2) = \mathrm{P}(D_1)\mathrm{P}(D_2)$.

Soient D_1, D_2 et D_3 trois descripteurs. Si $\mathrm{P}(D_1|D_3) = \mathrm{P}(D_1|D_2, D_3)$, on dit alors que la variable D_1 est indépendante de la variable D_2 pour D_3 donné. Si l'état de D_3 est donné, alors aucune variation dans la probabilité de D_1 n'affectera la probabilité de D_2, on dit que D_1 et D_2 sont conditionnellement indépendants. La relation d'**indépendance conditionnelle** entre deux variables est symétrique.

L'idée fondatrice des **réseaux bayésiens** (RB) est d'exploiter les indépendances conditionnelles entre les concepts afin de représenter la distribution de la probabilité jointe d'une manière plus compacte.

Un réseau bayésien est un graphe orienté sans cycle. Dans ce schéma, chaque nœud correspond à une variable aléatoire et les flèches entre nœuds permettent de définir les fils et les parents d'une variable. Ainsi, les arcs représentent les influences directes entre les variables. Et toute variable est alors indépendante de ses non-descendants sachant ses parents. Les dépendances respectent le principe de causalité : les flèches vont de la couche haute vers la couche basse.

Soit la probabilité jointe générale : $P^{\text{gen}} = \mathrm{P}(D_1, \ldots, D_j, \ldots, D_m)$ où D_j est le $j^{ème}$ descripteur de l'objet en question. La simplification des relations de dépendances statistiques permet d'exprimer la probabilité jointe générale par :

$$\mathrm{P}(D_1, \ldots, D_m) = \prod_{j=1..m} \mathrm{P}(D_j|\text{Parents}(D_j))$$

où Parents(D_j) est l'ensemble des parents du $j^{ème}$ descripteur.

Ainsi, pour la classification nous considérons toutes les probabilités jointes possibles. Au lieu d'un nombre de valeurs exponentiel par rapport au nombre de variables, l'opération nécessite, pour chaque variable, un nombre de valeurs exponentiel par rapport au nombre de ses parents.

4.1.2 Première hypothèse de simplification

La dépendance causale des concepts et des descripteurs bas observés permet de représenter la dépendance entre l'ensemble des concepts et des descripteurs bas sous la forme du RB dans la figure 4.1. L'état d'un concept C_j dépend uniquement des états des autres concepts. L'état d'un descripteur bas B_j dépend des états des concepts et des autres descripteurs bas. D'après la loi de Bayes :

$$\mathrm{P}(B, C) = \mathrm{P}(C)\mathrm{P}(B|C)$$

Le problème de classification revient donc à maximiser le produit de la probabilité d'observer les descripteurs bas sachant les concepts et de la probabilité marginale de l'ensemble des concepts. La probabilité marginale $\mathrm{P}(C)$ est apprise par une méthode simple de comptage. La probabilité conditionnelle $\mathrm{P}(B|C)$ est déterminée par un modèle de classification comprenant au minimum $2\,m_C$ classes (lorsque tous les concepts sont binaires) et m_B descripteurs bas. Cependant, ce type de modèle rend la tâche d'apprentissage complexe. Tout d'abord, le modèle de classification comporte trop de données ce qui diminue les performances de classification en terme de temps de calcul et de résultats. De plus, la base de données d'apprentissage doit comporter un échantillon assez large d'exemples pour chaque classe. Enfin, il est nécessaire d'apprendre le modèle sur une seule base annotée par tous les concepts de la taxonomie.

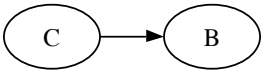

FIG. 4.1 – Relations de dépendance entre les concepts et les descripteurs bas d'un objet.

Modèle "multinet"

La deuxième hypothèse de simplification est émise pour des raisons techniques. Seulement un petit nombre de descripteurs numériques est utilisé pour l'extraction de chaque concept. Ainsi il est considéré qu'un descripteur numérique ne dépend que du concept qu'il permet de déterminer et des autres descripteurs numériques liés à ce concept.

Dans, [Nap00] les auteurs placent tous les concepts au même niveau sémantique. Ils peuvent appartenir à n'importe lesquelles des catégories de base : objets(voitures, homme, hélicoptère), lieu (extérieur, plage), ou des événements (explosion, homme qui marche). Ils proposent un réseau de concepts appelé multinet au sein duquel toutes les relations entre concepts sont prises en compte.

Le multinet autorise les relations d'interdépendance entre descripteurs, puisque cyclique. Il ne peut donc pas être représenté par un RB. Le désavantage du multinet est qu'il doit envisager la relation entre tous les concepts du système de traitement. Dans notre cas le nombre important de concepts et de descripteurs rend les calculs prohibitifs. Une solution est un modèle de fusion permettant de limiter les relations de dépendance entre les descripteurs.

4.1.3 Réseau bayésien naïf

Pour modéliser le contenu d'un objet par un RB naïf et rendre notre modèle calculable, nous allons faire un ensemble d'hypothèses sur les dépendances conditionnelles concernant les différents descripteurs :

Hypothèse 1 Les dépendances statistiques entre les nœuds du modèle respectent la hiérarchie sémantique du contenu i.e. le réseau bayésien associé à un objet est calqué sur la structure hiérarchique (bas, moyen, haut) des descripteurs. Cette hypothèse rend compte de la relation entre niveaux sémantiques.

Hypothèse 2 Le processus de génération d'un descripteur bas dépend uniquement du concept auquel il est rattaché. C'est ce qui permet de mettre en œuvre le partage de paramètres.

Hypothèse 3 Un concept moyen dépend uniquement du concept haut associé et pas des descripteurs bas de ce nœud père. Cela induit une simplification supplémentaire par rapport à l'hypothèse 1.

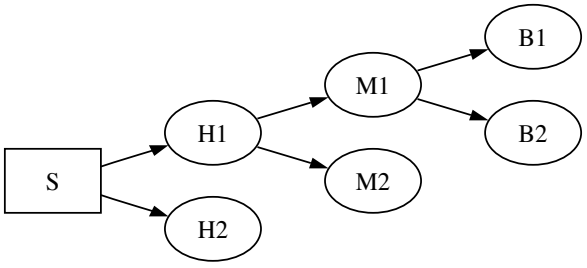

FIG. 4.2 – Relations hiérarchiques de dépendance des descripteurs du contenu

Ces hypothèses permettent de représenter les relations de dépendance entre descripteurs par le RB 4.2 et de réduire la complexité de l'inférence et de l'apprentissage par rapport à un modèle où toutes les dépendances statistiques sont prises en compte. Il correspond bien au schéma du contenu hiérarchique proposé précédemment au paragraphe 3.2.

Nous supposons que ces hypothèses sont trop réductrices et que la prise en compte des relations de dépendance entre descripteurs améliore les performances de classification. Il est donc nécessaire d'adopter un modèle hybride entre le multinet et le réseau bayésien qui considère ces relations complexes. Dans les chapitres qui suivent nous présentons ce modèle pour différents ensembles de descripteurs.

4.2 Modèle de contenu de niveau moyen

La classification du contenu moyen consiste à déterminer la valeur prise par un concept de niveau moyen. Nous envisageons d'abord le cas d'une seule classification à l'aide d'un descripteur par concept.

4.2.1 Classification d'un concept monomédia par un descripteur bas

La plupart des concepts sont modélisés par des variables binaires. Cependant l'efficacité des systèmes de classification permet de traiter le cas de concepts multiclasse. Nous envisageons ici l'ensemble de ces cas de figures. Nous considérons aussi les deux types de taxonomie de classification : le partitionnement et la classification hiérarchique.

Modèle à deux classes

Soit M un concept moyen binaire et B un descripteur associé. Nous définissons les deux hypothèses Hyp_0 et Hyp_1 comme l'hypothèse nulle (concept absent) et l'hypothèse vraie (concept présent). Nous appelons M la variable aléatoire binaire qui vaut 1 si le concept est présent et 0 s'il est absent. Ce modèle est particulièrement adapté aux tâches de classification par partitionnement en deux classes denses (e.g. Intérieur/Extérieur). Ou dans le cas, évidemment, qui se résume à la présence ou non de concepts dans l'espace des données : comme la présence de texture : herbe, ciel, ou d'objets : voiture, visage.

La relation de dépendance entre le descripteur numérique et son concept peut être décrite par le réseau bayésien $(M \rightarrow D)$. Ainsi la probabilité jointe dans ce cas s'écrit sous la forme :

$$P^{\text{gen}} = \mathrm{P}(B|M)\mathrm{P}(M)$$

Et nous attribuons à M la valeur qui maximise la probabilité jointe générale.

$$\widehat{M} = \arg \max_{M=\{1,0\}} (\mathrm{P}(B|M)\mathrm{P}(M))$$

Et $\mathrm{P}(B|M=0)$ et $\mathrm{P}(B|M=1)$ dénote la fonction de densité de probabilité conditionnelle du descripteur conditionnée sur l'hypothèse nulle (concept absent) et l'hypothèse vraie (concept présent). Ces deux probabilités sont estimées par un modèle de classification présenté dans l'état de l'art. La probabilité que le concept soit présent sachant la valeur du descripteur bas associé est notée $P^1(B) = \mathrm{P}(M=1|B)$ s'écrit donc :

$$P^1(B) = \frac{\mathrm{P}(B|M=1)}{\mathrm{P}(B)}$$

Modèle à plusieurs classes

Les partitionnements sont réalisés le plus souvent sur des modèles à plusieurs classes. Soit M un concept moyen binaire et $\mathrm{T} = \{m_1, \ldots, m_k, \ldots, m_p\}$] est la taxonomie associée. Dans ce cas, certains pratiquent une hiérarchisation de la taxonomie afin de produire des arbres de partitionnement binaires. Mais, cela entraîne des désavantages : les performances de classification sont réduites en présence d'un grand nombre de classes ; de plus, dans le cas où il est nécessaire d'enlever ou de rajouter des classes cela complique beaucoup la tâche et ne donne pas toujours de meilleurs résultats. Nous devons donc envisager la classification pour la détermination de concepts multiclasses.

La relation de dépendance entre le descripteur numérique et son concept peut être décrite comme précédemment par le réseau bayésien $(M \rightarrow B)$. Ainsi nous attribuons à M la classe m_k qui maximise la probabilité jointe générale P^{gen}.

$$\widehat{M} = \arg \max_{k=1..p} (\mathrm{P}(B|M=m_k)\mathrm{P}(M=m_k))$$

Et la probabilité que le concept soit de classe m_k sachant la valeur du descripteur bas associé noté $P^k(B) = \mathrm{P}(M=m_k|B)$ s'écrit :

$$P^k(B) = \frac{\mathrm{P}(B|M=m_k)}{\mathrm{P}(B)}$$

Modèle de classification hiérarchique

Dans notre système nous pratiquons plusieurs classifications de taxonomie hiérarchique. En effet, la classification hiérarchique est le modèle de données le plus répandu. La plupart des standards de classification de produits ou de services sont basés sur des schémas hiérarchiques. Dans le cadre de notre étude nous utiliserons ce type de taxonomie pour la segmentation du son en clip, par exemple. Il est nécessaire de définir les relations de dépendance qui existent au sein de ces classifications. Soient M_1, M_2, M_3 trois concepts moyens à valeur discrète et B_1, B_2, B_3, les ensembles de descripteurs associés. Pour rendre la chose plus concrète, supposons que ces concepts décrivent une photo et que M_1 = *intérieur/extérieur*, $M_2 = M_{\text{int}}$ = *parking/maison/commerce*, $M_3 = M_{\text{ext}}$ = *ville/nature*.

En général, la classification d'un objet suivant ces concepts se fait de manière hiérarchique comme suit : tout d'abord le concept M_1 est déterminé, la photo est d'intérieur ou d'extérieur. La hiérarchie

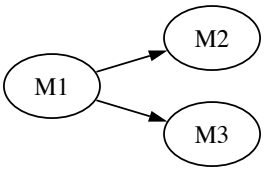

FIG. 4.3 – Relations de dépendance d'un modèle de classification hiérarchique

permet suivant la valeur de M_1 de déduire certaines hypothèses sur les valeurs possibles de M_2 et M_3. Par exemple si $M_1 =$ *intérieur* alors le concept M_3 ne peut pas prendre les valeurs ville ou nature et la valeur du concept M_2 est déterminée par classification.

On suppose que la relation de dépendance entre ces descripteurs peut être représentée par le réseau bayésien hiérarchique en 4.3. Les concepts M_2 et M_3 par définition ne sont pas denses par rapport à l'espace des données, ce qui interdit l'intégration de la hiérarchie dans le modèle par maximum de vraisemblance globale, comme on l'a défini. Il faut donc créer des concepts denses en rajoutant une variable nulle : les concepts deviennent donc $M_1 =$ *intérieur/extérieur*, $M_2 = M_{\text{int}} =$ *extérieur/parking/maison/commerce*, $M_3 = M_{\text{ext}} =$ *intérieur/ville/nature*. Les valeurs estimées de M_1, M_2, M_3 sont issues de la maximisation de la probabilité jointe générale.

$$\{\widehat{M1, M2}, M3\} = \arg \max_{\{M1,M2,M3\}} (P^{\text{gen}})$$

Sachant la probabilité générale :

$$P^{\text{gen}} = \underbrace{\mathrm{P}(M_1)\mathrm{P}(D_1|M_1)}_{M_1}\underbrace{\mathrm{P}(M_2|M_1)\mathrm{P}(D_2|M_2)}_{M_2}\underbrace{\mathrm{P}(M_3|M_1)\mathrm{P}(D_3|M_3)}_{M_3}$$

$$\Leftrightarrow$$

$$P^{\text{gen}} = P_1^{\text{gen}} * \omega_{(M_2,M_1)} P_2^{\text{gen}} * \omega_{(M_3,M_1)} P_3^{\text{gen}} \quad \text{sachant} \quad \omega(M_2, M_1) = \frac{\mathrm{P}(M_2|M_1)}{\mathrm{P}(M_2)}$$

Les relations de dépendance définies entraînent l'expression de la probabilité jointe en trois termes. Le premier correspond à la probabilité jointe générale associée au partitionnement du premier concept noté P_1^{gen}. Le deuxième et le troisième correspondent aux probabilités jointes générales associées aux partitionnements des concepts 2 et 3 (notés P_2^{gen} et P_3^{gen}) multipliés par des **taux de dépendance** (notés ω). Ces taux traduisent les liens de dépendance entre les concepts d'un niveau de précision donné et leur concept père. Ils sont appris sur la base d'exemples annotés par comptage. Dans les faits, ils permettent d'influencer la classification d'un concept par la classification des autres concepts. Ces influences sont de deux ordres :

- **Descendante**, la classification de M_1 influence celle de M_2 et M_3. Par exemple, $\omega_{(\textit{maison,extérieur})}$ est nulle, car la probabilité de se trouver à l'intérieur d'une maison et à l'extérieur est nulle ; si une image est présentée en question elle ne pourra pas être classée dans ces deux classes.
- **Ascendante**, si le premier terme favorise la décision M_1, alors les deux autres termes peuvent amener à une décision inverse. Par exemple, si M_1 est classé en intérieur par ses descripteurs bas, la classification $M_2 =$ *ville* et $M_3 =$ *extérieur* par leur descripteurs associés peut modifier la décision.

Modèle de deux concepts indépendants

Nous avons défini les relations qui existent entre un descripteur et son concept (à deux classes ou multiclasse), ainsi que les relations entre concepts d'une classification hiérarchique. Il nous faut donc maintenant définir les relations entre concepts moyens non liés par une relation hiérarchique. Soit M_1

et M_2, deux concepts moyens de ce type. L'hypothèse est que ces concepts sont indépendants. Cette relation est représentée par le réseau bayésien, (M_1 M_2), où aucun lien n'existe entre les deux concepts. Ce qui permet de simplifier la probabilité jointe générale :

$$P^{\text{gen}} = P_1^{\text{gen}} P_2^{\text{gen}} \quad \text{sachant} \quad P_j^{\text{gen}} = \mathrm{P}(M_j)\mathrm{P}(D_j|M_j)$$

Dans ce cas les classifications des deux concepts sont indépendantes l'une de l'autre. Et nous remarquons que la probabilité peut être pondérée par les taux de précédemment définis :

$$P^{\text{gen}} = \omega_{(M_1,M_2)} P_1^{\text{gen}}\ \omega_{(M_2,M_1)} P_2^{\text{gen}}$$

où

$$\omega_{(M_2,M_1)} = \omega_{(M_1,M_2)} = 1$$

4.2.2 Fusion du contenu de niveau bas

La plupart des descripteurs de niveau moyen sont extraits par classification de plusieurs descripteurs bas.

Classification par plusieurs descripteurs bas : early fusion

Soit M un concept moyen. L'ensemble des descripteurs de niveaux bas associés à ce concept est noté $B = \{B_1, \ldots, B_j, \ldots, B_m\}$ où B_j est le $j^{ème}$ descripteur bas de l'objet. Les hypothèses definies précédemment impliquent les relations de dépendance représentées par le réseau bayésien 4.4.

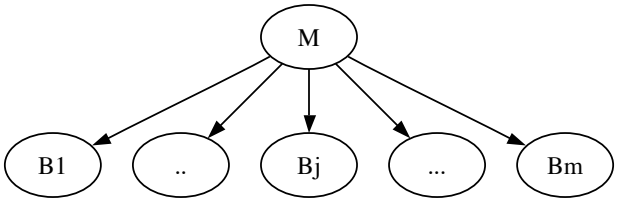

FIG. 4.4 – Relations de dépendance entre descripteurs bas

Ce réseau est appelé réseau naïf de Bayes et est fondée sur l'hypothèse que tous les descripteurs bas sont indépendants sachant le concept. Certaines modifications de l'espace des descripteurs bas tendent à les rendre décorrélés (PCA) ou indépendants (ICA). C'est pourquoi il est raisonnable d'envisager cette hypothèse. La probabilité jointe générale s'écrit

$$P^{\text{gen}} = \mathrm{P}(M) \prod_{j=1..m} \mathrm{P}(B_j|M).$$

Elle est le produit des probabilités jointes individuelles.

En un sens, la classification naïve de Bayes est optimale, car toute l'information disponible est utilisée pour effectuer une induction. De façon en apparence surprenante, on a constaté qu'elle montrait d'aussi bonnes performances dans les cas où les descripteurs ne sont pas indépendants. Cependant, expérimentalement, les hypothèses d'indépendance conditionnelle définissant chaque descripteur ne sont pas parfaitement valides. Généralement, ces dépendances sont faibles, mais l'hypothèse d'indépendance surévalue l'importance des descripteurs en réalité dépendants. Les modéliser explicitement conduit à des calculs supplémentaires, voire à des problèmes de surapprentissage.

Certains approchent ces dépendances simplement dans l'expression de la probabilité jointe des variables en remplaçant la loi de Bayes naïve par la loi de Bayes pondérée. La probabilité jointe générale s'écrit alors

$$P^{\text{gen}} = \mathrm{P}(M) \prod_{j=1..m} \mathrm{P}(B_j|M)^{\omega_j}$$

Les poids ω_j sont compris entre 0 et 1 et d'autant plus proches de 1 que les descripteurs correspondants sont conditionnellement indépendants des autres. L'utilisation de ces poids est justifiée du point de vue probabiliste et plusieurs algorithmes d'apprentissage ont été présentés dans la littérature pour les estimer (voir [Han01]). Ces techniques donnent des résultats assez satisfaisants, mais nous pensons qu'elles ne considèrent pas de façon appropriée toutes les dépendances entre les descripteurs bas. Notamment, parce que les poids ainsi définis sont indépendants de la valeur du concept. Il serait intéressant, lorsque le descripteur D_j est caractéristique de la $k^{ème}$ classe, d'apporter plus de poids à ce descripteur quand le concept M vaut cette classe m_k et moins quand $M = m_{k'}$.

C'est pourquoi nous proposons de prendre en compte les relations de dépendance entre les descripteurs bas sachant le concept. Ces relations de dépendance peuvent être représentées par le réseau bayésien $(M \rightarrow B)$. La probabilité jointe générale s'écrit alors :

$$P^{\text{gen}} = \mathrm{P}(M)\mathrm{P}(B|M).$$

où la probabilité conditionnelle $\mathrm{P}(B|M)$ est estimée par normalisation du score de classification obtenu par un modèle multidimensionnel (e.g. SVM, PPV). L'utilisation de ces modèles permet de considérer les dépendances complexes entre descripteurs. Nous remarquons, que la classification SVM multiclasse OPC, décrite au chapitre 2.2.3, permet d'accorder un poids variable aux descripteurs bas sachant la valeur du concept. Ceci est dû au fait que, pour chaque classe, un modèle de classification indépendant est déterminé.

Cette méthode de fusion à montré son efficacité pour de nombreuses applications de classification de descripteurs bas. Par la suite, nous montrerons qu'elle est applicable pour toutes les fusions envisagées dans notre étude.

Classification par plusieurs descripteurs bas : late fusion

Dans certaines situations, il est envisageable de modéliser un concept pour chaque descripteur (ou ensemble de descripteurs) de façon indépendante et de fusionner la décision des classifications ensuite : c'est la late fusion. Différents modèles du même concept sont créés pour chaque descripteur, ce qui résulte en de multiples classifications auxquelles sont associés différents résultats.

Ainsi nous définissons le concept moyen M et B l'ensemble des descripteurs bas associés. Pour simplification, nous considérons que M est binaire. Il est aisé de prolonger la théorie au cas multiclasse. On appelle M_j le concept moyen lié à l'ensemble des descripteurs bas B_j de B. Le choix des ensembles de descripteurs est à la discrétion du chercheur. Il existe plusieurs variantes : un concept intermédiaire par dimension, un par descripteur multidimensionnel ou bien un par media.

En général, l'hypothèse est que les concepts M_j ne dépendent que de M, et que les ensembles de descripteurs bas associés, notés B_j, ne dépendent que de leur concept M_j. Les relations de dépendance statistiques peuvent donc être représentées par le RB en figure 4.5. La probabilité jointe générale s'écrit alors :

$$P^{\text{gen}} = \mathrm{P}(M) \prod_{j=1..m} \mathrm{P}(M_j|M)\mathrm{P}(B_j|M_j)$$

$$P^{\text{gen}} = \mathrm{P}(M) \prod_{j=1..m} \omega_{(M_j,M)} P_j^{\text{gen}}$$

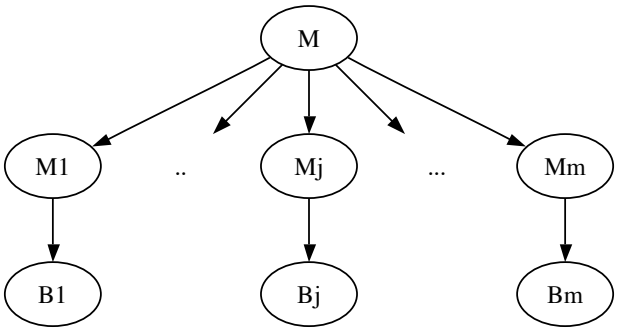

FIG. 4.5 – Relations de dépendance entre descripteurs de la late fusion

Les taux $\omega_{(M_j,M)}$ traduisent le lien de dépendance entre les concepts intermédiaires et le concept moyen. Dans les faits, ils traduisent le poids de chaque descripteur bas dans la classification du concept moyen. Par exemple, un taux associé faible suppose que la classification par ce descripteur entraîne de nombreuses erreurs et qu'elle ne doit pas trop influencer la classification finale.

Mais on l'a vu, l'indépendance entre descripteurs bas n'est pas vérifiée. Il en est de même pour celles des concepts associés. La probabilité jointe générale s'exprime donc en un seul terme comme :

$$P^{\text{gen}} = \mathrm{P}(M)\mathrm{P}(\{M_1,\ldots,M_m\},\{B_1,\ldots,B_m\}|M)$$

Dans ce cas nous préférerons donc réaliser la fusion des concepts par classification SVM du vecteur scores de M. La $j^{ème}$ dimension de ce vecteur contient le score de classification s_j^1 correspondant à la probabilité conditionnelle que le concept M_j soit présent sachant ses descripteurs bas associés (noté $P_j^1 = \mathrm{P}(M_j = 1|B_j)$).

Remarquons que par rapport au réseau bayésien naïf, les SVM ne considèrent pas l'influence descendante du concept moyen sur les concepts intermédiaires. Ceci n'a que peu d'influence sur le résultat final, car, dans le cas de la late fusion, la classification des concepts intermédiaires ne nous intéresse pas.

4.3 Modèle de contenu de niveau haut

La classification du contenu haut consiste à déterminer la valeur prise par un concept de niveau haut. Dans ce paragraphe nous étudierons les relations de dépendance statistique suivantes :

- entre un concept haut et les descripteurs associés pour sa classification,
- entre concepts moyens et descripteurs bas associés au concept haut,

Nous soulignerons l'influence de ces relations sur l'algorithme de classification globale.

4.3.1 Classification d'un concept haut monomedia

Le cas d'une classification à l'aide d'un descripteur moyen et d'un descripteur bas est d'abord envisagé. Soit H un concept haut binaire. Nous notons M le concept moyen associé à H et B le descripteur bas correspondant à M. Les hypothèses proposées précédemment impliquent plusieurs relations de dépendance entre ces descripteurs. La relation entre le concept haut et le descripteur moyen associé peut être décrite par le réseau bayésien $(H \rightarrow M)$. De même la relation entre le concept moyen et le descripteur bas est décrite par $(M \rightarrow B)$. L'ensemble de ces relations est représenté par le réseau bayésien $(H \rightarrow M \rightarrow B)$. La probabilité jointe générale s'écrit alors :

$$P^{\text{gen}} = \mathrm{P}(H)\,\mathrm{P}(M|H)\mathrm{P}(B|M)$$

$$P^{\text{gen}} = \mathrm{P}(H)\omega_{(M,H)}P_M^{\text{gen}}$$

Elle est évaluée pour toutes les valeurs du couple (M, H) et les valeurs qui la maximisent sont sélectionnées.

4.3.2 Fusion du contenu pour la classification haute

Pour la plupart des modèles, les descripteurs de niveau haut sont extraits par classification de plusieurs descripteurs du signal de niveaux sémantiques bas et moyen. Soit H un concept haut, et $M = \{M_1, \ldots, M_j, \ldots, M_m\}$ l'ensemble des concepts moyens utilisés pour la classification de H. Pour simplification nous supposons qu'ils sont binaires, ce qui est le cas la plupart du temps. Nous notons B_H l'ensemble des descripteurs bas associés à la classification de H. De même nous appelons $B_M = \{B_1, \ldots, B_j, \ldots, B_m\}$ l'ensemble des descripteurs bas associés aux classifications des concepts moyens où B_j est liée au $j^{ème}$ concept M_j.

Fusion des concepts moyens

Supposons tout d'abord que la classification de H se fait par fusion de concepts uniquement (B_H est vide). Les hypothèses simplificatrices supposent que les concepts moyens sont indépendants entre eux sachant H. De plus, les B_j ne dépendent que de leurs concepts respectifs M_j. Ainsi nous pouvons représenter le modèle de classification sous la forme d'un réseau bayésien dit naïf en figure 4.6.

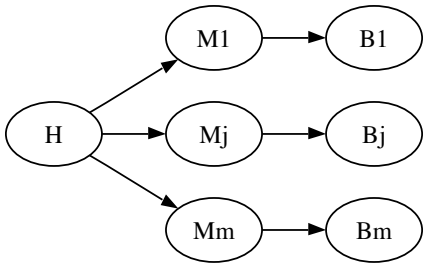

FIG. 4.6 – Relations de dépendance entre concepts moyens

Nous notons P_j^{gen} la probabilité jointe de la classification du $j^{ème}$ concept. Et, nous définissons le taux $\omega_{(M_j,M_H)}$ de dépendance de H sur M_j. Dans ce cas la probabilité jointe s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H) \prod_{j=1..m} \mathrm{P}(M_j|H)\mathrm{P}(B_j|M_j)$$

ainsi,

$$P^{\text{gen}} = \mathrm{P}(H) \prod_{j=1..m} \omega_{(M_j,M_H)} P_j^{\text{gen}}$$

Les probabilités conditionnelles $\mathrm{P}(M_j|H)$ et $\mathrm{P}(B_j|M_j)$ sont déterminées par comptage sur la base de données et par un modèle SVM de classification de concepts moyens. La probabilité marginale $\mathrm{P}(H)$ est estimée aussi par comptage. Les taux $\omega_{(M_j,H)}$ traduisent le lien de dépendance entre les concepts moyens atomiques et le concept haut. Dans les faits, ils permettent d'influencer sur la classification d'un concept par la classification de l'autre de façon ascendante (les M_j influencent H) et descendante (H influence les M_j).

Cependant l'hypothèse d'indépendance entre concepts moyens n'est pas valable dans les faits. Et la classification par réseau bayésien surévalue l'importance des concepts dépendants. Afin d'évaluer cette hypothèse, nous comparerons le modèle par réseau bayésien et le modèle par classification des scores utilisé dans le cadre de la late fusion et adapté à l'extraction de concept haut. Pour ce modèle les relations

de dépendance entre les descripteurs sont représentés par le RB $(H \rightarrow \{M, B\})$, et la probabilité jointe générale s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H)\mathrm{P}(M, B|H)$$

L'algorithme de classification SVM estime la probabilité conditionnelle générale (notée $\mathrm{P}(M, B|H)$) par classification du vecteur score de présence S^1 associé à H. La $j^{ème}$ dimension de ce vecteur contient le score de présence s_j^1 correspondant à la probabilité conditionnelle que le concept M_j soit présent sachant ses descripteurs bas associés (noté $P_j^1 = P'(M_j = 1|B_j)$). Ce qui permet de prendre en compte la dépendance entre concepts moyens pour la détermination d'un concept haut.

Mais, remarquons que, comme pour la late fusion, les SVM ne prennent pas en compte la relation descendante entre le concept haut et les concepts atomiques. Dans le cas où les concepts atomiques sont annotés manuellement, cela ne posera aucun problème. Par contre, si la classification de ces concepts est automatique, la perte de cette influence peut diminuer les performances du système.

Fusion des concepts moyens et des descripteurs bas

Nous nous plaçons ici dans le cadre de la détermination d'un concept haut H par un ensemble de descripteurs bas B_H et de concepts moyens M. Les notations définies précédemment sont conservées : B est l'ensemble de descripteurs associés à M. Pour simplifier il est courant de supposer que les descripteurs de l'ensemble B_H sont indépendants des concepts moyens de M sachant le concept de classification H. De plus, nous considérons que les descripteurs B_j ne dépendent que de leur concept respectif M_j. Ainsi les relations de dépendance du modèle de classification peuvent être représentées sous la forme du réseau bayésien en figure 4.7.

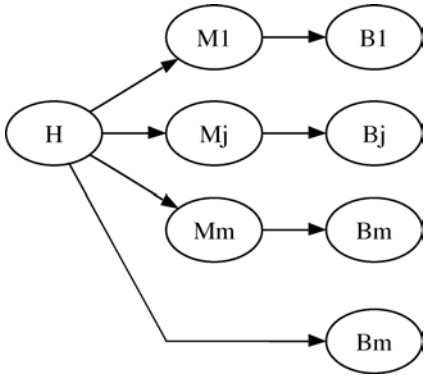

FIG. 4.7 – Relations de dépendance du modèle de concept haut

Nous appelons P_{BH}^{gen} la probabilité jointe du modèle de classification de H par B_H. De même, nous notons P_j^{gen} la probabilité jointe de la classification du $j^{ème}$ concept. Enfin, nous définissons $\omega_{(M_j,H)}$, le taux de dépendance de H sur M_j. La probabilité jointe globale s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H)\ \mathrm{P}(B_H|H) \prod_{j=1..m} \mathrm{P}(M_j|H)\mathrm{P}(B_j|M_j)$$

ainsi,

$$P^{\text{gen}} = \mathrm{P}(H)\ P_{B_H}^{\text{gen}} \prod_{j=1..m} \omega_{(M_j,M_H)} P_j^{\text{gen}}$$

Les taux $\omega_{(M_j,H)}$ traduisent le lien de dépendance entre les concepts moyens atomiques et le concept haut. Dans les faits, ils permettent d'influencer sur la classification d'un concept par la classification de l'autre. Ces influences sont de deux ordres :

– **Descendante**, la classification de H influence sur celle de M_j. Par exemple, $\omega_{(\textit{voiture}=1,H=\textit{intérieur})}$ est quasiment nulle, car la probabilité de trouver une voiture à l'intérieur est faible (à part dans un parking sous-terrain) ; si une image possède d'autres caractéristiques de niveau bas d'un intérieur, ce taux va favoriser la classification *voiture* $= 0$.
– **Ascendante**, si le premier terme, celui associé aux descripteurs bas, favorise la décision $H =$ *intérieur*, par exemple, alors la classification de *voiture* $= 1$ avec un fort taux de confiance peut amener à une décision inverse. C'est la caractéristique la plus intéressante pour nous puisque notre but est d'estimer des concepts hauts.

Mais contrairement à l'ensemble des descripteurs bas B_H dont les éléments peuvent être proches de l'indépendance, l'ensemble $\{B_H, M\}$ est le théâtre de relations de dépendances souvent complexes. Et dans ce cas, l'hypothèse proposée d'indépendance des descriptions basse et moyenne sachant le concept de classification H ne semble pas valable. Et la probabilité jointe s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H)\mathrm{P}(B_H, M, B|H)$$

En pratique, l'évaluation de $\mathrm{P}(B_H, M, B|H)$, la probabilité conditionnelle de la description sachant le concept H, est une tâche complexe.

Nous proposons pour cela de réaliser une "early" fusion décrite pour les descripteurs bas au paragraphe 2.7.1. La probabilité conditionnelle de la description est estimée par classification de H par le vecteur produit de la concaténation du vecteur descripteur de B_H et du vecteur score de présence de M. Ce vecteur que nous appelons **vecteur descripteur global** est noté $\mathbf{D}_{\text{glob}}$.

A ce point, il est envisageable de réaliser une modification de l'espace des descriptions associées à H par une chaîne appropriée de traitement de l'espace des données. Nous le verrons, cette chaîne peut contenir plusieurs traitements : transformation de l'espace avec ou sans diminution de la dimension, élimination de descripteurs.

Afin de comparer ce modèle aux méthodes existantes, nous analyserons les performances du modèle de "late" fusion décrit par [Luo01]. Les auteurs utilisent deux nouveaux concepts hauts intermédiaires : un concept H_B associé à l'ensemble des descripteurs bas de H ; et un concept H_M lié aux concepts moyens. Les hypothèses d'indépendance représentées par le RB dans la figure 4.8 supposent que les

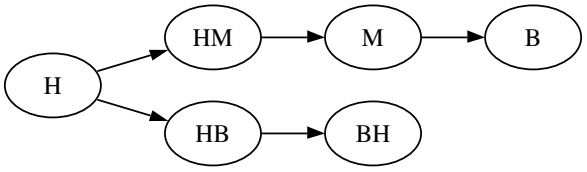

FIG. 4.8 – Relations de dépendance du modèle de late fusion moyen/bas

deux concepts intermédiaires sont indépendants sachant H. De même, l'ensemble des concepts moyens M ne dépend que du concept intermédiaire associé H_M, et les descripteurs bas B_H ne dépendent que de H_B. Dans ce cas la probabilité jointe générale s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H)\ \mathrm{P}(H_B|H)\mathrm{P}(B_H|H_B)\ \mathrm{P}(H_M|H)\mathrm{P}(M, B|H_M)$$

ainsi :

$$P^{\text{gen}} = \mathrm{P}(H)\ \omega_{(H_B,H)} P_B^{\text{gen}}\ \omega_{(H_M,H)} P_M^{\text{gen}}$$

Les taux $\omega_{(H_B,H)}$ et $\omega_{(H_M,H)}$ associés aux concepts intermédiaires traduisent ici le poids de chaque ensemble de descripteurs dans la classification du concept haut. Par exemple, un faible taux associé aux

descripteurs bas suppose que la classification par cet ensemble entraîne de nombreuses erreurs et qu'elle ne doit pas trop influencer la classification finale.

Cependant, nous avons souligné que l'indépendance entre descripteurs bas et moyens n'est pas vérifiée. Nous supposons donc qu'elle ne l'est pas non plus pour les concepts intermédiaires associés. la probabilité jointe générale s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H)\mathrm{P}(H_B, B_H, H_M, M, B|H)$$

Afin d'estimer la probabilité conditionnelle de l'ensemble de description de H sachant H, nous proposons de réaliser la fusion des concepts par classification SVM du vecteur score "late" de H. La première dimension de ce vecteur contient le score de classification du concept haut H_B par les descripteurs bas ; et la deuxième le score de classification de H_M par les concepts moyens.

4.3.3 Fusion du contenu multimedia

Nous l'avons vu, les relations des descripteurs au sein d'un seul media sont complexes. Dans ce cadre, la classification nécessite des modèles appropriés aux dépendances statistiques de la description. Il en va de même pour les descripteurs provenant de media différents. Nous étudierons deux modèles génératifs différents. Le premier issu de la littérature [Ben98, Aqa02] réalise la "late" fusion des descripteurs. Et nous proposons un deuxième modèle qui place le contenu de l'image et du son au même niveau et réalise la fusion "early" de toute la description.

Nous définissons H un concept haut multimedia, M l'ensemble des concepts moyens associé à H et B l'ensemble des descripteurs bas de H. Pour la simplification des notations, nous oublierons les descripteurs bas associés aux concepts moyens. De plus, nous appelons M_{Son} et B_{Son} les ensembles de descripteurs (moyens et bas) du contenu audio ; de même pour l'image et M_{Ima} et B_{Ima}.

Late fusion multimedia

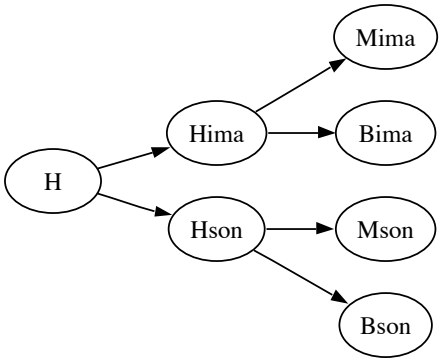

FIG. 4.9 – Relations de dépendance du modèle de late fusion multimedia

Le premier modèle introduit deux nouveaux concepts hauts monomedia intermédiaires : un concept H_{Son} associé à l'ensemble des descripteurs audio de H ; et un concept H_{Ima} lié aux descripteurs images. Les hypothèses de dépendance représentées par le RB dans la figure 4.9 supposent que les deux concepts intermédiaires sont indépendants sachant H. Dans ce cas la probabilité jointe générale s'écrit :

$$P^{\text{gen}} = \mathrm{P}(H)\ \mathrm{P}(H_{\text{Son}}|H)\mathrm{P}(M_{\text{Son}}, B_{\text{Son}}|H_{\text{Son}})\mathrm{P}(H_{\text{Ima}}|H)\mathrm{P}(M_{\text{Ima}}, B_{\text{Ima}}|H_{\text{Ima}})$$

ainsi,

$$P^{\text{gen}} = \mathrm{P}(H)\ \omega_{\text{Son}} P^{\text{gen}}_{\text{Son}}\ \omega_{Ima} P^{\text{gen}}_{\text{Ima}}$$

Où la probabilité jointe du modèle de classification de H par le son est notée $P^{\text{gen}}_{\text{Son}}$ et celle de l'image, P^{gen}_{j} ; et ω_{Ima}, ω_{Ima}, les taux de dépendance associés. Ces taux associés à chacun des media traduisent le poids des modalités dans la classification du concept haut. Par exemple, un taux associé au son faible suppose que la classification par ce média entraîne de nombreuses erreurs et qu'elle ne doit pas trop influencer la classification finale. Nous remarquons que la relation ainsi définie est **linéaire** et donc indépendante des taux de confiance associés à chaque media.

C'est pourquoi, en pratique, nous préférons à cette technique la technique de classification SVM des scores des concepts intermédiaires monomedia pour la late fusion. Ceci afin de mieux considérer les relations de dépendance entre les descripteurs audio et image. Comme nous l'avons déjà souligné, les SVM ne considèrent pas l'influence de H sur les concepts monomédia. Cependant, seul H importe et nous ne sommes pas intéressés par la détermination des concepts intermédiaires.

Early fusion multimedia

Nous proposons un deuxième modèle fondé sur la fusion "early" décrite précédemment. Celui-ci suppose que les éléments de la description associés au concept haut multimedia sont dépendants. Ce qui est représenté par la RB $(H \rightarrow \{M, B\})$. La probabilité jointe s'écrit alors :

$$P^{\text{gen}} = \mathrm{P}(H)\mathrm{P}(M, B|H)$$

La probabilité conditionnelle de la description $\mathrm{P}(M, B|H)$ est estimée par classification de H sur le vecteur produit de la concaténation du vecteur des descripteurs bas (audio et image) et du vecteur des scores moyens. Ce vecteur que nous appelons **vecteur descripteur global** est noté $\mathbf{D}_{\text{glob}}$.

Il est envisageable de réaliser une modification de l'espace des descriptions associées à H par une chaîne appropriée de traitement de l'espace des données. Cette chaîne peut devenir assez complexe dans ce cas où les concepts et les descripteurs proviennent de media différents.

4.3.4 Résumé du modèle de fusion

Résumons maintenant le modèle de fusion de descripteurs au vu des tâches à accomplir et du potentiel de chacun des algorithmes proposés.

Le but du modèle de fusion est d'obtenir une classification globale des concepts sachant les descripteurs observés. La méthode consiste à maximiser la probabilité jointe des concepts et des observations. L'estimation de la probabilité jointe globale est complexe, nous nous proposons de la simplifier par la définition des relations de dépendances qu'entretiennent les descripteurs.

Tout d'abord nous avons discrétisé l'échelle d'abstraction des descripteurs du contenu en trois niveaux sémantiques : bas, moyen et haut.

Ensuite, nous avons mis en évidence trois types de relations entre ces descripteurs :

1. La première concerne les concepts réunis au sein d'une taxonomie hiérarchique, ceux de la classification d'un lieu en *intérieur/extérieur*, *extérieur/ville/nature*... par exemple. Cette relation est bidirectionnelle (ascendante et descendante) entre les niveaux de précision. La classification du deuxième concept en *ville* interdit la classification du premier concept en *intérieur.* La classification en *intérieur* interdit la classification en *ville*.
2. La seconde concerne un descripteur et le concept qu'il influence au niveau supérieur. La présence d'une *voiture* influence le concept haut *intérieur/extérieur*. Cette relation est bidirectionelle (ascendante et descendante) entre les niveaux sémantiques, puisque, de la même façon, *intérieur/extérieur* peut influencer la présence d'une *voiture*.
3. La troisième concerne les descripteurs associés à un même concept de niveau supérieur. La *couleur d'une image* et la présence d'une *voiture* sont liés par des relations de corrélation complexes. Celles-ci influencent la classification du concept supérieur *intérieur/extérieur*.

Enfin, nous avons envisagé l'application de deux modèles de classification hiérarchique, le réseau bayésien (RB) et les support vector machine (SVM). Ces modèles diffèrent par leur capacité à considérer ces relations dans le cadre d'une fusion. Pour résumer :

- Les réseaux bayésiens définissent des relations de dépendances conditionnelles entre les descripteurs. Ils autorisent les relations ascendantes et descendantes : (1) et (2). Mais, la relation (3) de corrélation entre descripteurs est fortement limitée.
- Les "support vector machine" hiérarchiques pratiquent plusieurs classifications successives ; les probabilités intermédiaires obtenues par une classification servent à la classification du niveau sémantique supérieur. L'avantage des SVM est qu'ils considèrent la corrélation entre les descripteurs. Les relations de type (3) sont donc bien prises en compte. Cependant, ils interdisent les influences bidirectionnelles (ascendante-descendante). Ainsi, les relations de type (1) sont envisagées du haut vers le bas et de type (2), du bas vers le haut.

Les expériences réalisées par la suite nous permettront de conclure sur les capacités des deux modèles de classification, en terme de performance de classification. Une attention particulière sera aussi portée à l'évaluation des modèles de "late" fusion décrits dans ce chapitre.

Chapitre 5

Fusion de descripteurs bas monomédia

Le but de ce cinquième chapitre est de valider nos hypothèses de modélisation concernant la fusion des descripteurs de niveau sémantique bas. Pour cela, nous réalisons des expériences de classification du contenu à partir de ces descripteurs. Dans le paragraphe 5.1 nous présentons les modèles de classification pour les concepts appris. Puis nous montrons les résultats de classifications obtenus pour chacun des concepts dans le paragraphe 5.2. Nous comparons les hypothèses de modélisation concernant l'indépendance des descripteurs bas, l'utilité de la modification des descripteurs bas et la late fusion. Ce qui nous conduit en particulier à sélectionner les modèles utilisés par la suite parmi tous les modèles testés.

Les résultats d'identification des concepts associés à la classification parole/musique ont été diffusés en partie dans l'article [Del03].

5.1 Modèle de classification

Afin de valider ou non les hypothèses concernant la fusion de descripteurs bas, les résultats de classification de deux concepts classiques sont étudiés. Pour nos expériences, nous avons choisi deux classifications audio nécessaires à l'analyse de films : voix/musique et nom du personnage. Les caractéristiques des tâches d'extraction automatique permettent de traiter l'ensemble des problèmes posés.

5.1.1 Le contenu utilisé

Classification voix/musique

Le modèle de classification présenté est un modèle classique de concepts moyens audio. Il s'agit d'une classification hiérarchique dont l'ontologie est présentée dans la figure 5.1. Bien que celle-ci comporte plusieurs concepts, elle est souvent appelée classification parole/musique.

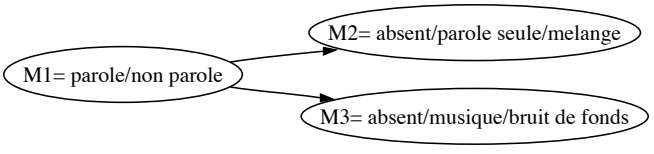

FIG. 5.1 – Classification parole/musique

L'ensemble M des concepts de la classification est défini par :
- $M_1 =$ *parole/non parole*
- $M_2 =$ *absent/parole seule/mélange*
- $M_3 =$ *absent/musique/bruit de fond*

Les **descripteurs bas** utilisés pour cette classification sont extraits à partir de segments temporels (frames) audio de taille fixée (3s dans notre cas). Tous les concepts sont extraits à partir des mêmes descripteurs bas : soit $B = \{B_1, \ldots, B_m\}$ l'ensemble de ces descripteurs. Ils sont issus de plusieurs domaines de la recherche : du traitement de la parole, de l'analyse musicale et de la psycho-acoustique... Pour un segment d'audio donné, il s'agit de statistiques (moyenne, écart type) de descripteurs locaux calculés sur des segments successifs de 10ms : énergie, volume, zero-crossing rate ratio (ZCR), fréquence fondamentale, flux spectral, sous-bandes MFCC (Mel-Frequency Cepstrum Coefficients), centroïde spectral, spectral rolloff point, sous-bandes d'ondelette. Ces descripteurs capturent les caractéristiques à court terme du signal audio. D'autres descripteurs sont extraits sur l'intégralité d'une frame : high zero-crossing rate ratio (HZCRR), low short time energy ratio (LSTER), volume dynamic range (VDR), modulation à 4Hz (4ME).

L'ensemble de ces descripteurs a été présenté dans l'article [Del03] et dans l'état de l'art au paragraphe 2.1.4.

Reconnaissance de personnages

Il s'agit de la détermination du locuteur d'un clip de parole. Cette tâche permet de reconnaître un acteur ou une voix off dans le signal audio d'un élément de film. Elle est aussi utilisée pour de la reconnaissance de personne par l'audio dans le cadre d'applications de sécurité, par exemple.

Soit H = nom de l'acteur, le **concept de classification**. Pour notre étude le modèle a été appris sur quatre films en version française. Nous choisissons, pour des raisons pratiques une taxonomie à 8 classes. Sept classes de la taxonomie $\mathrm{T} = \{c_0, \ldots, c_7\}$ contiennent des segments prononcés par un acteur dont le modèle de voix a été appris. La classe « autre » est constituée des segments qui n'appartiennent à aucune des sept premières classes.

- c_1 : Angelina Jolie (Tomb Raider)
- c_2 : Catherine Deneuve (Belle Maman)
- c_3 : Vincent Lindon (Belle Maman)
- c_4 : Seann William Scott (American Pie)
- c_5 : Jason Biggs (American Pie)
- c_6 : Jacques Gamblin (Les enfants du marais)
- c_7 : Jacques Villeret (Les enfants du marais)
- c_0 : autre

Soit B' l'ensemble des **descripteurs bas** associés aux concepts pour la classification. Les descripteurs numériques utilisés basé sur les MFCC sont issus de la recherche en reconnaissance de la parole et sont présentés par [Bim04] dans le cadre de la reconnaissance du locuteur. Ils sont extraits à partir de segments d'audio de taille variable de 3 à 21s. Ce sont les moyennes et écarts types de l'énergie des bandes du « spectre » MFCC et du rapport de l'énergie d'une bande sur le produit des énergies des bandes supérieure et inférieure. Si nous appelons MFCC_i l'énergie de la $i^{ème}$ bande du spectre MFCC, ce rapport s'écrit :

$$\frac{\mathrm{MFCC}_i}{\mathrm{MFCC}_{i-1}\mathrm{MFCC}_{i+1}}$$

5.1.2 Modèle statistique

Classification voix/musique

Le modèle de taxonomie hiérarchique présenté au paragraphe 4.2.1 est appliqué à la classification *parole/musique*. La fusion des descripteurs bas est assurée par le modèle présenté au paragraphe 4.2.2.

Reconnaissance du personnage

Le modèle statistique de détermination de concept par descripteurs bas présenté au paragraphe 4.2.2 est appliqué. Le concept H peut prendre 8 valeurs notées $(c_0, \ldots, c_7)$. Les relations de dépendance entre le concept et les descripteurs associés sont représentées par le RB : $H \rightarrow B'$.

5.1.3 Transformation de l'espace des descripteurs

De nombreuses études décrites au paragraphe 2.3 ont montré l'utilité de transformer les descripteurs bas d'une classification, pour cela nous réalisons une modification de l'espace des descripteurs. Tout d'abord la distribution de chacun des descripteurs est standardisée en fixant sa moyenne à 0 et sa variance à 1. Puis, pour chacune des classifications, un modèle de fusion adapté à l'ensemble des descripteurs bas associé est étudié.

Classification voix/musique

Les modèles de transformation de l'espace des descripteurs sont appris sur le regroupement du concept $M =$ *parole seule/mélange(parole-autre)/musique/bruit de fond* dont les classes correspondent à celles de la classification hiérarchique étudiée. Notre but étant de garder les mêmes descripteurs bas pour l'ensemble de la hiérarchie. Remarquons que dans le cas de classification en arbre plus complexe, cela risquerait de nuire à la classification, car certains descripteurs pourraient s'avérer non adéquats pour certains concepts. Il est donc indispensable dans ce cas de choisir des ensembles de descripteurs différents pour chaque concept de la hiérarchie.

$B = \{B_1, \ldots, B_m\}$ est l'ensemble des descripteurs bas associés aux concepts de la classification. A ce stade, $m = 96$. Il semble nécessaire d'éliminer certains descripteurs. En effet, utiliser un grand nombre de descripteurs peut réduire les performances du système lorsque certains d'entre eux contiennent peu d'information sur la classification considérée. De plus, le calcul de ces 96 descripteurs est prohibitif et ralentit fortement le traitement des données.

Nous testons plusieurs algorithmes d'**élimination de descripteurs**. Ils sont fondés sur l'Information Mutuelle, la PCA et la LDA : les descripteurs qui contiennent le moins d'information sur la classification de M sont éliminés. L'analyse des résultats de tests réalisés sur une base de données annotée permettra de conclure sur l'utilité de l'élimination de descripteurs et sur la validité des modèles d'information.

Afin d'améliorer la tâche de classification, il est souvent nécessaire de **modifier l'espace des descripteurs**. De façon générale, ceci est fait par projection de l'espace vers un nouvel espace numérique de même dimension. Cette transformation a été développée afin de réaliser deux tâches principales : rendre les descripteurs moins dépendants (PCA) ; et augmenter la séparabilité entre classes (LDA).

L'analyse des classifications utilisant ces techniques de fusion des données permettra d'établir certaines propriétés sur les avantages de l'élimination de descripteurs et sur l'utilité de la transformation de l'espace. De plus, nous conclurons sur la position de chacun de ces algorithmes dans la chaîne de traitement des données.

Reconnaissance de personnages

Soit l'ensemble des descripteurs bas associé au concept « nom du personnage ». A ce stade, il comprend $m = 68$ descripteurs. Les descripteurs bas employés pour cette classification sont déterminés à partir du même signal, le spectre MFCC du segment considéré. Dans ce cas, l'extraction du spectre est accomplie une fois pour tous les descripteurs, il n'est donc pas nécessaire d'appliquer un algorithme d'**élimination**, puisque le temps de calcul économisé est insignifiant.

Afin d'améliorer la tâche de classification, nous **modifions** l'espace des descripteurs pour réaliser la décorrélation des descripteurs (PCA) et l'adaptation aux 8 classes (LDA).

Nous avons vu que l'existence d'un grand nombre de descripteurs peut perturber la tâche de classification, c'est pourquoi nous cherchons à **diminuer** la dimension de l'espace des descripteurs par transformation linéaire fondée sur la PCA ou la LDA.

La transformation des données permet d'obtenir un ensemble de l descripteurs $B' = (B'_1, \ldots, B'_l)$. L'analyse de plusieurs classifications utilisant ces techniques de fusion de données numériques permettra de conclure sur l'utilité de la transformation de l'espace des données avec et sans diminution de la dimension.

5.1.4 Late/early fusion

Dans le paragraphe 2.4, la combinaison parallèle de classifications a été proposée comme une voie de recherche permettant d'améliorer les performances de détermination de concepts. Nous testons cette technique sur la première classification parole/musique. La « late » fusion envisage de modéliser un concept pour chaque descripteur de façon indépendante et de fusionner la décision de ces classifications. Cette fusion est réalisée par classification SVM du vecteur score de ces n classifications.

La comparaison des résultats de late et early fusion pour la classification proposée permettra de conclure sur l'efficacité de la combinaison de classifications.

5.2 Expérience et validation des hypothèses de modélisation

Dans notre première expérience, nous étudions les résultats de classification de plusieurs modèles de classification. Ces classifications visent à extraire la valeur estimée d'un concept pour un élément de structure de film à partir de descripteurs bas.

5.2.1 Base de données

Cette expérience nécessite la création d'une base de données annotées selon les concepts choisis. Pour chaque classification, cette base est séparée en deux. La première partie d'objets sert à l'apprentissage des modèles de traitement des descripteurs et de classification. La deuxième partie sert de base de test des classifications.

Classification voix/musique

La base de données est constituée de 500 segments de sons de 3s. Ils sont extraits à partir de nombreuses sources : plus particulièrement de films, mais aussi de programmes TV et de radio.

Le son est échantillonné à 22,5kHz, quantifié sur 16 bits et mixé en mono. Les segments d'audio sont annotés manuellement par les concepts de la classification hiérarchique.

Nous avons porté une attention particulière pour que la base représente une grande variété de signaux. Ainsi la classe de parole est composée de voix d'hommes ou de femmes, parlant en Anglais, Français, Allemand et Japonais. La classe de musique comprend de nombreux styles musicaux : jazz, pop, rap, techno, classique, avec ou sans parole. La classe des sons environnementaux contient plusieurs ambiances de ville, nature, café, bureau... De plus, nous avons essayé d'obtenir un même nombre d'exemples pour chaque classe.

Pour cette expérience, la base d'apprentissage comporte 300 sons et la base de tests 200.

Reconnaissance du personnage

La base de données pour la reconnaissance du locuteur est constituée de 400 clips d'audio de taille variable de 3s à 21s. Ils sont extraits à partir de quatre films dont le choix n'a pas été vraiment réfléchi : *Belle maman, Tomb Raider, American pie, Les enfants du marais*.

Le son est échantillonné à 22,5kHz, quantifié sur 16 bits et mixé en mono. Les clips sont obtenus par classification HMM du signal des films. Les segments homogènes sont annotés à la main du nom du personnage qui parle ou de « autre ».

Toutes les voix sont en Français. Une variété de personnages est représentée : hommes ou femmes, d'âges variés. Les films considérés sont issus de plusieurs styles cinématographiques : comédie, action..., ce qui a plusieurs incidences sur les caractéristiques des segments appris : le fond sonore derrière les voix est assez varié (musique, bruits), notamment. Nous avons essayé d'obtenir un même nombre d'extraits pour chaque voix apprise, et un plus grand nombre de voix classées dans « autre ».

Pour cette expérience la base d'apprentissage comporte 250 sons et la base de tests 150.

5.2.2 Critères d'évaluation

Afin de comparer les différents modèles, nous analyserons leurs résultats de classification. Nous définissons la matrice de confiance : $\mathbf{C} = \mathbf{C}_{ij}$ où $\mathbf{C}_{ij}$ est le pourcentage d'objets annotés de la $i^{ème}$ classe et classés par le modèle dans la $j^{ème}$. Ainsi les éléments $\mathbf{C}_{ii}$ de la diagonale correspondent au taux de bonnes classifications pour la classe i. Le taux de classification globale $\mathbf{GC}$ est définit comme la moyenne arithmétique des $\mathbf{C}_{ii}$.

5.2.3 Résultats

Classification voix/musique

Pour l'**élimination de descripteurs**, deux algorithmes sont comparés. Ils sont fondés sur l'analyse discriminante linéaire (LDA) et l'information mutuelle (MI). Nous choisissons, pour le test, de garder les 5 dimensions (sur les 96 initiales) qui contiennent le plus d'information concernant le concept choisi.

Pour chacun des algorithmes, les descripteurs sélectionnés pour la suite des opérations sont présentés dans le tableau 5.1 dans l'ordre de classement.

MI	$MFCC13.1$	$SPF1.1$	$4ME1.1$	$MFCC6.0$	$SPF1.0$
swLDA	$MFCC13.1$	$4ME1.1$	$SPF1.0$	$SPF1.1$	$MFCC12.0$

TAB. 5.1 – Descripteurs sélectionnés par les deux algorithmes

Il s'agit de l'écart type de la modulation à 4Hz ($4ME1.1$), du flux spectral ($SPF1.1$) et de la 13ème dimension des MFCC ($MFCC13.1$) ; et de la moyenne du flux spectral ($SPF1.0$) et de la 12ème dimension des MFCC ($MFCC12.0$).

Les descripteurs sélectionnés par les deux algorithmes sont sensiblement les mêmes. La LDA et l'information mutuelle caractérisent la même information : la discrimination des classes. C'est pourquoi, nous n'observons que peu de variation dans le classement des descripteurs. Tous concernent des caractéristiques essentielles de la parole [Sch00a]. Prenons l'exemple de la modulation à 4Hz, le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hz. Ce paramètre a des valeurs plus élevées pour les segments de parole que pour les segments sans parole.

Les matrices de confiance de la classification du premier concept $M_1 =$ *parole/nonparole* sont présentées dans les tableaux 5.2. Ils affichent des performances variées selon le modèle choisi, nous pouvons en tirer deux observations.

classe	96 descripteurs		5 des. MI		5 des. LDA	
parole	**95**	5	**92**	8	**92**	8
non parole	7.5	**92.5**	13.5	**86.5**	15.5	**84.5**

TAB. 5.2 – Performances de classification *parole/non parole* par les descripteurs sélectionnés

Premièrement, la classification par l'intégralité des descripteurs obtient de meilleurs résultats que celle réalisée à partir des 5 descripteurs sélectionnés. L'utilisation d'un plus petit nombre de descripteurs tend à réduire la quantité d'information disponible pour discriminer les classes, ce qui diminue les performances de classification. Cette perte d'information peut s'avérer importante. Notamment, pour la classe non-parole, nous constatons de l'ordre de 7% d'écart dû au fait que les descripteurs sélectionnés caractérisent plutôt la présence de parole. Lors d'expériences complémentaires, nous avons observé une réduction de cet écart lorsque plus de descripteurs sont sélectionnés et ceci de manière exponentielle.

Deuxièmement, la sélection des descripteurs par l'Information mutuelle (MI) semble fournir de meilleurs résultats que celle par analyse discriminante linéaire (LDA). Mais l'écart n'est pas très significatif puisqu'il est de 2% au maximum pour la classe non-parole. Nous sélectionnons donc les 5 descripteurs choisis par l'Information mutuelle pour la suite.

Pour la **transformation de l'espace** des descripteurs, deux algorithmes sont testés. Leur but est de modifier les données afin de décorréler les dimensions (PCA) ou de les adapter aux classes (LDA).

Sans modification

1.00	0.66	0.59	0.36	0.16
0.66	1.00	0.61	0.66	0.04
0.59	0.61	1.00	0.36	-0.01
0.36	0.66	0.36	1.00	-0.06
0.16	0.04	-0.01	-0.06	1.00

PCA

1.00	0.10	-0.39	0.06	-0.02
0.10	1.00	-0.18	0.20	0.79
-0.39	-0.18	1.00	0.10	-0.01
0.06	0.20	0.10	1.00	0.40
-0.02	0.79	-0.01	0.40	1.00

LDA

1.00	-0.75	0.46	0.51	-0.32
-0.75	1.00	-0.71	-0.81	-0.02
0.46	-0.71	1.00	0.45	0.14
0.51	-0.81	0.45	1.00	0.15
-0.32	-0.02	0.14	0.15	1.00

TAB. 5.3 – Matrices de corrélation des 5 descripteurs sélectionnés

Le tableau 5.3 montre la corrélation des 5 descripteurs sélectionnés avant et après application de l'algorithme de transformation.

Nous remarquons que les descripteurs avant transformation sont assez corrélés. Après une PCA, la corrélation des descripteurs est fortement diminuée, c'est le but recherché. Après une LDA, réalisée avec ou sans PCA, les descripteurs restent fortement corrélés, ce qui peut avoir une incidence pour certains algorithmes de classification qui ne prennent pas en compte les relations entre descripteurs (e.g. classification naïve de Bayes).

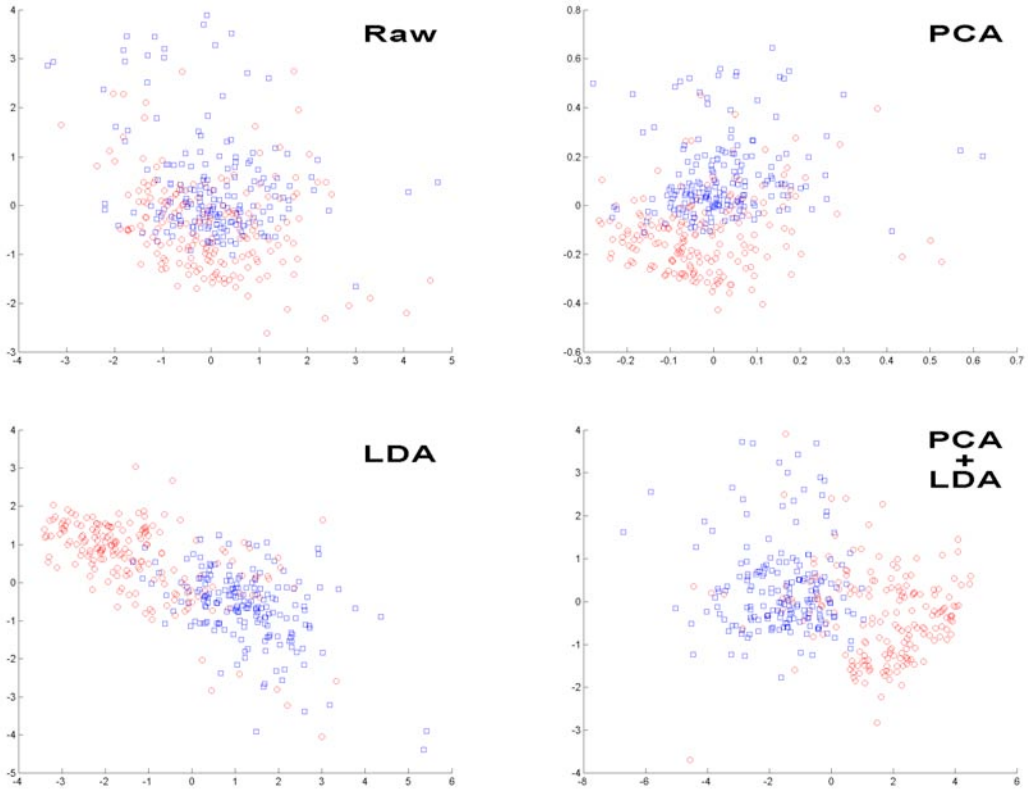

FIG. 5.2 – Espace des descripteurs avant et après la transformation.

La figure 5.2 représente les points de la base d'apprentissage pour les deux premières dimensions du vecteur descripteur. Les segments de paroles sont représentés en bleu et les segments de non-parole en rouge.

Nous constatons tout d'abord que la discrimination entre les deux classes est visible à l'œil nu. Pour la plupart des points la classification semble évidente, sauf pour certains d'entre eux situés dans la « zone de séparation » des deux classes. La PCA tend à rendre les distributions gaussiennes et à variances fixes (égale à 1). Cependant, elle semble amener une perte de discrimination par rapport à l'espace initial. Nous remarquons aussi que la LDA augmente fortement la séparation entre les classes, et que la décorrélation des descripteurs par PCA avant une transformation par LDA la diminue.

classe	PCA		LDA		PCA+LDA	
parole	**93**	7	**96.5**	3.5	**97**	3
non parole	12.5	**87.5**	6.5	**93.5**	5.5	**94.5**

TAB. 5.4 – Performances de classification *parole/non parole* par les descripteurs transformés

Les matrices de confiance de la classification du concept *parole/nonparole* par les descripteurs transformés sont présentées dans le tableau 5.4. Selon le modèle choisi, les performances varient. Deux remarques peuvent être faites.

Premièrement, la transformation des descripteurs par analyse en composante principale (PCA) amène une amélioration non-significative des performances de classification (1% environ). La décorrélation ne semble donc pas apporter d'avantages certains pour l'algorithme SVM.

Deuxièmement, l'analyse linéaire discriminante (LDA) conduit à une augmentation sensible des per-

formances de classification : 8% pour la classe non-parole. L'information sur la distribution des classes apportée par la LDA, augmente la discrimination des données et ajoute, ainsi, une connaissance supplémentaire au modèle de classification SVM.

Pour la suite nous considèrerons que la transformation optimale de l'espace des descripteurs comporte une PCA, suivie d'une LDA.

Afin d'évaluer l'intérêt de la classification par **late fusion** des descripteurs bas, les performances obtenues par ce modèle et le modèle de early fusion sont comparées. Le tableau 5.5 montre les taux de bonne classification des deux modèles appliqués au concept *parole/non-parole*.

	parole	non parole
early fusion	92	**86.5**
late fusion	**93**	86

TAB. 5.5 – Performances des modèles de fusion "early" et "late"

Ce tableau dévoile que les performances varient peu selon le modèle choisi. Au vu de ces résultats et de ceux observés par d'autres études, nous remarquons que la late fusion n'apporte pas d'amélioration significative des performances de classification, dans le cadre de la fusion de descripteurs bas.

Reconnaissance du personnage

Il s'agit ici de la classification du concept haut : nom du personnage. H peut prendre 8 valeurs, 7 sont des noms d'acteurs, et la $8^{\text{ième}}$ classe correspond aux voix « autres ».

personage	1	2	3	4	5	6	7	0	**GC**
rien	89	86	85	86	84	68	88	80	$\mathbf{83.2}$
PCA+LDA	94	92	89	91	91	90	95	89	$\mathbf{91.4}$

TAB. 5.6 – Performances de la reconnaissance du personnage par le son

Le tableau 5.6 présente les taux de bonnes classifications pour les 8 classes de la taxonomie. La première ligne montre les résultats obtenus sans **transformation** de l'espace des descripteurs, la deuxième expose les performances de classification après fusion des descripteurs par PCA, puis LDA. Ce tableau dévoile que les performances varient selon le modèle choisi.

Tout d'abord, la détermination automatique du nom du locuteur par classification SVM des descripteurs obtient de bonnes performances. Ces résultats, compte tenu de la présence de bruit de fond, correspondent à ceux obtenus par d'autres chercheurs, notamment pour des applications dans la sécurité.

Ensuite, le traitement des données par PCA+LDA apporte une amélioration significative des performances de classification, notamment en ce qui concerne la classe « autre ». Comme pour la classification parole/musique, la transformation des descripteurs permet de mieux prendre en compte la discrimination entre les classes.

Pour la **réduction de la dimension** de l'espace des descripteurs, deux algorithmes sont testés. Ils consistent à transformer l'espace des descripteurs initial par projection vers un espace de dimension inférieure. Les axes de ce nouvel espace correspondent aux k directions «principales » (PCA) ou « discriminantes » (LDA) de l'ensemble des données.

Les taux de classification pour la classification du concept par $k = 11, 30, 68$ descripteurs, sont présentés dans le tableau 5.7. Ils montrent que les performances varient selon la dimension choisie pour l'espace des données.

nbre de des.	11	30	68
PCA	67	76	84.5
LDA	79	85	91

TAB. 5.7 – Performances de la reconnaissance du personnage après réduction de la dimension

Remarquons, tout d'abord, que la réduction de la dimension entraîne une diminution sensible des performances de classification. La perte de l'information contenue dans les dimensions éliminées semble corrompre fortement la classification. C'est le cas notamment pour le traitement par PCA, nous constatons un écart de 16% entre la classification par 11 et 68 descripteurs.

Deuxièmement, le traitement des données par LDA, apporte des résultats meilleurs comparés à ceux obtenus par PCA (8% environ). Cela est dû au fait que les dimensions de projections correspondent, pour la PCA, aux dimensions qui expliquent le maximum de variance des données. Or, la variance ne donne aucune information sur la discrimination des classes : une forte variance assure uniquement une bonne discrimination entre les points. Dans certains cas, une dimension inadéquate pour la classification peut être considérée comme un axe « principal ». La LDA détermine les directions « discriminantes » pour le concept choisi, c'est pourquoi les descripteurs sélectionnés montrent de bonnes performances de classification. Nous constatons, notamment, un écart assez faible (5%) entre les résultats obtenus pour $k = 30$ et $k = 68$ (tous les descripteurs).

5.2.4 Résumé des résultats

En résumé, les résultats de ce chapitre assurent que les modèles de fusion de descripteurs bas capturent bien les caractéristiques utiles pour la classification. Les performances obtenues pour la classification *parole/musique* sont comparables à celles des meilleurs algorithmes existants. En ce qui concerne la reconnaissance de la voix de personnages, les algorithmes testés donnent des résultats bons, équivalents à ceux obtenus aujourd'hui par les chercheurs.

Les exemples de ce chapitre nous permettent de conclure sur les hypothèses prises précédemment :

La **réduction de la dimension** de l'espace des descripteurs entraîne une forte perte d'information nécessaire à la classification. Dans notre cas, où le nombre de descripteurs reste faible (inférieur à 100), elle n'est justifiée que lorsque le temps de calcul est trop important pour la classification choisie (c'est le cas, pour le premier modèle). Dans cette situation, l'élimination des descripteurs fondée sur l'information mutuelle semble la plus efficace.

La **transformation de l'espace** des descripteurs sans réduction de la dimension par PCA, ou LDA, apporte une amélioration des performances de classification. La fusion des données LDA, parce qu'elle ajoute de l'information sur les classes au système d'apprentissage, fait nettement progresser les performances. En revanche, la fusion par PCA, n'entraîne que peu d'amélioration ; elle semble en effet redondante avec certaines opérations de fusion réalisées par les SVM.

La **« late » fusion** dans le cadre de la classification de concepts monomédias par des descripteurs bas n'amène pas de meilleurs résultats que la « early » fusion. En effet, cette technique semble plus appropriée pour la fusion de descripteurs bas et moyens et la fusion de descripteurs provenant de médias différents,

Ces résultats seront considérés comme admis pour les expériences présentées dans les chapitres suivants. Les exemples de ce chapitre servent de repères expérimentaux des performances des classifications utilisées dans le cadre de la segmentation de film (chapitre 8).

Chapitre 6

Fusion du contenu bas et moyen monomédia

Le but de ce sixième chapitre est de valider nos hypothèses concernant la fusion de descripteurs bas et moyens pour la catégorisation automatique du contenu haut. Dans le paragraphe 6.1 nous présentons les modèles de classification appris. Puis, dans le paragraphe 6.2, nous comparons les résultats de classification obtenus pour les concepts hauts choisis. L'expérience de classification des descripteurs nous conduit en particulier à sélectionner les modèles de fusion utilisés par la suite parmi tous les modèles testés.

6.1 Modèle de classification

Afin de valider ou non les hypothèses concernant la fusion de descripteurs bas et moyens, les résultats de determination automatique du lieu d'une image sont étudiés. Pour nos expériences, nous avons choisi une classification hiérarchique nécessaire à l'analyse de films. Elle est présentée dans la figure 6.1. Les caractéristiques des tâches d'indexation permettent de traiter l'ensemble des problèmes posés.

6.1.1 Le contenu utilisé

Le modèle de classification présenté est un modèle classique de concepts hauts image. Il s'agit d'une classification hiérarchique des images en lieu. La détermination du lieu où a été capturée une image ou une séquence d'images est une tâche essentielle du traitement des films.

De nombreuses recherches considèrent le modèle de concept haut d'image *intérieur/extérieur*. Pour notre système, il est le premier concept d'une annotation plus vaste qui concerne la caractéristique de lieu. L'ensemble H des concepts de cette classification comprend trois éléments :

- $H_1 =$ *intérieur/extérieur*
- $H_2 =$ *absent/parking/maison/magasin*
- $H_3 =$ *absent/nature/ville*

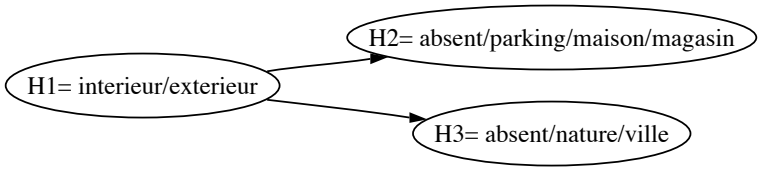

FIG. 6.1 – Modèle de classification hiérarchique des lieux.

Les **descripteurs numériques** utilisés pour la classification sont extraits à partir de l'intégralité de l'image. Tous les concepts hauts sont déterminés à partir des mêmes descripteurs bas. Soit $B = \{B_1, \ldots, B_m\}$, l'ensemble de ces descripteurs. Ils se présentent sous la forme d'un histogramme composé de deux parties : une partie texture et une partie couleur. La partie texture est fondée sur les motifs des contours locaux (local-edge pattern LEP). La partie couleur, quant à elle, est un histogramme RGB : R, G et B sont quantifiés chacun en 4 valeurs, ce qui donne un histogramme de 64 composantes. L'histogramme final est donc constitué de 576 composantes.

Les **descripteurs moyens** sont determinés par annotation manuelle ou automatique de l'image considérée. L'ensemble M des concepts moyens utilisés pour la classification des lieux comprend 8 concepts :

- $M_1 =$ bâtiment $= 1/0$
- $M_2 =$ voiture $= 1/0$
- $M_3 =$ lumière artificielle $= 1/0$
- $M_4 =$ route $= 1/0$
- $M_5 =$ ciel $= 1/0$
- $M_6 =$ herbe $= 1/0$
- $M_7 =$ arbre $= 1/0$
- $M_8 =$ personne $= 1/0$

Ces concepts sont représentés par des variables binaires. Celles-ci valent 1 lorsque l'objet ou la texture associée au concept est présent, 0 sinon. Dans le cadre de notre expérience, les 7 premiers concepts sont déterminés par annotation manuelle et le dernier (« personne ») par détection automatique de visage. L'algorithme, développé par C.Millet au CEA, est fondé sur les techniques « Adaboost » [Vio02]. En ce qui concerne les autres concepts $\{M_1, \ldots, M_7\}$, plusieurs recherches ont permis de rendre automatique la tâche d'annotation. Les algorithmes employés sont fondés sur la classification de textures (herbe, ciel, arbre, route, bâtiment) ou de formes (voiture). De facon générale, ceux-ci obtiennent des performances de détection de l'ordre de 80% de bonnes classifications. Mais faute de disposer à l'époque d'algorithmes performants, l'annotation est réalisée manuellement.

6.1.2 Modèle statistique

Les modèles statistiques présentés au chapitre 4 sont appliqués pour la détection des concepts hauts de la classification des lieux. Plusieurs modèles de fusion des données sont testés afin de valider nos hypothèses de modélisation.

Modèle naïf de Bayes

Pour un concept donné H_j, les relations de dépendance entre les concepts moyens et les descripteurs bas et moyens associés sont représentées par le réseau bayésien en figure 6.2. Les hypothèses de ce

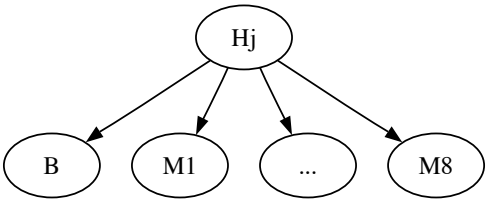

FIG. 6.2 – Modèle naïf de Bayes de classification d'un concept haut.

modèle sont :

1. Les descripteurs bas $\{B_1, \ldots, B_m\}$ associés à un concept H_j ne sont pas indépendants sachant ce concept. Cette hypothèse a été vérifiée dans le chapitre précédent pour deux classifications audio.

2. Les concepts moyens associés au concept haut H_j sont indépendants sachant ce concept.
3. L'ensemble des descripteurs bas B associés à un concept haut H_j est indépendant des concepts moyens associés, sachant ce concept.

Afin de valider ou non les deux dernières hypothèses de modélisation, ce modèle sera comparé aux modèles de fusion décrits dans le chapitre 4. En premier lieu l'hypothèse 2 d'indépendance des concepts moyens est testée par un modèle de fusion des scores. Ensuite la fusion des descripteurs bas et moyens est expérimentée.

Fusion des concepts moyens

Dans les faits, l'hypothèse d'indépendance entre concepts moyens ne semble pas valable. Le modèle naïf de Bayes surévalue l'importance des concepts dépendants. C'est pourquoi nous appliquons un modèle de classification qui prend en compte la dépendance entre concepts atomiques. Pour un descripteur haut donné, les relations de dépendance entre les concepts moyens et les descripteurs bas associés sont représentées par le réseau bayésien dans la figure 6.3.

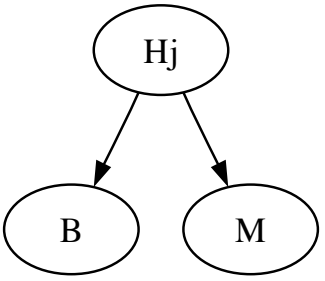

FIG. 6.3 – Modèle de fusion des concepts moyens pour la classification d'un concept haut.

Un modèle de classification SVM des scores obtenus par les concepts moyens permet d'évaluer les probabilités conditionnelles associées.

"Early" fusion des descripteurs bas et moyens

Pour simplifier, nous avons supposé que les descripteurs bas de l'ensemble B étaient indépendants des concepts moyens de M sachant H. Cependant, il est possible de remettre en cause cette hypothèse. Dans ce cas, pour un descripteur haut donné, les relations de dépendance entre les concepts moyens et les descripteurs bas associés sont représentées par le réseau bayésien $H \rightarrow \{M, B\}$.

En pratique il est possible de fusionner les descripteurs bas et moyen au sein d'un descripteur global du concept. Ce descripteur est représenté par un vecteur combinaison du vecteur score $\mathbf{S}$ des concepts et du vecteur $\mathbf{B}$ des descripteurs bas. Et un modèle SVM appris sur la base annotée permet la détermination de la probabilité jointe globale.

Il est aussi envisageable de réaliser la fusion des descripteurs par un modèle de « late » fusion décrit au chapitre 4.3.2.

"Late" fusion des descripteurs bas et moyens

On l'a vu, la combinaison parallèle de classifications a été proposée comme une voie de recherche permettant d'améliorer les performances de détermination de concepts. La late fusion envisage de modéliser un concept haut « intermédiaire » H_M pour l'ensemble des concepts moyens et un concept H_B pour les descripteurs bas. Les décisions de classification de ces concepts sont ensuite fusionnées par un modèle de classification (naïf de Bayes ou SVM).

Si on suppose que les deux concepts H_M et H_B ne dépendent que de H (modèle naïf de Bayes), alors les relations de dépendance sont représentées par le réseau bayésien de la figure 6.4. Cependant, nous préférerons ne pas considérer cette hypothèse et réaliser la classification SVM des scores des deux concepts.

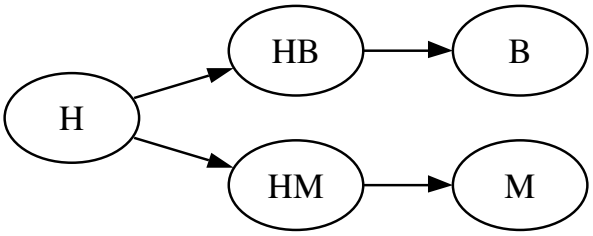

FIG. 6.4 – Modèle de late fusion des concepts moyens et des descripteurs bas.

6.2 Expérience et validation des hypothèses de modélisation

Dans notre deuxième expérience nous étudions les résultats de classification de ces modèles de concepts hauts. La classification a pour but d'extraire la valeur estimée des concepts hauts pour un élément de structure de film à partir de descripteurs bas et moyens.

6.2.1 Base de données

Cette expérience nécessite la création d'une base de données annotée selon les concepts choisis. Cette base est séparée en deux : une base d'apprentissage et une base de test. La première partie des objets sert à l'apprentissage des modèles de normalisation des scores et de fusion par réseau bayésien ou SVM. La dernière partie est la base de test des classifications.

La base de données est constituée de 400 images fixes tirées de vidéos personnelles, de films et de bases de données du CEA. Elle est annotée manuellement des concepts de la classification. Nous avons porté une attention particulière pour que chaque classe de la taxonomie contienne une quantité représentative d'images à apprendre, et que le nombre d'images ne varie pas trop au sein des concepts.

Pour cette expérience la première base comporte 300 images et la base de test 100.

6.2.2 Critères d'évaluation

Afin de comparer les différents modèles, nous analyserons leurs résultats de classification. La matrice de confiance est définie : $\mathbf{C} = \mathbf{C}_{ij}$ ou $\mathbf{C}_{ij}$ est le pourcentage d'objets annotés de la $i^{ème}$ classe et classés par le modèle dans la $j^{ème}$. Ainsi les éléments $\mathbf{C}_{ii}$ de la diagonale correspondent au taux de bonnes classifications pour la classe i. Le taux de classification globale $\mathbf{GC}$ est défini comme la moyenne arithmétique des $\mathbf{C}_{ii}$.

6.2.3 Résultats

Classification par les descripteurs bas

Tout d'abord, nous considérons la classification par les **descripteurs bas** (TestureLEP) du premier concept haut de la hiérarchie. Afin de mettre en valeur les performances des techniques de fusion envisagées nous ne modifierons pas les descripteurs dans ce cas. La matrice de confiance de la classification du premier concept $H_1 =$ *intérieur/extérieur* est présentée à gauche du tableau 6.1. Pour comparaison,

les performances de classification du même algorithme testé par [Mil04] sur une base de 20000 images, dont 2500 d'apprentissage, sont présentées à droite.

intérieur	**82**	18	**89.8**	10.2
extérieur	15	**85**	9.2	**90.8**

TAB. 6.1 – Performances de classification *intérieur/extérieur* par les descripteurs bas

Ces tableaux montrent que les performances obtenues varient pour les deux tests. Deux observations peuvent être faites.

Premièrement, la classification des images d'extérieur fournit de meilleurs résultats que celle des images d'intérieur pour les deux tests. Les photographies d'intérieur mal classées correspondent à des photos dont les lumières et les couleurs semblent révéler une image d'extérieur. Il s'agit principalement de photos comportant une fenêtre, d'images dont le haut est plus clair que le bas indiquant la présence d'un "ciel" ou de photos de grands espaces très éclairés. Les photographies d'extérieur mal classées représentent, en général, des gros plans sur des objets ou des paysages dont la vue est obstruée par un élément du décor sombre : un mur, une forêt. Ainsi, il semble que ce genre de cas "pathologiques" soient plus courants pour les images d'intérieur que pour celle d'extérieur, ce qui expliquerait une telle différence dans les performances observées.

Deuxièmement, nous remarquons, que les performances fournies par C.Millet sont bien supérieures aux nôtres. Nous supposons que cette différence est dûe à l'écart de la quantité d'images d'apprentissage. Notre base comportant moins d'objets d'exemple, le modèle de classification se fait une représentation moins précise des classes.

Classification par les descripteurs moyens

En suite, nous expérimentons la classification par les **descripteurs moyens** du premier concept haut de la hiérarchie. La catégorisation automatique du concept haut H_1 est réalisée par classification SVM des scores obtenues par les concepts moyens $\{M_1, \ldots, M_8\}$.

Ces scores proviennent de plusieurs modèles de classification, il est donc nécessaire de les **normaliser**. La détection des visages produit un coefficient de confiance à valeur réelle centré sur 0 : Le concept est présent si le taux est supérieur à 0, absent sinon. Afin de normaliser ces scores et de les projeter sur l'intervalle $[0, 1]$, la fonction sigmoïde de Genoud est appliquée.

La figure 6.5 montre les distributions des scores de présence de visage de la base d'apprentissage de la classification des lieux avant (en haut) et après (en bas) la normalisation par la fonction de Genoud. Nous remarquons que la normalisation du score de classification entraîne une modification forte de sa distribution. La forme de la courbe de Genoud augmente la séparabilité des points dont le score est proche de la valeur seuil 0 et diminue celle des points dont la classification est plus évidente.

Ainsi pour chaque objet de la base, nous obtenons un vecteur score des concepts moyens normalisés. Au $i^{ème}$ score correspond la pseudo probabilité jointe que le $i^{ème}$ concept soit présent sachant les descripteurs bas associés à ce concept.

Les matrices de confiance de la classification de $H_1 =$ *intérieur/extérieur* par les concepts moyens sont présentées dans les tableaux 6.2. Le tableau à gauche montre les résultats obtenus par le modèle naïf de Bayes ; celui de droite expose les performances de classification SVM des scores.

intérieur	**86**	14	**87**	13
extérieur	11	**89**	9	**91**

TAB. 6.2 – Performances de classification *intérieur/extérieur* par les descripteurs moyens

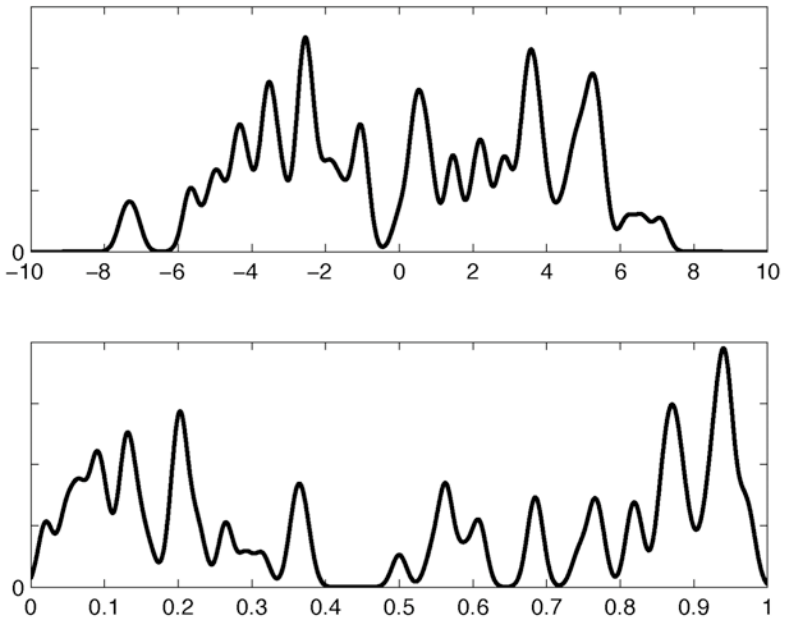

FIG. 6.5 – Distribution des scores de présence de visage des images de la base d'apprentissage

Ces tableaux dévoilent que les performances varient peu selon le modèle choisi. Cependant nous notons plusieurs observations.

Tout d'abord la catégorisation automatique du lieu d'une image par des concepts moyens obtient de bonnes performances. Comparées à celles obtenues par classification des descripteurs bas, nous constatons une augmentation nette de 5% environ. Les concepts moyens semblent contenir plus d'information sur la classification que les descripteurs bas TLEP. Mais l'utilisation de ces concepts, extraits de facon manuelle ou automatique, entraîne une augmentation du temps de calcul (et d'annotation).

Cependant, pour certaines images, la classification reste complexe. Nous remarquons par exemple qu'une photographie de parking sous-terrain contenant une voiture est classée comme extérieur, ce qui constitue une erreur. Cela est dû au fait que la plupart des voitures présentes dans les images de la base d'apprentissage sont à l'extérieur. Lorsque la présence de certains concepts semble fortement indiquer la classe de l'image, le système commet des erreurs.

Remarquons aussi que l'absence de concept dans une image ne donne en réalité que peu d'indication sur la nature du lieu. Or, dans les faits, les deux algorithmes classent une telle image en "intérieur". Ce qui entraîne des erreurs pour les images d'extérieur ne contenant aucun des concepts choisis.

Enfin, la fusion des données par le modèle naïf de Bayes et par SVM semble fournir des résultats similaires. Toutefois, nous remarquons un léger avantage pour le modèle de classification des scores par SVM (1% environ) dû à une meilleure prise en compte des relations de corrélation entre concepts. Par exemple, une image d'extérieur contenant un arbre et de l'herbe, mais dont la détection automatique se trompe et obtient pour ces concepts un taux de confiance de 45%, est classée en intérieur par le modèle naïf et en extérieur par les SVM.

Fusion des descripteurs moyens et bas

Afin de mettre en valeur les améliorations de classification par fusion des descripteurs bas et moyens, nous effectuons la suite de ces tests sur les deux autres concepts hauts de la classification hiérarchique des lieux, i.e. $H_2 =$ *parking/maison/magasin* et $H_3 =$ *nature/ville*. Nous utilisons à partir de maintenant le modèle SVM pour la fusion des concepts moyens.

Les matrices de confiance de la classification de H_2 et H_3 par les **concepts moyens** sont présentées dans le tableau 6.3. A gauche, le tableau montre les résultats obtenus pour la classification des images d'intérieur en *Parking/Maison/Magasin*. Et à droite, il présente ceux obtenus pour la classification des images d'extérieur en *Nature/Ville*.

Parking/Maison/Magasin		
22.5	0	78.5
0	0	100
0	0	100

Nature/Ville	
62.5	37.5
12.0	88.0

TAB. 6.3 – Performances de classification des lieux par les descripteurs moyens

Ce tableau montre des performances disparates selon le concept considéré. Plusieurs remarques peuvent être faites.

Premièrement, la classification de H_2 par les concepts obtient de mauvais résultats. Les descripteurs moyens utilisés décrivent des objets (et textures) d'extérieur et une grande partie des images d'intérieur ne contiennent aucun des concepts. C'est pourquoi, il semblent inadéquats pour la classification des images d'intérieur. Seuls quelques éléments de la classe "parking sous-terrain" ont été bien classés grâce à la présence de voitures dans l'image.

Deuxièmement, la classification de H_3 en *nature/ville* obtient d'assez bons résultats, Les concepts moyens employés pour la classification réalisent une bonne discrimination des classes.

Pour comparaison, les matrices de confiance de la classification des concepts H_2 et H_3 par les **descripeurs bas** sont présentées dans le tableau 6.4.

Parking/Maison/Magasin		
70.5	0	29.5
0	79	21
0	15	85

Nature/Ville	
87.5	13
9	92

TAB. 6.4 – Performances de classification des lieux par les descripteurs bas

Nous remarquons que pour les deux concepts hauts, la classification des descripteurs bas montre de meilleures performances que celle des concepts moyens. Les descripteurs TLEP réalisent donc une assez bonne discrimination des classes des deux concepts hauts H_2 et H_3. Cependant, ces résultats ne sont pas encore très bons. Et rappelons qu'en ce qui concerne la classification en *intérieur/extérieur*, les concepts ont obtenu de meilleurs résultats que TLEP. Notre hypothèse est que les informations contenues dans les descripteurs bas et dans les concepts moyens sont **complémentaires**. Ainsi, la fusion des deux ensembles de descripteurs pour la classification doit entraîner une amélioration des performances.

Nous expérimentons donc enfin la classification par fusion des concepts moyens et des descripteurs bas. Plusieurs modèles de fusion ont été présentés dans le chapitre 4. Les matrices de confiance de la classification des concepts considérés sont présentées dans les tableaux 6.5, 6.6 et 6.7. Le premier tableau montre les resultats obtenus pour le modèle naïf de Bayes. Le deuxième tableau exhibe les mêmes resultats pour l'identification des lieux par classification SVM du vecteur descripteur global, produit de la concaténation des descripteurs bas et des scores. Le troisième tableau expose les performances obtenues par la « late » fusion.

Ces tableaux dévoilent que les performances varient peu selon le modèle choisi. Cependant, ces résultats amènent à plusieurs réflections.

Premièrement, la fusion des concepts moyens et des descripteurs bas améliore les performances de classification de la plupart des concepts de « lieu » (3% en moyenne). Nous constatons que les concepts

Intérieur/Extérieur	
87	13
10	90

Parking/Maison/Magasin		
50.0	5,5	45.5
2	60.5	37.5
2	15	83

Nature/Ville	
86.5	13.5
10	90

TAB. 6.5 – Performances de classification des lieux : modèle naïf de Bayes

Intérieur/Extérieur	
89	11
9	91

Parking/Maison/Magasin		
77	3.5	19.5
1	78.5	20.5
1	14	85

Nature/Ville	
86.5	13.5
7.5	92.5

TAB. 6.6 – Performances de classification des lieux : modèle "early" fusion

apportent une information supplémentaire comparé au modèle de classification par les descripteurs bas uniquement.

Deuxièmement, pour certaines classes, la fusion semble diminuer les performances par rapport au modèle de classification basse, ou moyenne, seule. En ce qui concerne le modèle naïf de Bayes et la late fusion, le problème se pose lorsque les deux classifications (des descripteurs et des concepts) ne sont pas «d'accord» sur la détermination d'un concept, il y a ambiguïté. Pour le modèle naïf, les deux probabilités de classification sont multipliées sans pondération. Lorsque l'un des modèles de classification fournit de mauvais résultats (e.g. H_2 par les concepts moyens), la fusion n'apporte pas d'amélioration des performances. Le modèle «late» fusion semble mieux considérer ce phénomème, mais ses performances pour le concept H_2 restent inférieures à celles du modèle de classification des descripteurs bas. Le regroupement au sein de concepts intermédiaires entraîne une légère perte de corrélation entre les descripteurs moyen et bas. Ce qui est dramatique, dans ce cas, puisque les descripteurs moyens amènent de nombreuses erreurs pour deux des trois classes du concept.

Troisièmement, pour le concept H_2, seul le modèle de fusion early n'entraîne pas de réduction des performances de catégorisation. Si une dimension du vecteur descripteur n'est pas adéquate pour la classification, elle n'influera pas sur celle-ci. Ainsi les concepts moyens non-discriminants pour le concept haut ne sont pas considérés. Le concept « voiture » est le seul descripteur moyen utile pour la catégorisation de H_2 : sa présence indique que l'on se trouve dans un parking. Pour cette classe, la fusion early apporte une amélioration des performances. Pour la classification en maison ou magasin, aucun des concepts n'est pris en compte et nous observons des performances équivalentes avec et sans fusion. Nous en concluons que, lorsque les concepts moyens fournissent de mauvaises performances de classification, seule la fusion SVM "early" améliore de façon significative les performances de classification par rapport au modèle bas niveau.

Quatrièmement, les tableaux montrent des résultats globalement plus mauvais pour le modèle de Bayes comparé aux deux autres. Ce modèle ne prend pas bien en compte les relations de corrélation entre les concepts moyens et les descripteurs bas. Les modèles de classification SVM "early" et " late" obtiennent des résultats équivalents. La "late" fusion montre des performances légèrement meilleures, comparés à la "early" fusion, sauf pour le deuxième concept haut. Nous en concluons que, lorsque les concepts moyens et les descripteurs bas fournissent de bonnes classifications, les fusions SVM ("early" et "late") augmentent de façon significative le pouvoir de discrimination du modèle de classification par rapport aux modèles moyens ou bas considérés individuellement.

6.2.4 Résumé des résultats

En résumé, les résultats de ce chapitre assurent que les modèles de fusion des descripteurs bas et des concepts moyens capturent bien les caractéristiques utiles pour la classification des concepts hauts. Les

Intérieur/Extérieur	
87.5	12.5
8	92

Parking/Maison/Magasin		
72.5	4	23.5
2	75.5	23.5
1	14	85

Nature/Ville	
86.5	13.5
7.5	92.5

TAB. 6.7 – Performances de classification des lieux : modèle "late" fusion

performances obtenues pour la classification d'images en lieux sont améliorées par la fusion.

Les exemples de ce chapitre nous permettent de conclure sur les hypothèses prises précédemment.

Les **concepts moyens** choisis sont discriminants pour la classification des concepts hauts $H_1 =$ *intérieur/extérieur* et $H_3 =$ *nature/ville*. En ce qui concerne $H_2 =$ *parking/maison/magasin*, seul le concept « voiture » fournit une information appropriée.

La **fusion des descripteurs** bas et moyens par les modèles SVM ("early" et "late") apporte une amélioration significative des performances de classification (par rapport au modèle bas niveau), lorsque les concepts moyens sont pertinents pour la description du concept haut considéré (H_1 et H_3).

Seule la **fusion "early"**, concaténation des scores et des descripteurs bas, améliore nettement les performances de classification de H_2 pour lequel les concepts moyens sont peu discriminants. Cela est dû principalement à sa capacité à traiter l'information apportée par chaque descripteur de façon individuelle, sans tenir compte de son niveau sémantique. Au contraire, la fusion "late" entraîne une légère perte de corrélation entre les descripteurs moyens et bas. Pour la suite, nous utiliserons cette technique pour la catégorisation du lieu où a été tourné un plan de film, car elle fournit les meilleures performances de localisation (en moyenne).

Ces résultats seront considérés comme admis pour les expériences présentées dans les chapitres suivants. Les exemples de ce chapitre servent de repères expérimentaux des performances des classifications utilisées dans le cadre de la segmentation présentée au chapitre 8.

Chapitre 7

Classification du contenu haut multimedia

Le but de ce septième chapitre est de valider nos hypothèses concernant la fusion des media. Pour cela nous réalisons des expériences de classification du contenu à partir de descripteurs de l'image et du son. Nous présentons les modèles statistiques appris dans le paragraphe 7.1. Puis nous montrons les résultats de classification obtenus par les différents modèles dans le paragraphe 7.2. Nous comparons les hypothèses de modélisation de la fusion de plusieurs media. L'expérience nous conduit en particulier à sélectionner le modèle utilisé par la suite parmi tous les modèles.

7.1 Modèle de classification

Afin de valider ou non les hypothèses concernant la fusion des media, les résultats de la catégorisation automatique du « lieu » d'un plan de vidéo sont analysés. Pour nos expériences, nous avons choisi une classification hiérarchique nécessaire à l'analyse des films, présentée pour l'image au chapitre précédent. Les caractéristiques des tâches d'extraction automatique permettent de traiter l'ensemble des problèmes posés.

7.1.1 Le contenu utilisé

Le modèle de classification présenté est un modèle classique de concepts hauts. La détermination du lieu où a été capturé un plan (image et son) est une tâche essentielle du traitement des films. Dans ce cas, les concepts extraits sont dits « multimedia », ils décrivent des caractéristiques sémantiques qui concernent plusieurs media. Les plans regroupés au sein d'une classe de lieux possèdent des caractéristiques physiques communes, qui sont la signature visuelle et auditive du lieu. Cette classification comprend trois concepts hauts :

- $H_1 =$ *intérieur/extérieur*
- $H_2 =$ *absent/parking/maison/magasin*
- $H_3 =$ *absent/nature/ville*

Bien que cette classification comporte plusieurs concepts, elle est souvent appelée classification *intérieur/extérieur* par simplification. La caractéristique de lieu est essentielle à la segmentation de la structure des films, les scènes étant définies par une unité de lieu. Ces concepts sont une représentation de cette caractéristique. Ils sont estimés à partir de plusieurs descriptions physiques provenant de l'audio et de la vidéo.

contenu image

Le contenu image d'un plan est estimée par l'analyse de l'image « moyenne » de ce plan. L'algorithme de segmentation de plans, développé au CEA [Jos00], réalise la segmentation de niveau bas

du signal visuel des films par écrêtage d'une fonction d'observation fondée sur la couleur. L'image « moyenne » de chacun des plans correspond au minimum local de cette fonction entre deux coupures. Nous supposons que le contenu de cette image est représentatif du contenu global du plan, en raison des hypothèses d'homogénéité définies au paragraphe 3.3.

La présence d'objets ou de textures particulières dans cette image peut donner des indications fortes sur la classification. L'ensemble M_{IMA} des **concepts moyens image** employés pour la classification des lieux comprend 8 concepts. Ils ont été présenté dans le chapitre 6 :

- $M_1 =$ bâtiment $= 1/0$
- $M_2 =$ voiture $= 1/0$
- $M_3 =$ lumière artificielle $= 1/0$
- $M_4 =$ route $= 1/0$
- $M_5 =$ ciel $= 1/0$
- $M_6 =$ herbe $= 1/0$
- $M_7 =$ arbre $= 1/0$
- $M_8 =$ personne $= 1/0$

L'ensemble B_{IMA} des **descripteurs numériques image** est celui employé précédemment pour la localisation d'images. Ils se présentent sous la forme d'un histogramme de texture et de couleur (TLEP).

contenu sonore

En ce qui concerne l'audio, les descripteurs sont déterminés par l'analyse de l'intégralité du segment de son correspondant au plan considéré.

Le **contenu moyen** est constitué des concepts de la classification hiérarchique *parole/musique* présentée précédemment. L'ensemble M_{SON} est composé de 3 concepts moyens :

- $M_9 =$ *parole/non parole*
- $M_{10} =$ *absent/parole seule/mélange*
- $M_{11} =$ *absent/musique/bruit de fond*

L'ensemble des **descripteurs numériques audio** est composé de descripteurs fondés sur les MFCC et issus de la recherche en reconnaissance de la parole. Ils ont été appliqués par [Bim04] dans le cadre de la reconnaissance du locuteur et dans le cadre de la classification d'ambiance par [Nit97] sous une forme légèrement adaptée (descripteurs RASTA). Soit B_{SON} l'ensemble de ces 96 descripteurs bas audio.

7.1.2 Modèle statistique

Les modèles statistiques présentés au chapitre 4 sont appliqués pour la détection des concepts hauts multimedia. Plusieurs modèles de fusion des données sont testés afin de valider nos hypothèses de modélisation.

Modèle naïf de Bayes

Pour un concept haut donné, les relations de dépendance entre les descripteurs image et les descripteurs audio sont dictées par les quatre hypothèses :

1. Les descripteurs images sont indépendants des descripteurs audio sachant le concept haut auquel ils sont rattachés.
2. Les descripteurs bas image associés à un concept ne sont pas indépendants sachant ce concept (idem pour l'audio). Cette hypothèse a été vérifiée dans le chapitre 5 sur 2 concepts moyens.
3. Les descripteurs moyens image associés à un concept haut ne sont pas indépendants sachant ce concept (idem pour l'audio). Cette hypothèse a été vérifiée dans le chapitre précédent.

4. L'ensemble des descripteurs bas image associé à un concept haut sont dépendants des concepts moyens images associés au même concept. Cette hypothèse a été vérifiée dans le chapitre précédent dans le cas monomédia.

Afin de valider ou non l'hypothèse de modélisation 1, ce modèle sera comparé aux modèles de fusion décrits dans le chapitre 4. Les hypothèses d'indépendance des concepts moyens audio et image et d'indépendance des descripteurs bas (audio et image) sont testées sur trois modèles de fusion des scores et des descripteurs bas. Ceux-ci sont considèrent la fusion du contenu visuel et auditif.

Fusion du contenu multimedia

Afin de vérifier ou non les relations de dépendance entre media, nous testerons plusieurs modèles qui considèrent plus ou moins de dépendance entre descripteurs.

Le premier modèle réalise la « late » fusion de l'ensemble des descripteurs bas et de l'ensemble des descripteurs moyens, sans tenir compte du media dont ils sont issus. Ce modèle remet en cause l'hypothèse 4 dans le cadre de la classification multimédia. Les relations de dépendance du modèle naïf de Bayes peuvent être représentées par le RB de la figure 7.1.

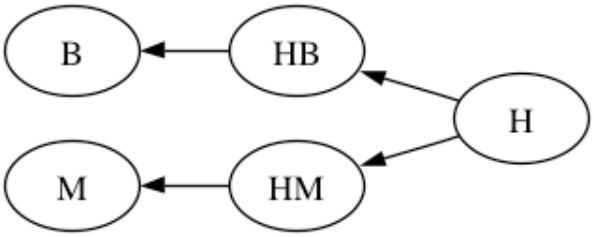

FIG. 7.1 – Modèle de late fusion des descripteurs bas et moyen.

Tout d'abord, nous envisageons la fusion du contenu bas des différents media. Nous avons souligné la nécessité de modifier ces descriptions afin d'améliorer les performances de classification. Nous appliquons le traitement des données présenté en figure 7.2 en haut.

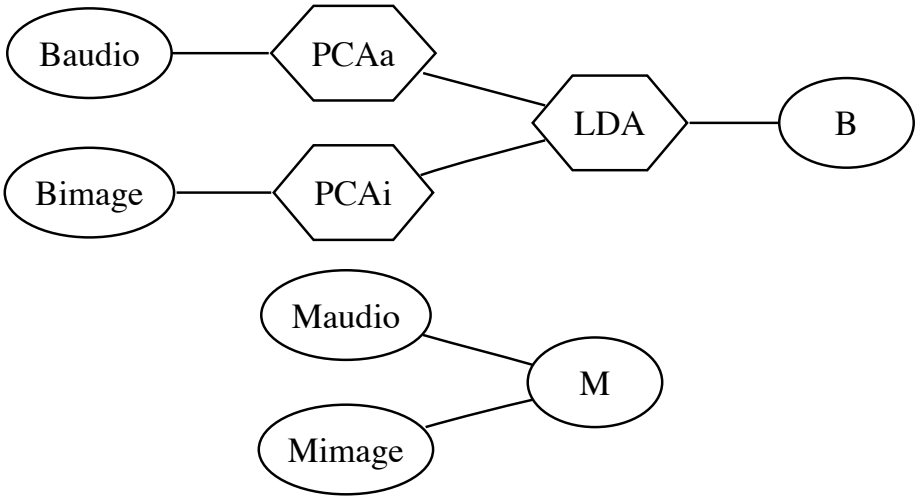

FIG. 7.2 – Modèle de fusion des descripteurs bas et des descripteurs moyens

La fusion des descripteurs bas audio et vidéo se fait en deux étapes. En premier lieu les descripteurs

de chaque media sont décorrélés séparément par PCA. Ensuite les vecteurs descripteurs, produits de la concaténation des deux descripteurs, sont adaptés aux classes de la taxonomie par un algorithme de LDA.

Enfin, pour prendre en compte les relations complexes qui existent entre les concepts moyens provenant de plusieurs media, les descripteurs scores des deux media sont fusionnés (concaténation) pour la classification SVM du concept H_M par les scores audio et vidéo (figure 7.2 en bas).

Le second modèle, présenté au chapitre 4.3.3, réalise la « late » fusion de l'ensemble des descripteurs audio et de l'ensemble des descripteurs image, il suppose que l'hypothèse 1 est vraie, c'est à dire que les deux médias sont indépendants. Cette fusion envisage de modéliser un concept pour chaque ensemble de media de façon indépendante, et de fusionner la décision de ces classifications. Pour un concept haut H, deux nouveaux concepts hauts intermédiaires sont définis H_{SON} et H_{IMA} correspondant aux modèles de classification par les descripteurs audio et vidéo. Le concept H est alors déterminé par une classification SVM des scores de ces concepts.

Le troisième modèle est le plus simple : il réalise la fusion par la classification du vecteur score global du plan. Celui-ci comprend les scores obtenus par les concepts moyens et les descripteurs bas extraits du signal image et son du plan considéré. Ce modèle suppose que l'hypothèse 1 est fausse.

Les résultats de classification des différents modèles présentés ici nous permettront de tirer des conclusions sur les relations d'interdépendance qui existent entre le contenu visuel et auditif des films.

7.2 Expérience et validation des hypothèses de modélisation

Dans la troisième expérience nous étudions les résultats de classification de plusieurs concepts multimedia. Cette classification a pour but d'extraire la valeur estimée de concepts hauts de la classification pour un plan de film.

7.2.1 Base de données

Cette expérience nécessite donc la création d'une base de données de plans indexés par le contenu choisi : annotation des concepts audio et vidéo choisis $(H_1, \ldots, H_p, M_1, \ldots, M_n)$, calcul des descripteurs bas.

Cette base est séparée en deux : une base d'apprentissage des modèles hauts et une base de test. La première partie des objets sert à l'apprentissage des modèles des concepts hauts et moyens, et du traitement des descripteurs bas. La deuxième partie est la base de test des classifications.

La base de données est constituée de 400 plans extraits de vidéos personnelles. Ces plans sont annotés manuellement des concepts de la classification. Nous avons porté une attention particulière pour que chaque classe de la taxonomie de classification ait une quantité représentative d'images à apprendre, et que le nombre de plans ne varie pas trop au sein des concepts.

Pour cette expérience, la première base comporte 300 plans et la base de test 100.

7.2.2 Critères d'évaluation

Afin de comparer les différents modèles, nous analyserons leurs résultats de classification. La matrice de confiance est définie : $\mathbf{C} = \mathbf{C}_{ij}$ ou $\mathbf{C}_{ij}$ est le pourcentage d'objets annotés de la $i^{ème}$ classe et classé par le modèle dans la $j^{ème}$. Ainsi les éléments $\mathbf{C}_{ii}$ de la diagonale correspondent au taux de bonnes classifications pour la classe i. Le taux de classification globale $\mathbf{GC}$ est définit comme la moyenne des $\mathbf{C}_{ii}$.

Intérieur/Extérieur	
89	11
10	91

Parking/Maison/Magasin		
77	3.5	19.5
1	78.5	20.5
1	14	85

Nature/Ville	
86.5	13.5
7.5	92.5

TAB. 7.1 – Performances de classification des lieux par les descripteurs de l'image

Intérieur/Extérieur	
88	12
9	91

Parking/Maison/Magasin		
87	4	9
1	92	7
0.5	5.5	94

Nature/Ville	
89,5	10.5
11.5	88.5

TAB. 7.2 – Performances de classification des lieux par les descripteurs du son

7.2.3 Résultats

Pour comparaison, les performances obtenues au chapitre précédent pour la classification SVM (early fusion) image des concepts hauts sont montrées dans le tableau 7.1. Ensuite, dans le tableau 7.2, nous présentons les performances de classification des mêmes concepts par les descripteurs de l'audio. Enfin, le tableau 7.3 expose les performances de classification de la fusion des descripteurs de l'image et du son par un modèle naïf de Bayes.

Intérieur/Extérieur	
93	7
8.5	91.5

Parking/Maison/Magasin		
86	4	10
1.5	93	5.5
1	11	88

Nature/Ville	
89	11
6.5	93.5

TAB. 7.3 – Performances de classification des lieux par fusion naïve des descripteurs image et son

Ces tableaux montrent que les performances varient selon les descripteurs employés pour la classification. Nous pouvons en déduire plusieurs observations.

Tout d'abord, la comparaison des résultats de classification monomedia nous montre que, suivant le concept haut étudié, l'information différente apportée par chaque media entraîne des performances variables. En ce qui concerne les concepts H_1 et H_3, le traitement de l'image est plus efficace que celui du son. Et pour le concept H_2, nous observons le phénomène inverse.

Remarquons que la fusion « naïve » des descripteurs audio et image améliore les performances de classification des concepts de lieux. Nous constatons un écart de 3 % entre la classification par l'image et la classification multimedia. Ainsi, l'information apportée par le contenu sonore donne un pouvoir de discrimination supplémentaire au modèle de classification.

Nous expérimentons enfin les trois modèles de fusion présentés au chapitre 4. Les taux de bonne classification des concept hauts considérés sont présentées dans le tableau 7.4.

Ils dévoilent que les performances varient peu selon le modèle choisi. Cependant, quatre remarques peuvent être faites.

Premièrement, la collaboration des classifications audio et image par late fusion fournit de meilleures performances comparée à la classification image uniquement (5% environ). Pour le signal sonore la fusion des descripteurs moyens et bas par classification SVM du vecteur descripteur global permet d'obtenir de bons taux de caractérisation, malgré quelques concepts moyens non pertinents. De plus, nous remarquons que la classification de l'audio fournit souvent de meilleurs résultats que celle de l'image. Ainsi, la fusion du son et de l'image améliore, définitivement, les performances de classification des plans de vidéo en lieux.

Modèle	H_1	H_2	H_3
Bas :A+V	88.5	81	92.5
Moyen :A+V	89.5	40	75
Late fusion :B+M	90	70	91
Audio	89,5	**91**	89
Image	89,5	80	89,5
Late fusion : A+I	**94,5**	90	**94**
Early fusion	91.5	88	93

TAB. 7.4 – Performances de classification des lieux : comparaison des modèles SVM

Deuxièmement, nous constatons une légère amélioration des performances des trois modèles de fusion par rapport au modèle naïf de Bayes, l'écart est de 2.5% environ. La prise en compte de la dépendance entre les descripteurs présente donc un interêt net dans le cadre de la classification multimedia.

Troisièmement, le modèle de "late" fusion des descripteurs image et son apporte de meilleurs résultats que celui réalisant la late fusion des descripteurs bas et moyens. Cela est dû principalement au fait que les concepts moyens audio ne semblent pas discriminants pour la classification des lieux ; de plus, les concepts moyens image ne sont pas tous pertinents pour les concepts hauts. Ainsi dans la "late" fusion des descripteurs bas et moyens, la classification par les concepts moyens multimedia n'a que peu de poids et elle n'apporte pas d'amélioration à la classification des descripteurs bas multimedia. Alors que pour la "late" fusion audio et image, les concepts moyens sont intégrés au sein d'ensembles comprenant des descripteurs bas. L'influence des concepts sur la classification est mieux prise en compte, notamment en terme de corrélation avec les descripteurs bas (voir le chapitre précédent).

Quatrièmement, contrairement aux expériences du chapitre précédent, la classification SVM du vecteur descripteur global, produit de la concaténation des scores et des descripteurs bas, obtient de moins bonnes performances que la « late » fusion du son et de l'image (1.5% en moyenne). Ceci est troublant car, a priori, la fusion early déconsidère les descripteurs non discriminants et dépasse la « late » fusion en terme de performances. Cependant, nous donnons une explication à cet écart : ce modèle prend bien en compte les phénomènes de masquage où l'image (ou le son) ne permet pas d'identifier le lieu de l'action. Par exemple : l'image est un gros plan sur une cravate, où le son est saturé par une personne qui parle. En général, ces phénomènes concernent un seul des deux media. La « late » fusion traite séparément les deux ensembles de descripteurs, et est ainsi plus efficace dans ces cas de figure. Remarquons qu'au cinéma ce genre de masquage est fréquent. Il n'est pas rare, que la bande sonore ne contienne aucune indication sur la localisation du plan (e.g. musique seule) ; de même pour l'image (e.g. gros plan). C'est pourquoi ce type de fusion est bien adapté aux films. Il serait intéressant de tester ces modèles pour d'autres types de documents audiovisuels : documentaires, journaux télévisés, rencontres sportives, pour lesquels le montage audiovisuel est plus "réaliste" dans sa représentation du monde : les émetteurs sont souvent présents dans les deux médias et le son et l'image sont synchrones. Il est possible que dans ces cas, la "late" fusion soit moins efficace, en raison de la perte de corrélation entre les deux médias.

7.2.4 Résumé des résultats

En résumé, les résultats de ce chapitre assurent que les modèles de fusion des descripteurs image et audio capturent bien les caractéristiques utiles pour la classification des concepts hauts. Les performances obtenues pour la classification de plans de vidéo en lieux sont améliorées par la fusion.

Les exemples de ce chapitre nous permettent de conclure sur les hypothèses prises précédemment.

Les **concepts moyens audio** choisis ne sont pas discriminants pour la classification des concepts hauts H_1 : *intérieur/extérieur*, H_2 : *parking/maison/magasin* et H_3 : *ville/nature*.

La **fusion des descripteurs image et son** pour la classification, apporte une amélioration des performances de classification des concepts de « lieux ». La complémentarité de l'information apportée par les deux media permet de valider les classifications monomedia lorsqu'elles sont identiques et de résoudre les ambiguïtés lorsque que l'un des media ne contient pas assez d'information sur la catégorisation choisie.

Les meilleures performances de classification sont obtenues par le modèle de **« late » fusion** de l'ensemble des descripteurs audio et image. En effet il semble mieux considérer les phénomènes de masquage d'un des media et ainsi réduit les erreurs, quand une perturbation environnementale ou le montage affectent un seul media. Pour la suite, nous appliquerons cette technique pour la catégorisation, par l'image et le son, du lieu où a été tourné un plan.

Ces résultats seront considérés comme admis pour les expériences présentées dans le chapitre suivant. Les exemples de ce chapitre servent de repères expérimentaux des performances des classifications utilisées dans le cadre de la segmentation présentée au chapitre 8.

Chapitre 8

Segmentation de la structure des films

L'extraction de la structure est essentielle pour le traitement automatique des films : détermination du contenu, recherche et résumé d'extraits.... Il est maintenant largement accepté que les films sont hiérarchiquement structurés en scènes, plans et images. Afin de mieux décrire la structure complexe des films de cinéma nous introduisons deux autres niveaux : groupe de plans et groupe de scènes. Plusieurs algorithmes de segmentation des plans obtiennent de bons résultats. Cependant il est toujours difficile d'extraire d'autres types de structures, notamment les scènes. De plus la plupart des algorithmes de segmentation de films ne prennent pas en compte le son.

Ce huitième chapitre présente une application des modèles de contenu à la segmentation hiérarchique de la structure des films de cinéma. Nous expliquons dans le paragraphe 8.1 comment segmenter les films par l'analyse des descripteurs de l'image et du son. Nous donnons ensuite les performances de segmentation de l'algorithme pour quelques exemples de films dans le paragraphe 8.2.

Ce chapitre apporte peu de nouveautés théoriques par rapport à l'état de l'art. Mais l'expérience de segmentation de films par les annotations textuelles n'a pas d'équivalent dans la littérature, à notre connaissance. Faute de disposer d'une base de données suffisante, nous présentons les résultats de segmentation de huit films. La performance de notre algorithme ne peut donc être comparée à celle des algorithmes existants. C'est pourquoi, nous introduisons un algorithme de référence inspiré des techniques classiques de segmentation.

8.1 Segmentation temporelle

Nous proposons l'étude d'un algorithme de segmentation des films qui permet de déterminer la structure complexe définie au chapitre 3. Le but est de segmenter le film selon la hiérarchie : groupe de scènes, scènes, groupe de plans (de clips), plans (clips). Les éléments de structure ainsi extraits sont multimedia, sauf aux deux premiers niveaux où la structure est dédoublée en clips pour le son et plans pour l'image.

8.1.1 Segmentation du premier niveau

Segmentation des plans

Les plans sont définis comme des séquences continues d'images prises sans arrêter la caméra. La segmentation en plans consiste à détecter tous les points de transition de l'image. De nombreux algorithmes ont été développés dans la littérature et ils obtiennent aujourd'hui de bonnes performances. Nous utilisons, pour cette tâche, l'algorithme développé au CEA par Pierre Josserand [Jos00].

Il s'agit d'un algorithme de segmentation de niveau bas par écrêtage d'une fonction d'observation. Celle-ci est déterminée par la distance entre descripteurs de deux segments consécutifs de 5 images

chacun. Ces descripteurs sont fondés sur la transformée en ondelettes de la couleur et de l'intensité lumineuse. Les points de coupure du signal sont placés sur les crêtes de cette fonction.

L'image « moyenne » de chacun des plans correspond au minimum local de la fonction entre deux coupures. Cette image est une sorte de résumé du plan qu'elle caractérise. Son contenu est, on le suppose, représentatif du contenu de l'ensemble du plan en ce qui concerne les descripteurs choisis pour segmenter.

L'algorithme du CEA obtient de bonnes performances pour une application aux programmes de télévision : documentaires, journaux télévisés... Aucun test n'a été réalisé sur des films de cinéma. Le problème est ardu, pour certaines séquences (travellings par exemple), les caractéristiques peuvent varier fortement durant un plan, ce qui amène de la sur-segmentation. De plus la présence de transitions complexes et parfois de longue durée entraîne des risques de non détection.

Segmentation des clips

Les clips sont des segments de son homogènes selon les concepts de la classification hiérarchique *parole/musique*. Cette structure permet de prendre en compte le montage de la bande son par addition d'objets sonores (parole, musique, bruit de fond etc...). Nous appliquons pour cette tâche un algorithme de segmentation par classification de segments de taille fixée à 3s. Cette classification est réalisée par un algorithme MMC dynamique appris selon le modèle de classification présenté au paragraphe 5.1.1.

Ces traitements fournissent la structure de niveau bas du film en entrée du système. La prochaine étape consiste dans l'extraction du contenu des objets ainsi délimités (plans et clips). Ce contenu servira à segmenter le film aux niveaux supérieurs de la structure. Dans le cadre de notre système, il est déterminé à partir de l'image moyenne des plans et de l'intégralité des clips. Il est nécessaire d'extraire de nouveaux descripteurs d'une part parce que ceux issus de la première segmentation ne sont pas adaptés à l'information mise en valeur par ces niveaux, et d'autre part il serait trop lourd de les manipuler, sachant qu'un film de 1h30 comprend $1.35 * 10^5$ images et 1800 frames.

8.1.2 Contenu du premier niveau

Le contenu descriptif bas, moyen et haut utilisé pour caractériser l'information des plans et des clips est noté D. Les descripteurs sont caractérisés par le media dont ils sont issus, leur niveau sémantique, et la hiérarchie de classification dont ils font partie. Ils ont tous été présentés précédemment dans les chapitres 5, 6 et 7, pour rappel :

Descripteurs Haut

En ce qui concerne le signal **image**, les concepts hauts développés concernent la classification hiérarchique du lieu. L'ensemble des concepts extraits est $H_{\text{Ima}} = \{H_1, H_2, H_3\}$ (voir 7) :

- $H_1 =$ *intérieur/extérieur*
- $H_2 =$ *parking/maison/commerce/bureau*
- $H_3 =$ *nature/ville*

Dans les faits, ces concepts sont multimedia, puisqu'ils sont extraits par classification du son et de l'image d'un plan. Cependant, nous le définissons comme un descripteur image pour des raisons de concordance avec le modèle de structure (plan, clip) choisi. Ainsi, il est considéré, dans ce cas, que l'audio aide à la classification de l'image. L'ensemble des descripteurs hauts image est noté $H_{\text{IMA}} = \{H_1, H_2, H_3\}$.

En ce qui concerne le signal **audio**, le concept H_4 : nom du locuteur a été introduit pour les segments de son comportant de la parole au chapitre 5. Les modèles de voix de sept acteurs et un modèle général de voix ont été appris. La classification du son permet de déterminer la valeur du concept haut dans la taxonomie $\{c_0, \ldots, c_7\}$, avec :

- c_0 : "autre"
- c_1 : Angelina Jolie (Tomb Raider)
- c_2 : Catherine Deneuve (Belle Maman)
- c_3 : Vincent Lindon (Belle Maman)
- c_4 : Seann William Scott (American Pie)
- c_5 : Jason Biggs (American Pie)
- c_6 : Jacques Gamblin (Les enfants du marais)
- c_7 : Jacques Villeret (Les enfants du marais)

L'ensemble des descripteurs hauts audio est noté $H_{\text{SON}} = \{H_4\}$.

Descripteurs moyens

Pour l'**image**, les concepts moyens (voir le chapitre 6) concernent la classification de textures et d'objets de façon manuelle ou automatique. L'ensemble des concepts moyens extraits est noté M_{IMA} et comporte 8 concepts :

- $M_1 =$ bâtiment $= 1/0$
- $M_2 =$ voiture $= 1/0$
- $M_3 =$ immeuble $= 1/0$
- $M_4 =$ route $= 1/0$
- $M_5 =$ ciel $= 1/0$
- $M_6 =$ herbe $= 1/0$
- $M_7 =$ arbre $= 1/0$
- $M_8 =$ visages $= n$

Le huitième concept, $visages$, identifie le nombre de personnages à l'écran.

Pour le **son**, les concepts développés sont associés à la classification hiérarchique *parole/musique* (voir le chapitre 5). L'ensemble M_{SON} des concepts audio est composé de trois concepts :

- $M_9 =$ *parole/non parole*
- $M_{10} =$ *parole seule/mélange*
- $M_{11} =$ *musique/bruit de fond*

Descripteurs bas

Les descripteurs de niveaux sémantique bas sont les descripteurs associés aux différents modèles de classification des concepts.

Pour l'**image** il s'agit de l'histogramme TLEP, décrit au chapitre 6. L'ensemble est noté B_{IMA}.

Pour le **son**, cinq descripteurs bas ont été sélectionnés dans le chapitre 5 pour la classification hiérarchique *parole/musique* . Nous avons aussi décrit 48 descripteurs fondés sur les MFCC employés à la détermination du locuteur (chapitre 5) et à la classification du lieu (chapitre 7). L'ensemble est noté B_{SON}.

Le premier niveau structurel est monomedia. Le signal image d'un film est décrit par une séquence de vecteurs que l'on notera $Plan = \{plan_i\}_{i=1..T_{\text{IMA}}}$, où T_{IMA} est le nombre de plans du film. Le $i^{ème}$ vecteur descripteur $plan_i$ contient les valeurs des descripteurs hauts, moyens et bas pour le plan i. Les dimensions de ce vecteur sont normalisées : moyenne nulle, variance fixée à 1. De même, le signal audio est représenté par la séquence des clips du films : $Clip = \{clip_i\}_{i=1..T_{\text{Son}}}$.

Les structures des niveaux supérieurs sont extraites par segmentation hiérarchique de ces séquences de plans et de clips.

8.1.3 Segmentation des niveaux supérieurs

Algorithme de segmentation hiérarchique

Le même algorithme est employé pour segmenter chaque niveau de la structure, en partant du niveau le plus haut : groupe de scènes, jusqu'au deuxième niveau : groupe de plans et groupe de clips . Comparée à une technique de segmentation du bas vers le haut, celle-ci permet de segmenter indépendamment chaque élément d'un niveau donné. Par exemple au niveau « groupe de plans », un modèle de segmentation appris sur l'intégralité du film est moins discriminant qu'un modèle appris sur une scène en particulier. Cela est dû au fait que l'apprentissage du modèle de classification des segments est réalisé sur une quantité limitée de données. Ainsi, les modèles de classes correspondent mieux à la variation de l'information au sein d'une scène.

L'algorithme de segmentation est fondé sur la classification MMC non-supervisée de la séquence de plans et de clips de façon indépendante. Cette classification permet de déterminer l'appartenance d'un objet de la séquence à l'une des classes du concept de classification C appris sur la séquence. La classification MMC, comparée à un regroupement simple, permet de considérer les probabilités de transitions entre groupes de plans, scènes, (etc. . .).

Cependant, les MMC sont des algorithmes itératifs qui nécessitent une initialisation. Un mauvais choix de paramètres initiaux peut fausser la classification. C'est pourquoi, nous appliquons le modèle présenté par [Foo03]. Ce modèle réalise une pré-segmentation de la séquence de plans par un regroupement (« k-moyenne »), lui-même initialisé par une première segmentation de bas niveau. Les classes (ou états) du modèle initial de segmentation sont apprises sur la séquence par l'algorithme de Baum-Welsh afin que chaque classe de la taxonomie de C représente une information trouvée dans différentes parties de la séquence. Cet apprentissage est réalisé sur les vecteurs descripteurs choisis pour la segmentation. L'ensemble de descripteurs doit donc décrire au mieux l'information commune aux éléments de chaque classe de la taxonomie. Ainsi, le choix des descripteurs utilisés est essentiel. Les probabilités de transition, sont apprises sur l'ensemble des huit films, pour chaque niveau structurel. Ceci afin de fixer la durée moyenne des éléments de la structure sur une valeur observée par l'annotation manuelle. En effet, il semble que l'apprentissage non-supervisé de ces probabilités produisent trop d'erreur de sur-segmentation, notamment dans les cas de structures alternées comme les scènes de dialogues (alternance champs/contre-champs).

Jusqu'ici, nous n'avons pas pris en compte l'aspect multimedia de la structure. Pour les niveaux supérieurs à deux, les coupures du son et des images coïncident : les éléments détectés sont multimedia. Il est donc courant de fusionner l'information provenant de l'image et du son afin d'améliorer les performances du système. Tout d'abord, la séquence des plans est segmentée pour obtenir les points de coupure image des structures du niveau considéré. Ensuite, les points de coupure audio sont déterminés. Enfin les points de coupure audio et image sont réunis par l'algorithme du plus proche voisin contraint. Pour une limite image, la limite audio la plus proche est sélectionnée ; si leur distance dépasse un seuil fixé, ils sont éliminés ; dans le cas contraire ils sont sélectionnés et un point de coupure multimédia est noté. Le choix de ce seuil dépend du niveau considéré de la structure et est déterminé d'expérience. La fusion des descripteurs audio et image par concaténation est impensable puisque au niveau bas de la structure les limites de plans et de clips ne correspondent pas. Les séquences de contenu associés aux plans et aux clips sont donc asynchrones.

Dans le cadre de nos tests, nous comparerons les résultats de plusieurs segmentations afin de déterminer le vecteur descripteur global utilisé pour représenter un plan et un clip à chaque niveau et media.

Algorithmes de comparaison

Les performances de cet algorithme de segmentation du bas vers le haut seront comparées à un modèle un peu plus complexe par son déroulement. Il correspond aux traitements classiques de la littérature : la segmentation des niveaux sémantiques est réalisée en 4 étapes :

- Segmentation de niveau bas des plans et des clips.
- Extraction des scènes par fusion des limites de plans et de clips.
- Segmentation MMC des scènes en groupes de plans.
- Regroupement MMC des scènes en groupe de scènes. Les scènes sont représentées par un vecteur descripteur composé de la moyenne et de la variance des descripteurs des plans de la scène considérée.

8.2 Expériences

Cette expérience nécessite la création d'une base de données de films indexée manuellement selon le modèle de structure et de contenu précédemment défini. Cette base est utilisée comme base de test. Les performances du système sont évaluées par comparaison des coupures de la segmentation automatique et de l'index d'apprentissage.

8.2.1 Base de données

La base de données est constituée de 8 films de cinéma : Belle maman, Les enfants du marais, 8miles, American Pie, Tomb raider, Yamakasi, Mortal Kombat, Dragon rouge. Au delà de tout jugement sur la qualité esthétique de ces œuvres, elles sont assez représentatives des films actuels. En effet, leurs caractéristiques sémantiques :

- style (action, comédie, film d'auteur) ;
- structure : montage (rapide/lent), narration ;
- personnages : hommes, femmes, enfants ;
- lieux : intérieur, extérieur. . . ;
- présence de musique (ou non) ;

et physiques : colorimétrie, textures, sont assez variées et couvrent un large panel.

Ces films sont stockées au format AVI. Le signal audio et image est ré-échantilloné, pour les besoins du traitement. Le son est échantillonné à 22.5 kHz, quantifié sur 16 bit et mixé en mono. Le signal image est une succession d'images fixes à la cadence de 25 images par secondes.

L'ensemble des films de la base est indexé de façon semi-manuelle : nous distinguons la partie structure et contenu.

En ce qui concerne la structure : les plans sont extraits par l'algorithme de segmentation développé au CEA par [Jos00]. Les clips sont segmentés par classification MMC du signal sonore (segment fixe de 3s). La structure de niveau supérieur (groupe de plans, scènes, groupe de scènes) est indexée manuellement.

En ce qui concerne le contenu : pour le premier niveau de la structure (plans, clips), les descripteurs bas sont extraits automatiquement et les concepts moyens et hauts sont annotés manuellement.

8.2.2 Critères d'évaluation

Les performances des modèles de segmentation des données sont analysées par l'évaluation de leurs résultats. Les performancees peuvent être décrites par deux valeurs déterminées par la comparaison d'une segmentation manuelle supposée juste et la segmentation réalisée par l'algorithme :

- Le taux de précision est le pourcentage exprimant le rapport entre le nombre de transitions pertinentes retrouvées par la segmentation et le nombre total de transitions trouvées par la segmentation.

Il vaut 100 lorsque toutes les transitions détectées correspondent à des transitions de référence et une valeur faible indique des problèmes de sur-segmentation. Ce taux est noté,

$$TP = \frac{\#\text{Transitions correctement détectées}}{\#\text{Transitions détectées}}$$

– Le taux de rappel est le pourcentage exprimant le rapport entre le nombre de transitions pertinentes retrouvées par la segmentation et le nombre total de transitions pertinentes que contient un film. Il vaut 100 lorsque toutes les transitions de références ont été détectées et une valeur faible indique une sous-segmentation. Ce taux est noté,

$$TR = \frac{\#\text{Transitions correctement détectées}}{\#\text{Transitions de références}}$$

8.2.3 Résultats

Les résultats obtenus pour la segmentation sont présentés dans ce paragraphe. Les performances de l'algorithme choisi sont analysées pour chaque niveau de la structure. Plusieurs vecteurs descripteurs sont comparés, ce qui nous permettra de conclure sur l'utilité de la description de la séquence des plans (et des clips) par des concepts moyens et hauts. Afin de valider l'hypothèse selon laquelle la coopération des media améliore les performances, les résultats de la segmentation du son, d'une part, de l'image, d'autre part, puis de la fusion des deux sont comparés.

Segmentation des plans

Cette première étape de la segmentation du film est réalisée par l'algorithme développé au CEA. Les performances obtenues par cette technique appliquée aux 8 films de la base sont présentées dans le tableau 8.1. Les résultats obtenus par l'auteur sur un ensemble de vidéos hétéroclites (journaux TV, films, animations…) sont aussi montrés.

	TP	TR
Base hétéroclite	90	99
Base de films	86	98

TAB. 8.1 – Performances de segmentation des plans

Nous remarquons que les performances sont sensiblement identiques pour les deux bases de test. Nos résultats sont légèrement inférieurs à ceux obtenus par les premiers tests de cet algorithme.

Les paramètres de cet algorithme, notamment le seuil de détection des points de coupure, sont fixés afin de détecter tous les changements de plans. Ce qui a pour conséquence de minimiser le nombre de sous-segmentations, alors que le nombre de **sur-segmentation** est assez important. Nous pouvons distinguer deux cas particuliers de sur-segmentation : les longs plans séquences et le passage d'un objet en gros plan dans l'image. La figure 8.1 montre un exemple de sur-segmentation pour chacun de ces cas. Pour un plan du film, les images «moyennes» obtenues par la segmentation appartiennent au même plan (présentation de la production, plan séquence «suivi du vélo»).

Segmentation des clips

La segmentation de la bande son en clips est réalisée par la classification MMC du signal audio. Dans le cadre de cette expérience, nous gardons les paramètres du modèle appris au chapitre 5. Il aurait été intéressant de réaliser l'apprentissage sur des sons provenant uniquement de films. Dans un film, le signal audio est issu d'un montage par addition d'amplitudes, ce qui le rend différent des sons de la base choisie

FIG. 8.1 – Cas de sur-segmentation des plans.

précédemment. En effet les objets de la base sont, pour la plupart, des sons « naturels » enregistrés par un seul capteur. Cependant, les performances de l'algorithme présentées dans le tableau 8.2, montrent des résultats équivalents à ceux obtenus précédemment.

	TP	TR
Sons de films	94	96

TAB. 8.2 – Performances de la segmentation des clips

Segmentation des groupes de scènes

La segmentation du signal **image** d'un film en groupe de scènes est réalisée par la classification MMC non-supervisée de la séquence des plans. Afin de mettre en valeur le rôle joué par les différents types de descripteurs (numérique et concepts), la segmentation est effectuée sur les ensembles préalablement cités : B_{Ima}, $M_{\text{Ima}} \cup H_{\text{Ima}}$ et $B_{\text{Ima}} \cup M_{\text{Ima}} \cup H_{\text{Ima}}$.

La figure 8.2 montre la matrice de similarité de la séquence des plans correspondant à l'intégralité du film « American Pie » extraite à partir des descripteurs bas images, et des concepts moyens et hauts image. A l'œil, il est possible d'identifier certaines structures du film qui se détachent de la représentation. Par exemple, le grand carré rouge en haut à gauche de l'image correspond à la séquence de plans délimitée par les plans 161 et 523. Cependant, si ces plans montrent une similarité pour les caractéristiques choisies pour segmenter, il n'est pas évident qu'il corresponde effectivement à un groupe de scènes. De plus, nous remarquons que les limites de cette séquence ne sont pas nettes, et sont difficiles à interpréter.

La figure 8.3 représente la segmentation obtenue par la classification non-supervisée « k-moyenne » de la séquence des plans extraites à partir des descripteurs moyens et hauts image des 30 premières minutes du film American Pie (500 plans environ). Nous ne montrons pas cette figure sur l'intégralité du film pour des raisons de lisibilité, nous ferons de même par la suite. Cette première segmentation sert à l'apprentissage du modèle de classification MMC non-supervisée. Les segmentations correspondant à plusieurs valeurs des seuils (w_1, w_2) sont montrées. Ces seuils sont définis dans [Foo03] ; ils correspondent à la distance minimum entre deux segments consécutifs séparés par la première segmentation bas niveau (w_1) et le regroupement "k-moyenne" (w_2).

Nous constatons que plus la valeur du seuil est basse, plus le nombre de segments extraits est important. Par expérience, réalisée sur plusieurs films, nous fixerons ces seuils à ($w_1 = 0.6$ et $w_2 = 0.9$). Ce choix est un compromis trouvé afin de minimiser le nombre de sur-segmentations et sous-segmentations.

La figure 8.4 montre la classification « k-moyenne » pour les matrices de similarité des concepts calculées sur une fenêtre de taille $w = 5$ plans et $w = 10$ plans.

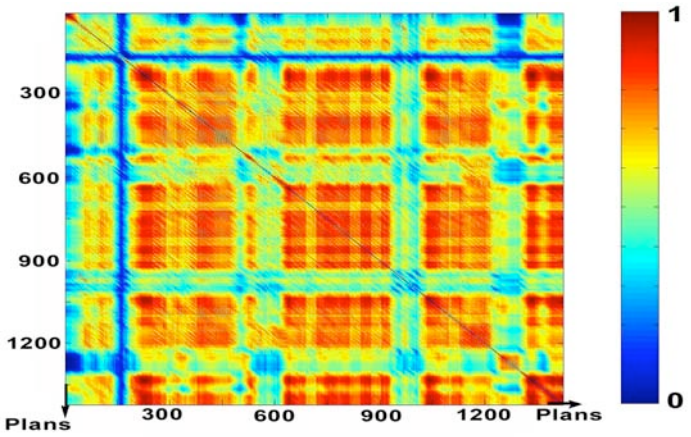

FIG. 8.2 – Matrice de similarité des plans d'*American Pie*.

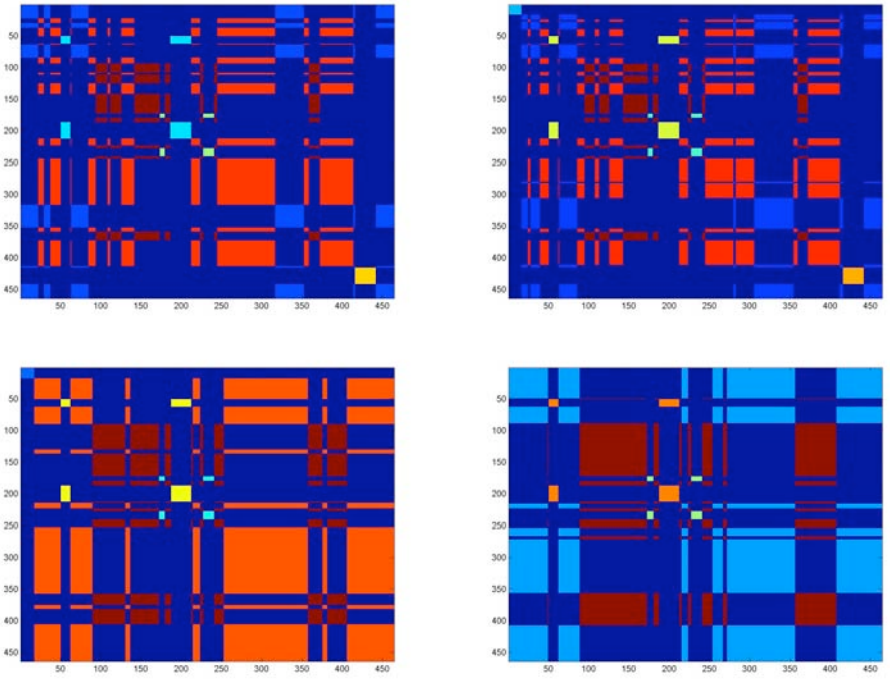

FIG. 8.3 – Segmentation par classification "k-moyenne", comparaison des valeurs de seuil : $(w_1 = 0.55, w_2 = 0.9), (w_1 = 0.6, w_2 = 0.8), (w_1 = 0.6, w_2 = 0.9), (w_1 = 0.6, w_2 = 0.95)$.

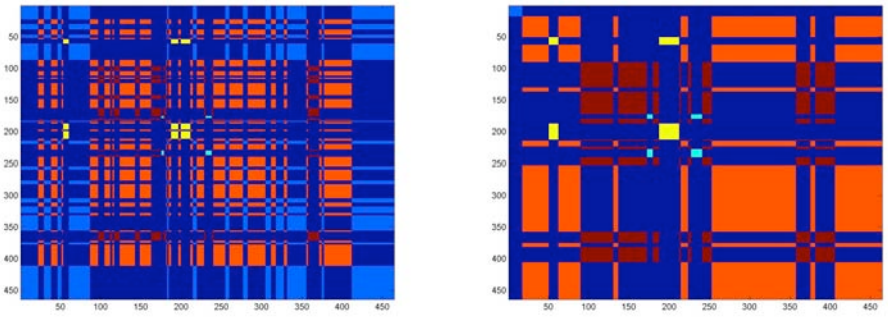

FIG. 8.4 – Segmentation par classification "k-moyenne" des concepts, comparaison des tailles de fenêtre : $w = 5$ et $w = 10$.

Nous remarquons que pour une taille de fenêtre de 5 plans, nous observons une sur-segmentation du signal. De même, cela n'est pas représenté ici, mais pour une taille de fenêtre supérieure de 15 plans, nous avons observé une forte sous-segmentation. Dans les faits, il est rare qu'un groupe de scènes dure moins de dix plans, il n'est donc pas nécessaire de choisir une taille de fenêtre inférieure à 10. De plus, si l'on choisit une taille trop grande, il y a des risques d'inclure dans un groupe de scènes, une scène courte qui ne lui appartiendrait pas. Dans le cadre de la segmentation en groupes de scènes, nous choisissons donc de fixer la taille de la fenêtre de lissage à 10 plans pour le calcul de la similarité.

Le tableau 8.3 présente les performances de segmentation en groupes de scènes par traitement des concepts $M_{\text{Ima}} \cup H_{\text{Ima}}$, d'une part, et des descripteurs bas B_{Ima}, d'autre part. Ces résultats sont déterminés par comparaison de la segmentation manuelle réalisée par l'auteur et de la segmentation automatique pour les huit films.

	TP	TR
Concepts	33	57
Descripteurs bas	17	28

TAB. 8.3 – Performances de la segmentation en groupe de scènes par classification "k-moyenne"

Nous remarquons que la segmentation par classification «k-moyenne» entraîne de nombreuses erreurs de sur-segmentation. Notamment, nous constatons que certains segments trop courts pour être des groupes de scènes sont détectés, ce qui explique des résultats de segmentation assez faibles.

La figure 8.5, montre le résultat final de la segmentation MMC de la séquence des plans des 30 premières minutes d' «American Pie», pour une description par les concepts hauts et moyens.

Nous remarquons que, par rapport à la classification "k-moyenne", plusieurs des segments « courts » disparaissent. De plus, le nombre de classes effectives, c'est-à-dire celles auxquelles sont attribués des plans, diminue. Le MMC réalise donc un « lissage » de la segmentation obtenue par l'algorithme «k-moyenne».

	TP	TR
Concepts	57	57
Descripteurs bas	43	43

TAB. 8.4 – Performances de la segmentation MMC en groupes de scènes

Les performances de segmentation MMC par les concepts présentés dans le tableau 8.4 nous permettent de constater une diminution du taux de sur-segmentation par rapport à une classification simple

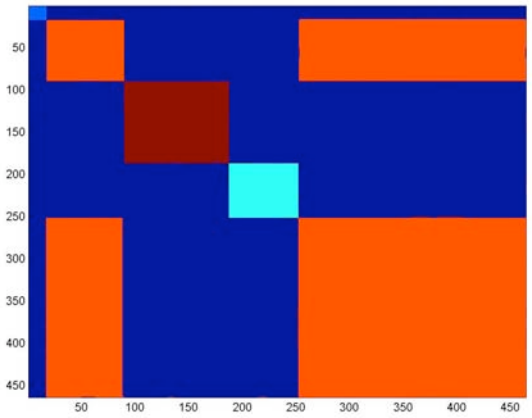

FIG. 8.5 – Segmentation par classification MMC.

par «k-moyenne». Cependant cette opération n'augmente pas le taux de bonne segmentation. Nous espérons donc que celui-ci sera amélioré par la fusion des descripteurs numériques et des concepts.

Le tableau 8.5 présente les performances de segmentation des groupes de scènes par traitement combiné des descripteurs numériques et des concepts. Deux algorithmes de fusion sont comparés, le premier est un algorithme de fusion des points de coupures par plus proche voisin et le deuxième réalise la segmentation sur la séquence des vecteurs descripteurs globaux comprenant les concepts et les descripteurs bas.

	TP	TR
PPV	50	28
Concaténation	66	57

TAB. 8.5 – Performances de la segmentation de l'image en groupe de scènes par les descripteurs numériques et les concepts

Les résultats montrent que la performance varie de facon notable selon le modèle. Deux constatations peuvent être faites.

Premièrement, les performances du modèle de segmentation par fusion des descripteurs numériques et textuels sont supérieures, en combinant les deux informations au sein d'un vecteur score global, plutôt qu'en fusionnant les deux ensembles de points de coupures par les PPV. L'utilisation du vecteur global semble donc importante pour l'identification des variations du contenu visuel. La fusion par PPV tend à supprimer les points de coupure qui n'ont pas été détectés par les deux ensembles de descripteurs : la fusion permet alors de réduire le taux de sur-segmentation, mais entraîne aussi une diminution des performances de bonne segmentation. Seule la segmentation par le vecteur global peut sélectionner des points de coupures ambigus (ceux détectés par un seul type de descripteurs). Les conséquences sont minimes sur le taux de sur-segmentation, car il est rare que les variations accidentelles du contenu se retrouvent dans les deux ensembles de descripteurs ; mais elles sont importantes sur le taux de bonne segmentation.

Deuxièmement, dans ce cas, la fusion des deux types de descripteurs amène une légère amélioration de la segmentation des groupes de scènes par rapport au modèle numérique seul. Les deux méthodes de fusion permettent de réduire la sur-segmentation. Seule la deuxième méthode améliore le taux de bonne segmentation, car elle prend bien en compte l'information apportée par les deux types de descripteurs. C'est pourquoi nous l'emploierons dans la suite pour cette catégorie de fusion.

Troisièmement, les résultats de la segmentation des groupes de scènes par les descripteurs images sont assez mauvais. Nous détectons à peine la moitié des points de coupure. Les groupes de scènes sont de grands blocs narratifs du film ; au sein d'un groupe de scènes les plans sont reliés par une sémantique complexe. Il semble, que le niveau d'abstraction du contenu image soit trop faible pour bien représenter l'homogénéité du contenu à ce niveau. Pour améliorer la segmentation, il nous faudrait extraire certaines caractéristiques plus abstraites liées à l'intériorité des personnages et du spectateur : tensions, émotions, buts, en un mot les sentiments. Ce qui est une tâche complexe pour un ordinateur.

La segmentation du signal **sonore** d'un film en groupes de scènes est réalisée par la classification MMC non supervisée de la séquence des clips audio. Des expériences similaires à celles décrites ci-dessus pour l'image ont été réalisées pour l'audio. Elles ont montré l'efficacité de la fusion des descripteurs numériques et textuels pour la segmentation du signal sonore des films. Pour l'audio, les deux méthodes de fusion permettent de réduire la sur-segmentation et seule la deuxième méthode, fondée sur la concaténation des descripteurs, améliore le taux de bonne segmentation.

Le tableau 8.6 montre que les performances de segmentation du son varient par rapport à celles obtenues pour l'image.

	TP	TR
Concaténation	50	71

TAB. 8.6 – Performances de la segmentation du son en groupe de scènes par les descripteurs numériques et les concepts

Nous remarquons que le traitement de l'audio est plus performant que celui de l'image pour la segmentation des groupes de scènes. Les variations des caractéristiques de la bande sonore semblent donc bien correspondre à ce niveau structurel. Au sein d'un film, la bande sonore varie de façon assez lente, ainsi les segments homogènes pour le contenu audio sont plus longs temporellement que les segments image, ce qui diminue la sur-segmentation. De plus, aujourd'hui, la musique joue un rôle prépondérant dans la narration : souvent, elle est présente tout le long du film (voir « Harry Potter »). Nous avons constaté que la structure de la bande musicale est calquée sur celle du film qu'elle accompagne, c'est d'autant plus le cas pour les musiques composées à cet effet. Or, par la musique "tous les efforts, toutes les émotions et les manifestations possibles de la volonté, tous ces processus intérieurs de l'homme, que la raison englobe dans la notion de "sentiment", peuvent être exprimées à l'aide de la multitude infinie des mélodies possibles, mais toujours exclusivement dans la généralité de la forme pure, sans la substance, toujours seulement en tant qu'en soi, et non en tant qu'apparence, en quelque sorte comme l'âme de l'apparence, incorporellement" [Sch19]. Ainsi, la capacité de la musique à illustrer l'intériorité des personnages par des représentations simples permet d'extraire la structure des groupes de scènes à partir de la séquence de descripteurs numériques. Les caractéristiques fondées sur le cepstre audio, le tempo, la fréquence fondamentale, semblent bien représenter la sémantique abstraite utile pour segmenter à ce niveau.

La **fusion du son et de l'image** pour la segmentation d'un film en groupes de scènes est réalisée par un modèle de combinaison des points de coupure audio et image fondé sur les plus proches voisins.

	TP	TR
Image	66	57
Son	50	71
Multimedia	75	57

TAB. 8.7 – Performances de la segmentation en groupe de scènes multimedia

Le tableau 8.7 compare les performances obtenues par les modèles de segmentation vidéo, audio et multimedia. Nous en tirons deux remarques.

Premièrement, les résultats de sur-segmentation du modèle multimedia sont supérieurs à ceux observés pour les modèles monomedia. L'utilisation du modèle de fusion par les PPV diminue ici aussi le nombre de fausses coupures.

Deuxièmement, en ce qui concerne les performances de bonne segmentation du modèle multimedia, nous remarquons une amélioration par rapport au modèle image. Cependant le modèle de fusion PPV tend à favoriser l'élimination de points de coupure, c'est pourquoi il n'atteint pas les performances de bonnes classifications réalisées par le modèle audio.

Pour conclure, il semble que dans le cas particulier des groupes de scènes, où la sémantique nécessaire pour segmenter est forte, l'utilisation du signal sonore soit indispensable ; les descripteurs du contenu image expriment mal la sémantique forte des caractéristiques utiles pour cette tâche.

Segmentation des scènes

La segmentation des groupes de scènes en scènes est réalisée par la classification MMC non supervisée de la séquence des plans et des clips. Afin de mettre en valeur le rôle joué par les différents media, la segmentation est réalisée sur les ensembles de descripteurs image, audio et multimedia. De plus nous utilisons les limites des groupes de scènes annotées à la main plutôt que celles détectées précédemment, afin d'évaluer les performances du système pour le niveau structurel concerné. Nous ferons de même pour les groupes de plans.

	TP	TR
Image	64	90
Son	72	80
Multimedia	80	80

TAB. 8.8 – Performances de la segmentation en scènes

Le tableau 8.8 compare les performances obtenues par les modèles de segmentations pour les trois ensembles de descripteurs. Plusieurs conclusions peuvent être tirées de cette étude.

Premièrement, les performances de segmentation des scènes observées sont supérieures à celles de la segmentation des groupes de scènes, et ceci pour les trois ensembles de descripteurs. Nous apportons plusieurs explications : la segmentation est réalisée sur chaque groupe de scènes du film de façon individuelle. Dans ce cas, les variations du contenu des scènes sont mieux mises en évidence que si l'intégralité du film était traitée en une fois. De plus, nous avons observé, en pratique, l'homogénéité des descripteurs choisis au sein des scènes. En effet, ils capturent efficacement la signature visuelle et auditive du lieu dont l'unité est une caractéristique essentielle à ce niveau structurel. Nous pouvons donc affirmer que l'ensemble des descripteurs bas et des concepts utilisés ici est bien adapté à la segmentation des films de cinéma en scènes. Enfin, nous constatons que les problèmes liés à la présence, au sein d'une scène, d'un plan (ou clip) dont le contenu est très différent du reste de la scène sont minimisés, car les informations ambiguës apportées par ce plan sont compensées par les plans de son voisinage. Ainsi la classification MMC et le choix de la fenêtre de lissage sont essentiels à la prise en compte de la continuité temporelle.

Deuxièmement, la comparaison des modèles audio et image montre qu'ils obtiennent des résultats différents. Tout d'abord, la segmentation de la bande sonore produit moins de points de coupures que celle de l'image, car le signal visuel est plus sujet à des variations brusques de son contenu et l'algorithme de segmentation détecte plus de classes que pour l'audio. L'algorithme audio présente ainsi un taux de sur-segmentation supérieur et de bonne segmentation inférieur comparé à l'algorithme de segmentation de l'image.

Troisièmement, nous constatons qu'en ce qui concerne la segmentation multimedia, les problèmes rencontrés précédemment persistent. La fusion des points de coupure audio et image par PPV entraîne une sous-segmentation flagrante. Les conséquences en sont une amélioration du taux de sur-segmentation par rapport au monomedia, et une baisse des performances de bonne segmentation par rapport à l'image.

Pour conclure, il semble qu'au niveau structurel des scènes, l'utilisation de concepts visuels liés à la description du lieu et des éléments du décors présents améliore fortement la segmentation par les descripteurs numériques. Ceci en raison de l'unité de lieu caractéristique des scènes.

Segmentation des groupes de plan (et clips)

La segmentation des scènes en groupes de plans et groupes de clips est réalisée par la classification MMC non supervisée de la séquence des plans et des clips. Nous utilisons les limites des scènes annotées à la main plutôt que celles détectées précédemment.

	TP	TR
GPlans	75	87
GClips	76	82

TAB. 8.9 – Performances de la segmentation en groupes de plans et groupes de clips

Le tableau 8.9 montre les performances obtenues par les deux modèles de segmentation. Deux remarques peuvent être faites.

Premièrement, les résultats de bonne segmentation montrent que le modèle utilisé est performant à ce niveau de la segmentation temporelle. La séquence des descripteurs choisis représente de façon efficace la variation du contenu liée à la structure des groupes de plans (et clips). En ce qui concerne l'image, les descripteurs numériques influencent fortement la segmentation, car ils contiennent les informations de couleur et de texture du fond et ainsi permettent de déceler les changements de la grosseur des plans (e.g. plan général vers gros plan) ou d'autres caractéristiques du changement de groupes de plans. De même, l'utilisation des concepts semble appropriée ici, en particulier, le concept « nombre de personnages », car l'entrée ou la sortie d'une personne indique souvent un point de coupure. En ce qui concerne la bande sonore, de façon similaire, nous avons noté l'importance des descripteurs numériques. Les variations des caractéristiques physiques simples de l'audio marquent souvent un changement de l'action ou de l'emplacement de l'action. Les concepts audio de la classification hiérarchique *parole/musique* apportent aussi une information essentielle pour la segmentation à ce niveau. Par exemple, une scène chez le médecin : le personnage attend dans la salle d'attente : brouhaha, éternuements etc. . . ; il entre dans le cabinet, le bruit cesse, une conversation commence ; de même pour une course poursuite, par exemple : la musique est absente au début de la scène, puis apparaît lorsque l'action démarre, ce qui a pour effet d'accentuer la tension.

Deuxièmement, le taux de sur-segmentation pour les deux media est assez élevé. Nous avons identifié plusieurs problèmes : tout d'abord l'algorithme de classification non-supervisée est faussé lorsqu'une scène comprend un seul groupe de plans (ou clips) ; le modèle de classification utilisé ne permet pas d'apprendre des concepts unaires et produit dans ce cas une segmentation erronée. Par exemple, souvent les scènes de dialogues présentent un seul groupe de clips constitué d'une alternance de voix différentes, et l'algorithme propose une segmentation calquée sur cette alternance. De plus, la segmentation est sensible aux plans (ou clips) pathologiques, car le choix de la taille de la fenêtre de lissage est limitée par la taille de la scène traitée. Nous avons fixé $w = 3$.

Pour conclure, il semble qu'à ce niveau, l'utilisation de concepts audio, comme la présence de musique ou de parole, et de concepts image, comme le nombre de personnages, améliore fortement la segmentation par les descripteurs numériques. Les informations complexes caractérisées par ces descripteurs

semblent essentielles pour déceler l'homogénéité du contenu des plans et des clips au sein de ces structures.

Algorithme de comparaison

Enfin les performances de notre algorithme de segmentation du bas vers le haut sont comparées à celles des modèles de référence de la littérature, dont les étapes sont :

- Extraction des scènes par fusion des limites de plans et de clips.
- Segmentation par écrêtage des scènes en groupes de plans.
- Regroupement MMC des scènes en groupe de scènes.

	TP	TR
Groupe de scènes	50	57
Scènes	53	83
Groupes de plans	66	76

TAB. 8.10 – Performances de la segmentation de la structure par les algorithmes de référence

Le tableau 8.10 montre que les performances de segmentation varient selon le modèle utilisé. Plusieurs observations en découlent.

Premièrement, en ce qui concerne le niveau des groupes de scènes, les résultats obtenus par l'algorithme de référence sont supérieurs à ceux de notre algorithme. Le modèle de référence est fortement avantagé, car l'information apportée par la segmentation manuelle des scènes facilite la segmentation. De plus, la fusion des media est réalisée par la concaténation des descripteurs audio et image issus des scènes, ce qui n'entraîne pas les problèmes de sous-segmentation des méthodes par PPV.

Deuxièmement, pour le niveau des scènes, notre algorithme dépasse l'algorithme de référence. La technique de segmentation des scènes par fusion des limites de plans et de clips détecte trop de points de coupure, car il n'est pas rare que deux limites coïncident sans que cela marque la fin d'une scène. Nous constatons donc un taux de sur-segmentation assez élevé pour ce modèle. De plus notre algorithme est avantagé par la connaissance a priori de la structure des groupes de scènes.

Troisièmement, pour la segmentation des groupes de plans, notre modèle présente des performances supérieures à celle du modèle de référence. Le modèle MMC non-supervisé semble donc approprié pour la segmentation des media. Pour sélectionner un point de coupure, le modèle de segmentation par écrêtage ne considère pas l'ensemble de la scène, mais seulement les plans situés au voisinage de celui-ci. Cette perte d'information peut se révéler nuisible. De plus, le seuil d'écrêtage a été fixé à la main après plusieurs expériences, mais il devrait être différent pour chaque scène car il est lié à la variation particulière du contenu au sein des groupes de plans. Les conséquences sont multiples. Nous constatons, notamment, une sur-segmentation lorsque le contenu varie beaucoup dans une scène, par exemple lorsque le champ/contre-champ d'un dialogue montre deux vues très différentes.

8.2.4 Résumé des résultats

Les résultats de ce chapitre peuvent se résumer par trois affirmations.

Premièrement, la prise en compte de l'information fournie par les **concepts** en plus de l'information des descripteurs numériques semble faciliter l'identification de la structure des films. L'amélioration par rapport au modèle numérique uniquement est particulièrement sensible pour la segmentation des scènes, car plusieurs des concepts choisis concernent la description du lieu de l'action, une caractéristique essentielle de ce niveau structurel. Globalement, la fusion des concepts aux descripteurs numériques apporte une information supplémentaire et significative sur le contenu des media traités.

Deuxièmement, l'information apportée par le signal sonore en sus du signal image perfectionne les capacités de segmentation du système. Notamment en ce qui concerne les groupes de scènes, car la

musique représente de façon "brute", sans apparence, certains concepts abstraits liés à l'intériorité du personnage. Les descripteurs numériques audio "portent" ainsi en eux une certaine abstraction utile pour la segmentation à ce niveau. De façon générale, les descripteurs audio procurent une information supplémentaire et appropriée à la segmentation de la structure des films.

Troisièmement, la comparaison des performances de segmentation du **modèle de référence** et de notre modèle nous permet de conclure sur la capacité de la segmentation hiérarchique MMC à capturer les caractéristiques essentielles de la structure des films.

Conclusion

Nous avons développé au cours de ce travail, un cadre probabiliste de construction et d'utilisation de modèles de classification multimédia, ensuite appliqué à l'indexation de films de cinéma. Nous avons d'abord montré que le modèle de fusion des descripteurs numériques et textuels permet de déterminer avec une performance satisfaisante plusieurs concepts monomédias présents dans les films, dans la mesure où les descripteurs choisis sont pertinents pour la classification. Ensuite, nous avons constaté une augmentation des performances de détermination d'ambiances, en combinant l'information auditive et visuelle du film au sein d'un modèle de fusion de données multimédias. Enfin, nous avons utilisé les résultats de classification pour la segmentation et l'indexation de films de cinéma, avec de meilleures performances que les méthodes numériques existantes, notamment pour la segmentation des scènes.

D'un point de vue théorique, nous avons combiné, dans les modèles de classification de concepts proposés, deux types de descripteurs habituellement considérés comme différents : les descripteurs numériques et les annotations textuelles dites « atomiques ». La fusion de ces informations est réalisée par un modèle probabiliste génératif inspiré par les réseaux bayésiens et les SVM (Support Vector Machine) hiérarchiques. Nous avons aussi décrit une méthode pour utiliser ces modèles efficacement dans le cadre de la classification de concepts multimédias. Nous proposons plusieurs extensions théoriques différentes à ce travail.

Premièrement, en ce qui concerne les annotations atomiques d'images, celles ci sont déterminées, pour la plupart, par un opérateur humain. Dans le cadre d'une indexation du contenu, il serait nécessaire d'automatiser cette tâche. Par exemple, il est possible d'extraire certains concepts de « textures », comme du ciel, de l'eau ou de l'herbe ; et de détecter la présence d'objets, comme une voiture ou une personne. Les taux de confiance de ces classifications gouverneraient la détermination des concepts de niveau sémantique supérieur. Une question se pose alors : quelle sera l'influence des erreurs de classification atomique sur la décision finale ? Les performances obtenues par les algorithmes de la littérature pour la détection d'objets ou de textures ne semblent, en effet, pas suffisantes pour améliorer les résultats de classification par les descripteurs numériques uniquement. D'autant plus que la présence de certains concepts atomiques influence fortement la classification du concept supérieur. Une solution serait de tenter d'améliorer les performances de détection d'objets par l'utilisation du signal sonore. Le modèle de fusion multimédia décrit ici pourrait s'appliquer à cette tâche. Une autre possibilité consisterait à modifier la distribution des scores obtenus par les concepts atomiques afin de minimiser le poids des fausses détections. Cette normalisation supplémentaire pourrait permettre au modèle de classification de ne pas considérer les concepts dont la présence est ambiguë et de favoriser d'autres concepts, ou les descripteurs numériques.

Deuxièmement, pour le signal sonore, la fusion des descripteurs numériques et des concepts atomiques choisis n'apporte pas de gain de performance par rapport au modèle numérique simple. Il serait intéressant de préférer des concepts audio atomiques appropriés à la catégorisation automatique des lieux. Par exemple, la présence de chants d'oiseaux, ou de klaxons de voitures pourrait influencer la classification. Les performances actuelles obtenues par les modèles de séparation et reconnaissance de sources audio permettent d'envisager cette application.

Troisièmement, en ce qui concerne la fusion des médias audio et visuel pour la classification des lieux, il serait utile de réfléchir à une méthode d'identification des plans ambigus dont l'un des médias perturbe la classification. Nous pouvons envisager la reconnaissance des images ou des segments sonores corrompus par un algorithme de classification appris sur les résultats de la classification des lieux. Ce qui permettrait de repérer les erreurs classiques du système dues à des gros plans ou à un son bruité, par exemple.

Quatrièmement, il serait utile, dans le cadre de la segmentation de films, de proposer une identification plus complexe des lieux. En effet, les classes choisies a priori pour cette classification ne sont pas toujours appropriées aux films à indexer. La quantité d'information disponible pourraient être mise à profit afin de réaliser un apprentissage non supervisé de modèles de lieux, permettant de distinguer entre deux plans de lieux différents et deux plans d'un même lieu présentés d'un point de vue (et d'écoute) différent.

D'un point de vue appliqué, nos résultats les plus originaux ont concerné la classification multimédia à partir de descripteurs numériques et d'annotations textuelles, ainsi que la segmentation de film de cinéma. Cependant, nous avons testé les performances de nos algorithmes dans des conditions assez limitées. Ces limitations proviennent principalement des difficultés liées à l'apprentissage des modèles de classe et en particulier au coût temporel de l'annotation.

Premièrement, nous avons testé nos algorithmes sur un seul type de concepts multimédias : les concepts de lieux. Pour ceux-ci, nous avons montré, sur des plans de vidéo, que le modèle développé de fusion des descripteurs fournit de bonnes performances de classification. Il aurait été intéressant d'expérimenter différentes classifications multimédias, comme l'identification de personnes ou la détection d'évènements. Dans le cas de l'identification de personnes, nous avons vérifié l'efficacité de l'algorithme de classification pour l'audio uniquement. Nous faisons l'hypothèse que les résultats obtenus par notre modèle de fusion sont valables dans le cas multimédia, car il répond aux problématiques de la fusion de médias différents comme les phénomènes de masquages. Les attentes sont donc prometteuses, mais cela ne remplace en rien des tests approfondis sur des cas réels.

Deuxièmement, hormis pour la classification de l'audio en parole/musique, nous avons testé toutes les tâches sur des bases d'exemples trop réduites. La collecte d'un nombre suffisant de sons ou de plans de films étiquetés pose un véritable problème. Pour l'identification audio de personnes, en raison des contraintes temporelles, il nous a été impossible d'apprendre les voix des acteurs choisis à partir de plusieurs films. Sur les exemples de voix que nous avons trouvés de cette façon, nous avons constaté la présence de bruit de fond ou de musique. Dans ces circonstances, l'homogénéité du fond sonore au sein d'un film fausse l'identification des voix. Le problème est encore plus manifeste en ce qui concerne l'identification multimédia des lieux. Dans ce cadre, nous avons essayé de créer une base de données adaptée contenant les exemples nécessaires pour l'apprentissage et l'évaluation des modèles de fusion multimédia. Mais le temps prohibitif de l'annotation manuelle ne nous a pas permis de réunir une quantité homogène et suffisante d'exemples, ce qui peut entraîner un sur-apprentissage des modèles de certains concepts peu représentés.

La troisième limitation concerne nos conditions de test de l'algorithme de segmentation multimédia. Elle est due à la difficulté de créer une base de données assez importante pour être représentative de l'ensemble des films de cinéma. Nous avons testé les tâches de segmentation sur un nombre trop limité de films. La performance de notre algorithme ne peut donc pas être comparée à celle des algorithmes existant pour le même type de segmentation.

Annexe A

Annexe

Cette annexe décrit plusieurs techniques intervenant au long de ce travail. Nous présentons dans le paragraphe A.1 un modèle de représentation temporelle des concepts de description. Ensuite, dans le paragraphe A.2 nous décrivons le modèle de classification Support Vector Machine (SVM) utilisé dans le cadre de l'indexation du contenu. Enfin, nous exposons dans le paragraphe A.3 les modèles de Markov cachés employés pour la segmentation du signal des films et pour l'indexation de la structure.

A.1 Représentation temporelle des concepts : modèles de strates

Afin de visualiser les variations de ces concepts dans le temps, certains utilisent un **modèle de représentation par strates**. Une strate est une liste de plans de vidéo auxquels est attaché un concept. Elle regroupe des plans d'images qui partagent une sémantique commune, représentée par un concept. Ces concepts sont issus des classifications de plans présentées ultérieurement. Chaque strate est associée à une liste de plans ordonnés chronologiquement. Comme le montre la figure 2, les strates d'une vidéo peuvent avoir des plans en commun. Cela signifie que les objets, évènements ou actions contenus dans les annotations respectives de ces deux strates apparaissent ou se produisent simultanément dans les segments d'images qui se chevauchent.

Les annotations peuvent être attachées à tous les niveaux de la structure cinématographique (scène, séquence) de la vidéo qui est modélisée elle-même par une hiérarchie de classes. Cette approche "tout objet" facilite la représentation d'objets complexes, l'identification des objets pour une éventuelle réutilisation, et permet l'héritage des attributs et des méthodes. Nous avons vu, de plus, son intérêt, dans le cadre de la détection de la structure cinématographique (scène, séquence) de la vidéo. En conclusion, la puissance d'expression d'un modèle d'indexation est donc liée à sa capacité à définir finement des strates, les éventuelles relations (ensemblistes, temporelles...) entre les strates, et les liens entre strates et annotations associées. Non moins important est le choix d'un formalisme de représentation de connaissances (logique, relationnel, objets, graphes conceptuels, réseaux sémantiques...) pour représenter les annotations chargées de la description sémantique de haut niveau.

A.2 Modèle de classification SVM

Fondamentalement, les méthodes SVM (figure A.2) projettent les données à classer dans un espace de grande dimension, en utilisant un critère linéaire.

Suivant les données d'entraînement, la projection appropriée Φ est choisie telle que la séparation linéaire entre classes soit effective. Le calcul est fait sans connaître une forme explicite de la projection, mais seulement grâce aux noyaux correspondant au produit scalaire entre projections. Le modèle est spécifié par le choix de ce noyau $\ker$: $\ker(X, X') =< \Phi(X), \Phi(X') >$, et f la fonction dont le signe donne

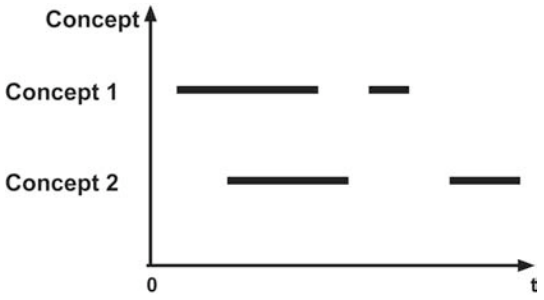

FIG. A.1 – Modèle de représentation des concepts par strates.

l'approximation de la classe d'un objet et la valeur absolue fournit une estimation du taux de confiance dans cette classification : $f(X) =< \omega, \Phi(X) > +b$, où ω est le vecteur orthogonal à l'hyperplan H de séparation. Dans la littérature, plusieurs types de noyaux ont été présentés, afin de décrire au mieux la structure topologique des données. Nous pouvons citer les noyaux linéaires, polynomiaux, gaussiens, sigmoïdes et un nombre important de variantes.

Soit Σ la base d'apprentissage, pour tout élément O de Σ le couple (D, C) est connu, où D est la position de O dans l'espace des descripteurs et C la valeur du concept de classification (ici -1 ou 1) du $i^{ème}$ objet. La détermination du modèle de classification consiste donc à déterminer ω afin de maximiser la marge entre les points d'entraînement et l'hyperplan les séparant : les SVM minimisent le risque structurel de mauvaise classification [Ada02]. Ce qui assure une bonne généralisation des propriétés de la vraie population.

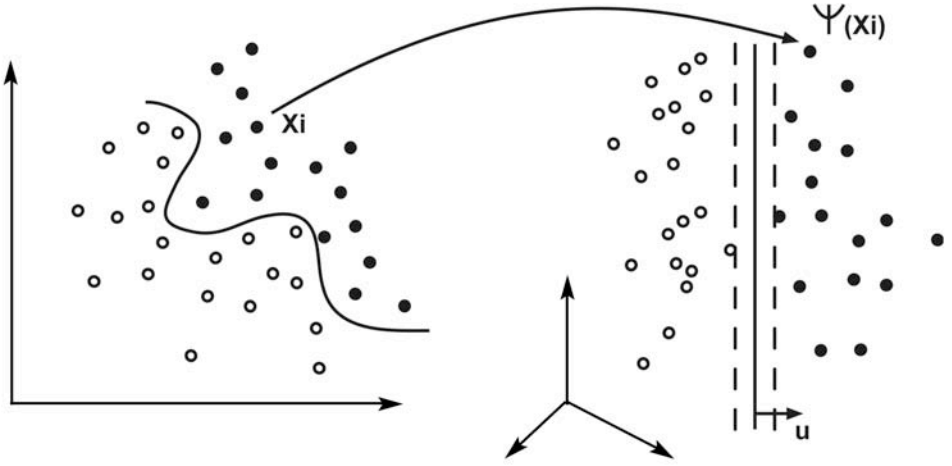

FIG. A.2 – Modèle de classification SVM.

Les SVM sont des approches dites discriminantes : elles se concentrent sur les caractéristiques des descripteurs qui permettent la séparation des classes du concept choisi.

La plus grosse limitation des approches par SVM tient au choix du noyau[Bur98]. Une fois qu'il est fixé, les SVM n'ont qu'un seul paramètre choisi par l'utilisateur, l'error penalty. Mais le noyau est une grosse boite noire dans lequel les paramètres sont cachés. Le meilleur choix pour ker est donc toujours un sujet de recherche. Dans le cadre de notre étude nous choisirons un noyau gaussien, pour des raisons de simplification. Une seconde limitation est la rapidité et la taille du modèle, pour les tests et entraînement : il nécessite toujours deux passages. Ce qui peut être gênant lorsque l'on dispose de millions de données à traiter.

A.3 Les modèles de Markov cachés : MMC

Le modèle de Markov caché est fortement apparenté aux automates probabilistes. Un automate probabiliste est une structure composée d'états et de transitions, et d'un ensemble de distributions de probabilités de transitions entre états. La figure A.3 représente un automate à 3 états, les flèches indiquent les transitions possibles entre états.

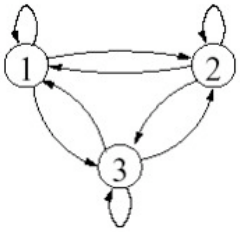

FIG. A.3 – Automate probabiliste à trois états.

La différence essentielle entre les MMCs et les automates probabilistes se situe dans le fait que, contrairement aux automates probabilistes, la génération des symboles du MMC se fait au niveau des états et non sur les transitions. A chaque état est associé une distribution de probabilité de transition sur l'ensemble de ces transitions. Les éléments caractéristiques d'un modèle MMC noté $\mathcal{MMC}_d(C, \mathbf{A}, \mathbf{B})$ sont :

C un concept à p états et $\mathcal{M}(\mathrm{T}, \mathrm{X}, \Psi)$ son modèle de classification associé,

$\mathbf{A} = \{\alpha_{j_1 j_2}\}$ la matrice des probabilités de transition entre états et

$\mathbf{B} = \{\beta_j(t)\}$ la matrice des probabilités d'émission des états.

Ces probabilités s'expriment par :

$$\begin{array}{c} \alpha_{j_1 j_2} = P(C_{t+1} = c_{j_2} | C_{t+1} = c_{j_1}) \\ et \\ \beta_j(t) = P(D | C_t = c_j) \end{array} \tag{A.1}$$

Pour une chaîne de Markov à temps discret du premier ordre, la probabilité d'occupation d'un état dépend uniquement de l'état précédent ; les probabilités de transition $\alpha_{j_1 j_2}$ sont donc indépendantes du temps. Le but est de déterminer la séquence la plus probable d'états qui correspond à la suite de segments. Comme pour la classification, ou la segmentation, il existe deux hypothèses distinctes pour ce faire.

Cas supervisé

- la taxonomie C est connue a priori.
- $\mathbf{A}$ est appris sur une base de séquences de référence.
- les $\beta_j(t)$ de $\mathbf{B}$ sont déterminés sur la séquence en question par classifications successives de segments.

En général les probabilités d'émission sont calculées à partir de modèles statistiques gaussiens. Mais elles peuvent être estimées à partir d'autres classifications comme les ppv ou les SVM après normalisation des taux de confiance. On se trouve alors dans le cas d'un problème de « **décodage** » : le modèle de classification et les probabilités d'émission sont connus, la séquence la plus probable reste à déterminer. Ceci est réalisé par l'**algorithme de Viterbi** qui extrait la séquence dont la probabilité d'émission est maximum.

Cependant, lorsque les structures isolées ne peuvent pas être décrites par des concepts déterminés a priori (c'est le cas pour la segmentation d'un film par exemple, où le modèle des scènes d'un ne peut pas être connu à l'avance), il est nécessaire d'apprendre le modèle MMC.

Cas non supervisé

On est ici dans le cas où la taxonomie de la segmentation n'est pas connue a priori. De même pour les matrices de transition et d'émission . Comme il a été vu, pour la classification, le problème consiste à trouver la taxonomie et son modèle de classification qui correspondent le mieux à la structure des données. C'est-à-dire qu'il faut déterminer le modèle MMC, $\mathcal{MMC}(C, \mathbf{A}, \mathbf{B})$ le plus probable sachant les observations produites par ce modèle. Nous nous trouvons donc dans le cadre d'un problème d'**apprentissage**. Lorsque le modèle est appris, le problème devient un problème de décodage, et l'algorithme de Viterbi est appliqué pour déterminer la séquence des concepts.

Généralement, le problème d'apprentissage revient à ajuster les paramètres MMC pour que les observations soit représentées le mieux possible pour l'application donnée. La quantité à optimiser durant le processus d'apprentissage peut être différente suivant les applications. Il existe deux critère d'optimisation courants dans la littérature : "maximal likelihood" (ML) utilisé par Baum et Welsh ou d'autres [Bag01] et le "maximum mutual information" (MMI). Dans le cadre de notre application nous choisirons l'algorithme de **Baum&Welsh**, pour des raisons de simplicité. Nous ne rentrerons pas dans les détails de ces algorithmes : voir [Gop98]. Mais il est possible de les résumer ainsi. Ce sont des algorithmes itératifs qui permettent de trouver le meilleur modèle de séquence pour une séquence donnée. Que ce soit l'algorithme de Baum Welsh ou les modèles par descentes de gradient, ils sont tous deux initialisés par une segmentation par classification non supervisée, sans prise en compte du temps. [Bre98] utilise un modèle de regroupement par k-mean pour le résumé musical, mais, on l'a vu, il existe de nombreux algorithmes plus efficaces pour cette tâche.

Bibliographie

[ace] *Acemedia : Integrating knowledge, semantics and content for user-centered intelligent media services, www.acemedia.org/aceMedia/.*

[Ach96] B. Achermann and H. Bunke. Combination of classifiers on the decision level for face recognition. In *Conference on Pattern Recognition*, 1996.

[Ada02] B. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C-Y. Lin, A. Natsev, M. Naphade, C. Neti, H.J. Nock, H.H. Permuter, R. Singh, J.R. Smith, S. Srinivasan, B.L. Tseng, Ashwin T. V., and D. Zhang. Ibm research trec-2002 video retrieval system. In *TREC*, 2002.

[Ajm02] J. Ajmera, A. McCowan, and H. Bourlard. Robust hmm-based speech/music segmentation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[Ala98] A.Aydin Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora. Image sequence analysis for emerging interactive multimedia services- the european cost 211 framework. *IEEE Transactions on Circuits and Systems for Video Technology*, 8 :7, 1998.

[Aqa02] W.H. Aqams, G. Iyengar, C-Y Lin, M.R. Naphade, C. Neti, H.J. Nock, and J.R. Smith. Semantic indexing of multimedia content using visual, audio and text cues. In *International Conference on Multimedia and Expo (ICME)*, 2002.

[Ass98] J. Assfalg, C. Colombo, A. Del Bimbo, and P. Pala. Embodying visual cues in video retrieval. In *IAPR International Workshop on Multimedia Information Analysis and Retrieval*, pages 47–59, 1998.

[Bab99] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based video indexing by intermodal collaboration. In *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, pages 1–9, 1999.

[Bag97] P. M. Baggenstoss. Structural learning for classification of high dimensional data. In *Int. Conf. Intelligent Systems Semiotics*, pages 124–129, 1997.

[Bag01] P.M. Baggenstoss. A modified baum-welch algorithm for hidden markov models with multiple observation spaces. *IEEE Transactions on Speech and Audio Processing*, 9(4), 2001.

[Baj97] R. Bajaj and S. Chaudhury. Signature verification using multiple neural classifiers. *Pattern Recognition*, 30(1) :1–7, 1997.

[Ben98] S. BenYacoub, J. Luttin, K. Jonsson, J. Matas, and J. Kittler. Audio-visual person verification. In *In Computer Vision and Pattern Recognition*, pages 580–585, 1998.

[Ben02] P. N. Bennett, S. T. Dumais, and E. Horvitz. Probabilistic combination of text classifiers using reliability indicators : Models and results. In *Special Interest Group on Information Retrieval (SIGIR)*, 2002.

[Bik99] D.M. Bikel, R. Schwartz, and R.M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34 :211–231, 1999.

[Bim98] J.A. Bimes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hmm. *ICSI TR-97-021*, 1998.

[Bim04] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Chagnolleau, S. Meigner, T. Merlin, J. Ortega, D. Petrovska, and D.A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on applied signal processing*, pages 430–451, 2004.

[Blu04] P. Blunsom. Hidden markov models, www.cs.mu.oz.au/460/materials/hmm-tutorial.pdf, 2004.

[Bor98] J.S. Boreczky and L.D. Wilcox. A hidden markov model framework for video segmentation using audio and image features. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

[Bre98] J.S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

[Bru95] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

[Bru99] R. Brunelli, O. Mich, and C.M. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2) :78–112, 1999.

[Bur98] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2) :121–167, 1998.

[Car94] T. Carron and P. Lambert. Color edge detector using jointly hue, saturation and intensity model for stained images. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 977–981, 1994.

[Cas98] M. La Cascia, S. Sethi, , and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998.

[Cas01] M.A. Casey. Reduced-rank spectra and minimum entropy priors as consistent and reliable cues for genralized sound recognition. In *Workshop for Consistent and Reliable Cues, Aalborg, Denmark*, 2001.

[Cer] Le Cerveau. *La perception visuelle dévoilée par les illusions d'optiques, http ://www.lecerveau.mcgill.ca*.

[Cha92] S. Chang and A. Hsu. Image information systems : Where do we go from here ? *IEEE Trans. on Knowledge and Data Enginnering*, 4(5) :431–442, 1992.

[Cha98] S. Chang, W. Chen, H.J. Horace, H. Sundaram, and D. Zhong. A fully automated content based video search engine supporting spatio-temporal queries. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 8(5) :601–615, 1998.

[Cha99] E. Chavez, G. Navarro, R. Baeza-Yates, and J. Marroquin. Searching in metric spaces. Technical Report TR/DCC-99-3, Dept. of Computer Science Univ. of Chile, 1999.

[Che98] S.S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA*, 1998.

[Che03a] S. Cheng and H. Wang. A sequential metric-based audio segmentation method via the bayesian information criterion. In *European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, 2003.

[Che03b] Y.-C. Cheng and S.-Y. Chen. Image classification using color, texture and regions. *image and Vision Computing*, 21 :759–776, 2003.

[Cla51] R. Clair. *Réflexion faite*. Gallimard, 1951.

[Coi94] R.R. Coifman and M.V. Wickerhauser. Adapted waveform analysis as a tool for modeling, feature extraction and denoising. *Optical Engineering*, 33(7) :2170–2174, 1994.

[Com01] MPEG-7 Committee. Overview of the mpeg-7 standard (version 6.0). Technical report, Report ISO/IEC JTC1/SC29/WG11 N4509, 2001.

[Coo02] M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*, 2002.

[Cou01] L. Couvreur, C. Ris, and C. Couvreur. Model-based blind estimation of reverberation time : Application to robust asr in reverberents environments. In *IEEE Transactions European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001.

[Dee90] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6) :391–407, 1990.

[Del03] B. Delezoide and X. Rodet. Audio features selection for speech/music discrimination, 2003.

[Dug96] R. Dugad and U.B. Desai. A tutorial on hidden markov models, 1996.

[Dum00] S. Dumais and H. Chen. Hierarchical classification of web content. In *Special Interest Group on Information Retrieval (SIGIR)*, 2000.

[Eis58] S. Eisenstein. *Journal d'un cinéaste*. Editions en langues étrangères, Moscou, 1958.

[Ens93] P.G.B. Enser. Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1) :25–39, 1993.

[Erp00] M. Van Erp and L. Schomaker. Variants of the borda count method for combining ranked classifier hypotheses. In *Proceedings of the seventh International Workshop on Frontiers in Handwriting Recognition*, pages 443–452, 2000.

[Fis36] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, (7) :179–188, 1936.

[Fle96] M.M. Fleck, D.A. Forsyth, and C.Bregler. Finding naked people. In *4th European Conference on Computer Vision*, pages 591–602, 1996.

[Foo94] J. Foote. *Decision-Tree Probability Modeling for HMM Speech Recognition*. PhD thesis, Cornell University, 1994.

[Foo99] J. Foote. An overview of audio information retrieval. *Multimedia Systems*, 7(1) :2–11, 1999.

[Foo03] J. Foote and M. Cooper. Media segmentation using self-similarity decomposition. In *SPIE Storage and Retrieval for Multimedia Databases*, volume 5021, pages 167–75, 2003.

[For02] D. A. Forsyth and J. Ponce. *Computer Vision : a modern approach*. Prentice-Hall, 2002.

[Fou66] M. Foucault. *Les Mots et les Choses, Archéologie des sciences humaines*. Gallimard, 1966.

[Fre98] B.J. Frey. Graphical models for machine learning and digital communication, 1998.

[Gen96] D. Genoud, F. Bimbot, G. Gravier, and G. Chollet. Combining methods to improve speaker verification decision. In *Proc. International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1756–1759, Philadelphia, PA, 1996.

[Ghi95] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. Query by humming : Musical information retrieval in an audio database. In *ACM Multimedia*, pages 231–236, 1995.

[Gir99] A. Girgensohn and J. Foote. Video classification using transform coefficients. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1999.

[Goh01] K-S. Goh, E. Chang, and Kwang-Ting Cheng. Svm binary classifier ensembles for image classification. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2001.

[Gop98] R.A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

[Got96] M. Goto and Y. Muraoka. Beat tracking based on multiple agent architecture a real time beat tracking system for audio signal. In *International Conference on Multi Agent Systems (ICMAS)*, pages 103–110, 1996.

[GP96] R. Zabih G. Pass. Histogram refinement for content-based image retrieval. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV 96)*, 1996.

[Gri04] B. Grilheres, S. Brunessaux, and P. Leray. Probabilistic combination of text classifiers using reliability indicators : Models and results. In *RIAO*, 2004.

[Gru92] T. Gruber. Theory of bibliographic-data. 1992.

[Gus01] F. Gustafsson. Segmentation of signals using piecewise constant linear regression models, 2001.

[Han01] D.J. Hand and K. Yu. Idiot's bayes - not so stupid after all ? *International Statistical Review*, 69(3) :385–398, 2001.

[Hat91] J.P. Haton, J.M. Pierrel, G. Perennou, J. Caelen, and J.L. Gauvain. *Reconnaissance de la parole*. Dunod éd., Paris, 1991.

[Hua95] Y.S. Huang and C.Y. Suen. A method of combining multipe experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1) :90–94, 1995.

[Hua98] J. Huang, Z. Liu, and Y. Wang. Integration of audio and visual information for content-based video segmentation. In *International Conference on Image Processing*, page 526 – 530, 1998.

[Hyv99] A. Hyvaarinen. Survey on independent component analysis. *Neural Computing Surveys*, 2(94-128), 1999.

[Idr97] F. Idris and S. Panchanathan. Review of image and video indexing techniques. *Journal of Visual Communication and Image Representation*, 8(2) :146–166, 1997.

[imd] The internet movie database, www.imdb.com.

[Int91] N. Intrator. *Feature extraction using an exploratory projection pursuit neural network*. PhD thesis, Brown Univ., Providence, RI, 1991.

[Iur01] U. Iurgel, R. Meermeier, S. Eickeler, and G. Rigoll. New approaches to audio-visual segmentation of tv news for automatic topic retrieval. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

[Iye00] G. Iyengar and C. Neti. Speaker change detection using joint audio-visual statistics. In *RIAO*, 2000.

[Jai95] R. Jain, R. Kasturi, , and B.G. Schunck. *Machine Vision*. McGraw-Hill, 1995.

[Jia00] H. Jiang, T. Lin, and H.J. Zhang. Video segmentation with the assistance of audio content analysis. In *International Conference on Multimedia and Expo (ICME)*, pages 1507–1510, 2000.

[Joi04] M. Joint, P.A. Moellic, P. Hède, and P. Adam. Piria : A general tool for indexing, search and retrieval of multimedia content. In *SPIE Storage and Retrieval for Multimedia Databases*, 2004.

[Jos00] P. Josserand. Detection de transitions à l'intérieur d'une séquence video en vue de son indexation. Master's thesis, Université du Littoral de Calais, 2000.

[Jou97] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9) :853–858, 1997.

[Kan00] M. Kankahalli and T. Chua. Video modeling using strata-based annotation. *IEEE Multimedia*, 7(1) :68–74, 2000.

[Kis96] C.O. Kiselman. Regularity properties of distance transformations in image analysis. In *Computer Vision and Image Understanding : CVIU*, 1996.

[Kit98] J. Kittler, M. Hatef, R.P.W.Duin, and J.Matas. On combining classifiers. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 20(3), 1998.

[Kou03] A.Z. Kouzani. Locating human faces within images. In *Computer Vision and Image understanding*, volume 91, pages 247–279, 2003.

[Kur99] Mikko Kurimo. Indexing audio documents by using latent semantic analysis and som. In Erkki Oja and Samuel Kaski, editors, *Kohonen Maps*, pages 363–374. Elsevier, 1999. IDIAP-RR 99-13.

[Kyp04] M. Kyperountas, Z. Cernekova, C. Kotropoulos, M. Gavrielides, and I. Pitas. Scene change detection using audiovisual clues. In *Norvegian Conference on Image Processing and Pattern Recognition*, 2004.

[Lee97] J.H. Lee. Analyses of multiple evidence combination. In *Research and Development in Information Retrieval*, 1997.

[Lee01] T. Berners Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5) :34–43, May 2001.

[Li97] J. Li, T. Ozsu, and D. Szafron. Modeling of moving objects in a video database. In *IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, pages 336–343, 1997.

[Li00] Stan. Z. Li. Content-based classification and retrieval of audio using the nearest feature line method. *IEEE Transactions on Speech and Audio Processing*, 2000.

[Li02] Y. Li and Jay Kuo C. Extracting movie scenes based on multimodal information. In *SPIE Proc. on Storage and Retrieval for Media Databases 2002 (EI2002)*, pages 383–394, 2002.

[Lie01] R. Lienhart. Reliable transition detection in videos : a survey and practioner guide. *International Journal of Image and Graphics*, 2001.

[Lin03] C.-Y. Lin. Ibm research trecvid-2003 video retrieval system. In *Proceedings of (VIDEO) TREC 2003 (Twelfth Text Retrieval Conference)*, 2003.

[Liu94] L.J. Liu, J.F. Lu, J.Y. Yang, K. Liu, Y.G. Wu, , and S.J. Li. Efficient segmentation of nuclei in different color spaces. In *SPIE Storage and Retrieval for Multimedia Databases*, pages 773–778, 1994.

[Liu98] Zhu Liu, Yao Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene segmentation and classification. *J. VLSI Signal Process. Syst.*, 20(1-2) :61–79, 1998.

[Liu99] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In *AVPA*, 1999.

[Liu01] L. Liu, H. Jiang, and H. Zhang. A robust audio classification and segmentation method. In *9. th. ACM Int. Conf. on Multimedia*, pages 203–211, 2001.

[Loo02] M. Loog and R.P.W. Duin. Non-iterative heteroscedastic linear dimension reduction for two-class data from fisher to chernoff, 2002.

[Loz98] R. Lozano and H. Martin. Querying virtual videos using path and temporal expressions. In *ACM symposium on Applied Computing*, 1998.

[LS55] C. Lévy-Straus. *L'analyse structurale du mythe*. Plon, 1955.

[Luc01] L. Lucchese and S.K. Mitra. Color image segmentation : A state-of-the-art survey. In *"Image Processing, Vision, and Pattern Recognition", Proc. of the Indian National Science Academy (INSA-A)*, volume 67 A, pages 207–221, 2001.

[Luo01] J. Luo and A. Savakis. Indoor vs outdoor classification of consumer photograph using low-level and semantic features. In *Int. Conf. Image Proc. ICIP01*, 2001.

[Ma99] Wei-Ying Ma and B. S. Manjunath. Netra : A toolbox for navigating large image databases. *Multimedia Systems*, 7-3 :184–198, 1999.

[Mac97] D.J.C. MacKay. Introduction to gaussian processes. Technical report, Cambridge University, 1997.

[Mac03] W. MacLean, A. Jepson, and R. Frecker. Recovery of egomotion and segmentation of independent object motion using the em algorithm. In *Proc. 5th British Machine Vision Conference*, pages 175–184, 2003.

[Mar98] K. Martin and Y. Kim. 2pmu9. instrument identification : a pattern-recognition approach. In *136th Meet. Ac. Soc. of America*, 1998.

[Mar99] J. Marques and P. Moreno. A study of musical instrument classification using gaussian mixture models and support vector machines. Technical Report TR/DCC-99-3, Cambridge, US, 1999.

[Mar00] M. Markkula and E. Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1 :259–285, 2000.

[MC99] I. Magrin-Chagnolleau and G. Durou. Time-frequency principal components of speech : application to speaker identification. In *IEEE Transactions European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 2, pages 759–762, 1999.

[McL04] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Paperback, 2004.

[McN96] R. McNab, L.A. Smith, I.H. Witten, C.L. Henderson, and S. Cunningham. Towards the digital music library : Tune retrieval from acoustic input. In *Digital Libraries*. Bethesda (MD, USA), 96.

[Mei01] S. Meignier, J-F. Bonastre, and S. Igounet. E-hmm approach for learning and adapting sound models for speaker indexing. *Chania, Crète*, 2001.

[Met68] Christian Metz. *Essais sur la signification au cinéma*. Klincksieck, 1968.

[Mil04] C. Millet. Génération de sémantiques spatiales de différents niveaux. Master's thesis, Commissariat à l'Energie Atomique. Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue, 2004.

[Mor00] P.J. Moreno and R. Rifkin. Using the fisher kernel method for web audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[Nag91] A. Nagasaka and Y. Tanaka. Automatic scene-change detection method for video works. In *2ndWorking Conference on Visual Database Systems*, pages 119–133, 1991.

[Nap00] M.R. Naphade, I. Kozintsev, and T. Huang. Probabilistic semantic video indexing. In *Neural Information Processing Systems (NIPS)*, 2000.

[Net00] C. Neti, G. Potamianos, J. Leuttin, I. Matthews, H. Glotin, D. Vergyri, J. Sisson, and A. Mashari. Audio-visual speech recognition. Technical report, CLSP Summer Workshop, 2000.

[Nit97] S. Nitin. Situational awareness from environmental sounds. Technical report, MIT Media Lab, 1997.

[Ore97] M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet template. In *Computer vision and pattern recognition*, 1997.

[Pap94] T.N. Pappas. An adaptive clustering algorithm for image segmentation. *IEEE Trans. on Signal Processing*, SP-40(4) :901–914, 1994.

[Pat96] N.V. Patel and I.K. Sethi. Audio characterization for video indexing. In *SPIE on Storage and Retrieval for Still Image and Video*, 1996.

[Pee00] G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of mpeg-7. In *ICMC : International Computer Music Conference, Berlin*, 2000.

[Pee02a] G. Peeters, A. La Burthe, and X. Rodet. Toward automatic music audio summary generation from signals analysis. In *International Conference on Music Information Retrieval (ISMIR)*, 2002.

[Pee02b] G. Peeters and X. Rodet. Automatically selecting signal descriptors for sound classification. In *International Computer Music Conference (ICMC)*, Goteborg (Sweden), 2002.

[Pol97] G. De Poli and P. Prandoni. Sonological model for timbre characterisation. *Journal of New Music Research*, 2, 1997.

[Qué99] G. Quénot and P. Mulhem. Two systems for temporal video segmentation. In *CBMIí99,Toulouse, France, October*, pages 187–193, 1999.

[Rab78] Rabiner and Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[Rab93] Rabiner and Juang. *Fundamental of Speech Recognition*. Prentice-Hall, 1993.

[Rao48] C.R. Rao. The utilization of multiple measurements in problems of biological classification. *J. Royal Statistical Soc.*, 10(B) :159–203, 1948.

[Rau91] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition : Recommendations for practitioners. *IEEE Trans. Pattern Analysis Machine Intelligence*, 13 :252–264, 1991.

[Rau01] Andreas Rauber and Markus Fruhwirth. Automatically analyzing and organizing music archives. *Lecture Notes in Computer Science*, 2163 :402–405, 2001.

[Rec99] A. Rector, P. Zanstra, W. Solomon, J. Rogers, R. Baud, W. Ceusters, W. Classen, J. Kirby, J. Ridrigues, A. Mori, E. Haring, , and J. Wagner. Reconciling users needs and formal requirements : Issues in developing a re-usable ontology for medicine. *IEEE Transactions on Information Technology in BioMedicine*, 1999.

[Ros76] E. Rosch. Classification d'objet du monde réel : origines et représentations dans la cognition. *Bulletin de Psychologie*, 242-250, 1976.

[Ros82] A. Rosenfeld and A. Kak. *Digital Picture Processing*. Academic Press, 1982.

[Ros00] S. Rossignol. *Segmentation et indexation des signaux sonores musicaux*. PhD thesis, Université de Paris VI, 2000.

[Row98] H.A. Rowley, S. Baluja, , and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1) :23–38, 1998.

[Sal01] P. Salembier and J. Smith. Mpeg-7 multimedia description schemes. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6) :748–759, 2001.

[Sap97] G. Sapiro. Color snakes. *Computer Vision and Image Understanding*, 68(2) :247–253, Nov. 1997.

[Sar97] C. Saraceno and R. Leonardi. Audio as support to scene change detection and characterization of video sequences. In *International Conference on Acoustics Speech and Signal Processing*, pages 2597–2600, 1997.

[Sau02] B. Le Saux and N. Boujemaa. Unsupervised robust clustering for image database categorization. In *IEEE-IAPR International Conference on Pattern Recognition*, 2002.

[Sch19] A. Schopenhauer. *Le monde comme volonté et comme représentation*. puf, 1819.

[Sch97] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, volume 2, 1997.

[Sch99] E.D. Scheirer. Sound scene segmentation by dynamic detection of correlogram comodulation. In *CASA*, 1999.

[Sch00a] E. Scheirer. *Music-Listening Systems*. PhD thesis, MIT Media Laboratory, 2000.

[Sch00b] E.D. Scheirer, R.B. watson, and B.L. Vercoe. On the perceived complexity of short musical segments. In *International Conference on Music Perception and Cognition*, 2000.

[Sch00c] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *In IEEE Computer Vision and Pattern Recognition*, 2000.

[Sco92] D. W. Scott. *Multivariate Density Evaluation*. New York : Wiley, 1992.

[Sla02] M. Slanney. Semantic audio retrieval. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[Sla03] M. Slaney, D. Ponceleon, and J. Kaufman. Understanding the semantic of media. *Video Mining. Kluwer*, pages 225–258, 2003.

[Sme01] A.F. Smeaton, P. Over, and R. Taban. The trec-2001 video track report. In *The Tenth TextRetrieval Conference (TREC 2001)*, 2001.

[Soh99] J. Sohn, N.S. Kim, and W. Sung. A statistical model-based voice activity detection. In *IEEE Signal Processing*, 1999.

[Sub98] S.R. Subramanya and A. Youssef. Wavelet-based indexing of audio data in audio/multimedia databases. In *4th Int'l. Workshop on Multimedia DBMS*, 1998.

[Sun00a] H. Sundaram and S. Chang. Audio scene segmentation using multiple features, models and time scales. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[Sun00b] H. Sundaram and S-F. Chang. Determining computable scene in film and their structure using audio-visual memory model. In *ACM Multimedia*, 2000.

[Sun00c] H. Sundaram and S.-F. Chang. Video scene segmentation using video and audio features. In *International Conference on Multimedia and Expo (ICME)*, pages 1145–1148, 2000.

[Tza99] G. Tzanekis and P. Cook. Multifeature audio segmentation for browsing and annotation. In *IEEE Workshop on Applications of Signal Processing to Acoustic and Audio*, 1999.

[Tza01] G. Tzanetakis, G. Essl, and P. Cook. Audio analysis using the discrete wavelet transform. In *WSES Int. Conf. Acoustics and Music : Theory and Applications (AMTA 2001)*, Skiathos, Greece,, 2001.

[Vai99] A. Vailaya, M. Figueiredo, A. Jain, and H.J. Zhang. Content-based hierarchical classification of vacation images. In *IEEE Multimedia Conference*, 1999.

[Vio02] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision - to appear*, 2002.

[Vla02] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, N. Koudas M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas. Non-linear dimensionality reduction techniques for classification and visualization. In *eighth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 645 – 651, 2002.

[Wac96] H.D. Wactlar, T. Kanade, M.A. Smith, and S.M. Stevens. Intelligent access to digital video : Informedia project. *IEEE Computer, Vol.29, No.3, pp.46-52*, 29(5) :46–52, 1996.

[Wac00] H.D. Wactlar, A.G. Hauptmann, M.G. Christel, R.A. Houghton, and A .M. Olligschlaeger. Complementary video and audio analysis for broadcast news archives. *Communications of the ACM*, 32(2) :42–47, February 2000.

[Wan00] Y. Wang, Z. Liu, and J-C. Huang. Multimedia content analysis, using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6) :12–36, 2000.

[War99] P.C.H. Wariyapola, S.L. Abrams, A.R. Robinson, K. Streitlien, N.M. Patrikalakis, P. Elisseeff, and H. Schmidt. Ontology and metadata creation for poseidon distributed coastal zone management system. In *Advances in Digital Libraries*, 1999.

[Whi03] B. Whitman. Semantic rank reduction of music audio. In *IEEE Workshop on Applications of Signal Processing to Acoustic and Audio*, 2003.

[Wol96] E. Wold. Content-based classification, search, and retrieval of audio. *IEEE Multimedia Magazine*, 3(3) :27–36, 1996.

[Xu92] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transaction on Systems, Man, And Cybernetics*, 22(3) :418–435, 1992.

[Yos01] A. Yoshitaka and M. Miyake. Scene detection by audio-visual features. In *International Conference on Multimedia and Expo (ICME)*, pages 49–52, 2001.

[Zha93] H. Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1) :10–28, 1993.

[Zha95] H.J. Zhang, S.Y. Tan, S.W. Smoliar, and G.Y. Hone. Automatic parsing and indexing of news video. *Multimedia Systems*, 2(66) :246–266, 1995.

[Zha98] T. Zhang and C. Kuo. Hierarchical system for contentbased audio classification and retrieval, 1998.

[Zha01] T. Zhang and C. Kuo. Content-based classification and retrieval of audio. In *SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations*, 2001.

[Zha03] D. Zhang and G. Lu. Evaluation of similarity measurement for image retrieval. In *IEEE International Conference on Neural Networks & Signal Processing*, 2003.