METHODES DE RESUME DE VIDEO A PARTIR D'INFORMATIONS BAS NIVEAU, DU MOUVEMENT DE CAMERA OU DE L'ATTENTION VISUELLE

Mickael Guironnet

UNIVERSITÉ JOSEPH FOURIER – GRENOBLE 1

$N°$ attribué par la bibliothèque

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER – GRENOBLE 1

Spécialité : «Signal, Image, Parole et Télécoms»

préparée au Laboratoire des Images et des Signaux de Grenoble

dans le cadre de **l'École Doctorale «Électronique, Électrotechnique, Automatique, Télécommunications et Signal»**

présentée et soutenue publiquement

par

Mickaël GUIRONNET

né le 7 mars 1979, à Tournon (Ardèche)

le 12 octobre 2006

Titre :

MÉTHODES DE RÉSUMÉ DE VIDÉO À PARTIR D'INFORMATIONS BAS NIVEAU, DU MOUVEMENT DE CAMÉRA OU DE L'ATTENTION VISUELLE

Directeurs de thèse : Denis PELLERIN et Patricia LADRET

JURY

Madame	M. ROMBAUT,	Présidente
Monsieur	O. COLOT,	Rapporteur
Monsieur	B. MERIALDO,	Rapporteur
Monsieur	P. TARROUX,	Examinateur
Monsieur	D. PELLERIN,	Directeur de thèse
Madame	P. LADRET,	Co-directrice de thèse

Remerciements

Je remercie tout d'abord mes encadrants, Denis et Patricia qui m'ont guidé tout au long de ma thèse par leur rigueur scientifique et leurs conseils. J'ai aussi apprécié la confiance qu'ils m'ont porté en m'accordant une grande autonomie à la réalisation du projet.

Je remercie également les membres du jury :
- Michèle Rombaut pour avoir présidé ma soutenance.
- Olivier Colot et Bernard Mérialdo pour avoir expertisé mon manuscrit de thèse et pour leurs remarques pertinentes qui ont contribué à la qualité du manuscrit.
- Philippe Tarroux pour avoir examiné ma thèse.

Je tiens aussi à remercier mes compagnons de thèse, Zakia, Corentin, Pierre et Emmanuel pour m'avoir accompagné durant ces années de thèse au LIS. Une pensée particulière à Nathalie pour ses nombreux conseils et son soutien. Un grand merci à Michèle pour sa disponibilité et ses conseils qui ont contribué à l'avancée de ma thèse. Ma reconnaissance va également à Georges Quénot et Christian Marendaz pour leur collaboration. Je souhaite enfin remercier l'ensemble du personnel, les thésards pour la bonne ambiance au laboratoire.

Je n'oublie pas mes amis qui m'ont permis de me changer les idées. Merci à Mathilde, Mickael, Karen, Hugo, Audrey, Jérôme, Florian, Sébastien, Alex et Olivier... pour les bonnes soirées passées ensemble. J'ai également une pensée pour le club de football de Coux qui a contribué à me divertir ainsi que l'ensemble des joueurs avec lesquels j'ai passé de bons moments.

Les remerciements vont évidemment à ma famille, en particulier mes parents, mon frère et ma belle-famille pour leur soutien et leur encouragement. Merci à Nadège et Brigitte pour avoir eu la gentillesse et « le courage » de relire le manuscrit.

Pour clore ces remerciements, je voudrais remercier profondément ma copine Marion qui m'a accompagné et m'a épaulé dans les moments difficiles. Merci de ta patience, de ton réconfort et de tes précieuses aides sans lesquelles je ne pourrais écrire ces remerciements que je dédie à toutes les personnes de mon entourage.

Table des matières

Chapitre 1

Introduction

Durant cette dernière décennie, la technologie « numérique » a révolutionné les moyens de communication avec l'arrivée entre autres des téléphones portables, de l'internet à haut débit et de la télévision numérique. L'explosion des moyens de communication a conduit à une augmentation spectaculaire des informations audiovisuelles. Dans ce contexte, il est apparu nécessaire de mettre à disposition des outils permettant de retrouver rapidement l'information désirée parmi tous les documents multimédia.

Dans le cadre de cette thèse, nous nous intéressons à décrire les vidéos afin de faciliter la recherche d'informations dans une base de données vidéo. Une manipulation aisée et une visualisation rapide du contenu des documents sont des critères essentiels auxquels répond le résumé de vidéo. L'objectif de cette thèse est donc de proposer des méthodes de résumé de vidéo pour fournir un aperçu rapide à l'usager et faciliter l'accès aux informations recherchées. Le résumé est une version courte de la vidéo qui doit contenir l'essentiel de l'information, tout en étant le plus concis possible.

La création automatique de résumé nécessite l'étude du contenu des vidéos pour être efficace. L'analyse du contenu des vidéos, domaine de recherche en pleine expansion, consiste à extraire des informations pertinentes issues de différents canaux (image, son et texte) afin de décrire judicieusement les documents multimédia. Les traitements peuvent être effectués séparément ou conjointement sur les différents canaux. Nous avons choisi de nous consacrer aux informations visuelles des vidéos. L'extraction de caractéristiques vise à indexer de manière efficace les vidéos et à créer des résumés pertinents. L'intérêt du résumé de vidéo réside aussi dans la grande variété des applications qui en découlent comme la recherche, la classification et la navigation dans des bases de vidéos.

1.1 Notre contribution

Dans cette thèse, nous avons étudié trois nouvelles méthodes de résumé de vidéo. La première s'appuie sur des caractéristiques de bas niveau alors que la deuxième et la troisième font appel à des index de plus haut niveau, le mouvement de caméra et l'attention visuelle. Pour cela, nous avons été amené à proposer un modèle de classification des mouvements de caméra et un modèle d'attention visuelle. Nous avons également abordé des applications du résumé de vidéo : la recherche par l'exemple, la classification et la navigation dans les vidéos.

Notre travail a permis d'apporter les contributions suivantes :

Nous avons tout d'abord conçu une méthode de résumé de vidéo à partir d'index bas niveau. De nouveaux descripteurs compacts et flous ont été créés et sont basés sur des caractéristiques de bas niveau : couleur, orientation et mouvement. Ils ont été utilisés pour segmenter la vidéo en unités homogènes et ainsi former le premier niveau du résumé. Puis, nous avons élaboré un algorithme de regroupement par similarité avec contrainte temporelle pour fournir un résumé hiérarchique afin de faciliter la navigation. Une étude sur la manière d'associer les index a été plus particulièrement menée. Celle-ci a conduit à une combinaison s'appuyant sur la logique floue. Nous avons montré l'apport de la combinaison sur une application du résumé, la recherche par l'exemple.

Nous avons ensuite développé une méthode de résumé de vidéo à partir du mouvement de caméra. En effet, le mouvement de caméra pensé par le réalisateur transmet un message et donc induit une information sur le contenu. En premier lieu, une méthode a donc été conçue pour extraire les mouvements de caméra dans chaque plan de la vidéo. Elle consiste, à partir d'une estimation paramétrique du mouvement, à combiner les paramètres selon le Modèle des Croyances Transférables pour identifier et décrire les mouvements de caméra. La méthode de résumé repose quant à elle sur l'amplitude et l'enchaînement des mouvements de caméra. Afin de juger de la qualité des résumés, nous avons conçu une méthode d'évaluation. Une expérience a ainsi été mise en place pour créer le résumé d'une vidéo afin de le comparer au résumé automatique créé par notre méthode. Pour obtenir une méthode de résumé automatique, nous avons aussi élaboré une méthode de détection des changements de plan fondée sur le Modèle des Croyances Transférables.

Nous avons également créé une méthode de résumé à partir de l'attention visuelle. Cette caractéristique de plus haut niveau qui permet de déterminer les régions où le regard est attiré, nous est apparue pertinente pour résumer des vidéos. Nous avons donc élaboré un modèle spatio-temporel d'attention visuelle que nous avons ensuite employé pour créer une méthode de résumé.

Une autre contribution de la thèse a été le développement d'une méthode de classification des vidéos. Elle consiste à annoter des extraits de vidéo suivant différents concepts (par exemple, bateau, plage, basketball, ...). L'originalité de la méthode réside dans la combinaison des sorties de différents classifieurs qui repose sur le Modèle des Croyances Transférables. Cette étude a été menée dans le cadre des expérimentations de TREC Video 2004.

1.2 Organisation de la thèse

Le coeur de cette thèse concerne la création de résumé de vidéo ainsi que quelques-unes des applications du résumé. La thèse s'organise de la façon suivante.

Dans le chapitre 2, nous dressons un état de l'art sur les différentes méthodes de création de résumé. Après une description de la structure de la vidéo, nous étudions les méthodes existantes suivant les deux grandes familles de résumé : résumé statique, issu d'une sélection d'images représentatives (images clés) et résumé dynamique, résultant d'une sélection de segments extraits de la vidéo, équivalent à une bande annonce.

La méthode de création de résumé à partir d'index bas niveau est présentée dans le chapitre 3. Pour obtenir un résumé de taille variable selon l'attente de l'utilisateur, nous proposons une méthode de résumé hiérarchique. Elle repose sur une micro-segmentation de la vidéo où chaque segment obtenu est homogène suivant un ou plusieurs descripteurs. Un algorithme de regroupement par similarité avec contrainte temporelle est ensuite conçu pour réunir les segments homogènes afin d'aboutir à une macro-segmentation. Notre méthode s'appuie sur de nouveaux descripteurs flous et notre principale contribution concerne la combinaison des descripteurs suivant la théorie de la logique floue. Enfin, une application à la recherche par l'exemple est étudiée pour vérifier la performance de notre méthode.

Dans le chapitre 4, nous cherchons à identifier les mouvements de caméra dans la vidéo, caractéristique pertinente pour analyser le contenu des vidéos. Nous développons une méthode de classification des mouvements de caméra qui s'appuie sur le Modèle des Croyances Transférables. Elle consiste à estimer le mouvement dominant entre paires d'images, puis à analyser le mouvement de manière plus globale sur un ensemble d'images. L'identification et la quantification des mouvements de caméra sont à la base d'une méthode de création de résumé de vidéo dans le chapitre 5. La sélection des images clés dépend de l'amplitude et de l'enchaînement des mouvements dans les plans de la vidéo. Pour mesurer la qualité des résumés, nous avons également proposé une méthode d'évaluation. Elle repose sur une expérience effectuée par des sujets humains pour construire le résumé afin de le comparer à celui de notre méthode. Par ailleurs, comme cette méthode de résumé nécessite la connaissance des plans, nous avons développé une méthode de détection des changements de plan décrite dans le chapitre 6. Elle s'appuie sur la combinaison de descripteurs selon le Modèle des Croyances Transférables pour détecter les différentes transitions dans les vidéos.

Dans le chapitre 7, nous décrivons la méthode de création de résumé de vidéo à partir de l'attention visuelle. Pour extraire les régions saillantes dans les images, un modèle spatio-temporel d'attention visuelle est construit et une expérience psychophysique est mise en place pour évaluer le modèle. Une méthode de résumé est ensuite proposée à partir de cette information.

Nous étudions dans le chapitre 8 une application du résumé, la classification des vidéos. Dans un contexte où les données audiovisuelles sont conséquentes (TREC Video 2004), nous développons un système de classification qui repose sur un apprentissage d'une Machine à Vecteurs de Support (SVM) et une combinaison des classifieurs SVM en utilisant le Modèle des Croyances Transférables.

Nous résumons dans le chapitre 9 les principales contributions de cette thèse puis nous proposons des améliorations possibles et des perspectives pour la poursuite de ce travail.

Chapitre 2

Etat de l'art sur les résumés de vidéos

Ce chapitre présente un état de l'art sur les méthodes de création de résumé de vidéo. Après une description de la structure des vidéos, nous citons les principales méthodes que constituent les deux grandes familles de résumé de vidéo : résumé statique, issu d'une sélection d'images représentatives (images clés) et résumé dynamique, résultant d'une sélection de segments extraits de la vidéo, équivalent à une bande annonce. Les approches présentées ci-après n'ont pas la prétention d'être exhaustives mais représentatives.

2.1 Structure des vidéos

Une vidéo se compose d'images affichées à une fréquence de 25 images (ou 30 images) par seconde, accompagnées d'une bande son. Dans le cadre de cette thèse, le son n'a pas été pris en compte et seule l'information apportée par les images a été étudiée. Suivant le regroupement des images, différentes entités peuvent être repérées. Mais, comme signalé dans [Ven02], le vocabulaire les décrivant est souvent ambigü. Cela traduit la difficulté à laquelle sont confrontés les auteurs pour représenter une vidéo. Dans la suite du manuscrit, nous allons considérer différents niveaux de segmentation.

Le plan, souvent considéré comme l'unité de base des vidéos, se définit comme une portion de vidéo filmée continûment sans effets spéciaux ni coupure. A partir du plan, différents niveaux de segmentation ont été proposés. Tout d'abord, les images adjacentes à l'intérieur de chaque plan sont regroupées suivant une caractéristique commune (par exemple, si elles ont un même mouvement de caméra) pour former une *micro-segmentation*, premier niveau de la segmentation. Le *plan* correspond au niveau suivant de la segmentation. Enfin le dernier niveau de la segmentation consiste à réunir des plans ou des micro-segments pouvant provenir de plans différents pour établir une *macro-segmentation*. Néanmoins, cette notion, qui diffère suivant les auteurs, demande un critère pour pouvoir regrouper des micro-segments ou des plans. Le critère ici choisi est la similarité des contenus visuels. Un autre critère aurait pu être le regroupement de plans contenant un même évènement. Parfois, la notion de *scène* est abordée mais là encore les définitions varient selon les auteurs. Dans [Oh04], la scène se compose de plans adjacents qui ont une similarité sémantique en objets, en personnes dans l'espace et dans le temps alors que la scène correspond pour d'autres travaux [Ngo03] à un regroupement de plans qui ont une même unité de lieu. La figure 2.1 illustre la structure des vidéos. L'extraction automatique de la structure d'une vidéo n'est pas une tâche facile et la difficulté augmente avec les niveaux de segmentation. En effet, les micro-segments ou les plans

peuvent être déterminés à partir du signal de la vidéo alors que les macro-segments sont plus difficiles à obtenir car ils dépendent du contenu sémantique des plans. Les entités définies pour structurer une vidéo seront manipulées dans les méthodes de création de résumé de vidéo.

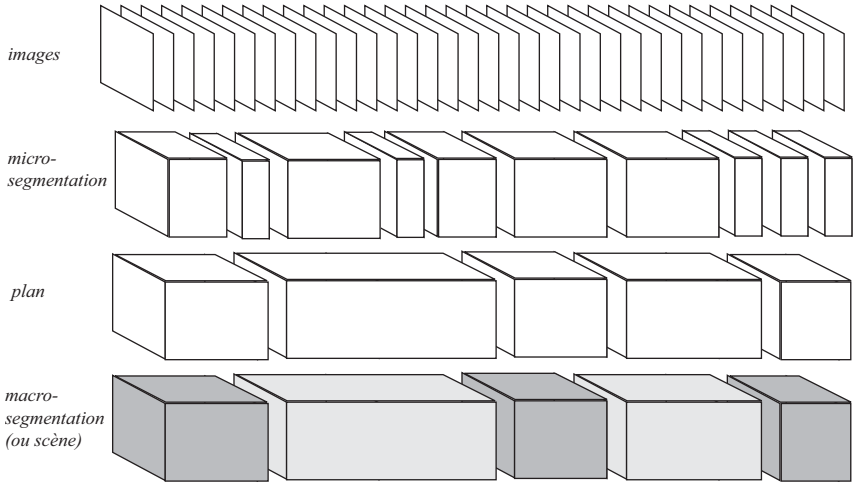


FIG. 2.1 – Structure des vidéos. Au niveau de la macro-segmentation, les parties de même couleur sont regroupées en macro-segments.

De plus, certains auteurs [Rui04] distinguent deux types de vidéo : les vidéos ayant un script (scripted content) et les vidéos sans script (unscripted content). Pour le premier type, il s'agit de vidéos qui possèdent une structure bien définie comme les journaux télévisés et les films. En revanche, pour le deuxième type, aucun script n'est écrit comme dans les vidéos de sport ou la surveillance vidéo. Nous avons, dans la suite de cette thèse, étudié principalement des vidéos possédant un script.

2.2 Résumé de vidéo

Devant le volume grandissant des données audiovisuelles, la construction automatique de résumé de vidéo est devenue un domaine de recherche en pleine expansion. Le résumé de vidéo a pour objectif de fournir des informations pertinentes et concises afin d'aider l'utilisateur à naviguer ou à organiser ses fichiers vidéos plus efficacement. Deux sortes de résumé peuvent être retrouvées dans la littérature [Li01, Li04] que nous appellerons en français : *résumé statique* (video summary) et *résumé dynamique* (video skimming). Le résumé statique consiste à sélectionner les images les plus représentatives de la vidéo. Ces images appelées *images clés* se présentent en général sous la forme d'un scénarimage (storyboard). Le résumé dynamique se compose d'extraits de la vidéo et correspond à une version courte de la vidéo originale. Cette technique est similaire à la construction automatique d'une bande-annonce donnant un aperçu du film.

2.2.1 Résumé statique

Le résumé statique de vidéo se compose d'un ensemble d'images qui représente le contenu de la vidéo. Beaucoup de travaux ont été réalisés ces dernières années et quatre familles de résumé statique ont été dégagées dans [Li01] : méthodes reposant sur l'échantillonnage, sur

les plans, sur les scènes et autres. Nous avons repris ces quatre familles de résumé que nous avons complétées avec des travaux plus récents.

2.2.1.1 Méthodes basées sur l'échantillonnage

Les premiers travaux [Mil92, Tan95] concernant les résumés de vidéos consistent à choisir les images clés en sous-échantillonnant uniformément ou aléatoirement la séquence originale. L'inconvénient des méthodes qui n'étudient pas le contenu, est la non représentation de certaines parties de la vidéo et la possible redondance de certaines images clés avec des contenus similaires.

2.2.1.2 Méthodes basées sur les plans

Des travaux plus élaborés tentent d'extraire des images clés en s'adaptant au contenu de la vidéo. La détection des plans est réalisée pour mieux ajuster la sélection des images clés au contenu de la vidéo. De nombreuses méthodes existent pour détecter les plans et une description détaillée de celles-ci peut être trouvée au chapitre 6. Une façon simple pour représenter les différents plans de la vidéo est d'extraire la première image du plan comme image clé [Nag91, Ton93] ou les première et dernière images du plan [Ued91]. Ces méthodes semblent efficaces pour décrire les plans stationnaires où le contenu varie peu. En revanche, elles ne fournissent pas une représentation satisfaisante pour les plans avec de forts mouvements de caméra.

D'autres travaux choisissent alors de représenter le contenu des vidéos en employant des caractéristiques visuelles de bas niveau comme la couleur, le mouvement ou la texture. Dans [Zha95, Gun97], le nombre d'images clés dépend du contenu des plans. La première image du plan est sélectionnée comme image clé. Puis, si la distance entre l'histogramme couleur de la dernière image clé sélectionnée et l'image courante est supérieure à un seuil alors une nouvelle image clé est choisie. La sélection de la première image n'est pas forcément judicieuse puisqu'elle peut être soumise aux effets de transition (fondu enchaîné) entre les plans. L'approche décrite dans [Zhu98] compare les images d'un plan suivant leurs histogrammes couleur puis réalise le rassemblement des images en plusieurs groupes. Seuls les groupes de taille assez importante sont conservés et les images les plus proches du centre de gravité de chaque groupe sont alors choisies comme images clés. Dans [Zha00], l'histogramme couleur est utilisé comme vecteur caractéristique pour représenter les images et la courbe dans cet espace caractéristique est considérée. L'approximation par des morceaux de droites conduit à la sélection des images clés. Choisir l'histogramme couleur a l'intérêt d'être invariant aux orientations, robuste aux bruits et rapide à extraire. Cependant, il n'est pas insensible aux mouvements de caméra.

Pour construire un résumé de vidéo, des approches exploitent le mouvement, caractéristique représentant directement la dynamique du plan. Wolf [Wol96] propose une méthode d'extraction basée sur le mouvement. Le flot optique entre chaque paire d'images du plan est estimé et permet de définir une mesure sur le mouvement. Les images clés sont sélectionnées à chaque minimum local sur la courbe traduisant le mouvement. Dans [Rav05], une image mosaïque est construite à partir de l'estimation de mouvement de caméra et elle est utilisée pour décrire le contenu des plans dynamiques. Cette approche fournit seulement une vue statique de l'arrière-plan et ne représente pas le déplacement des objets à l'intérieur du plan. De plus, la construction de l'image mosaïque dépend de l'estimation du mouvement de caméra et reste

peu efficace en présence de mouvements complexes.

Des travaux ont également étudié différentes caractéristiques de bas niveau. Dans [Dou00], plusieurs descripteurs (histogrammes couleur et mouvement) sont extraits pour chaque image du plan. Les vecteurs caractéristiques des différents descripteurs sont alors réunis pour créer un unique vecteur pour chaque image et former une courbe temporelle dans l'espace des caractéristiques. Les images clés sont obtenues en sélectionnant des points appropriés sur la trajectoire pour la caractériser. Une méthode similaire est discutée dans [Cio06, Cio05], où une courbe est définie et caractérise des différences d'images suivant trois descripteurs (histogrammes couleur, histogrammes des orientations et analyse par ondelettes pour décrire la texture).

L'inconvénient de travailler au niveau des plans est que le nombre d'images clés peut être trop important pour représenter la vidéo.

2.2.1.3 Méthodes basées sur les scènes ou macro-segments

Des travaux ont été réalisés en ne considérant pas comme unité de la vidéo le plan mais en définissant des unités de plus haut niveau. Suivant les auteurs, différents niveaux de hiérarchie sont étudiés pour créer le résumé de vidéo. Par exemple, le regroupement de plans par similarité peut être considéré comme un niveau plus élevé que la segmentation en plans et aura pour conséquence de sélectionner moins d'images clés.

Des méthodes de résumé ont été conçues pour décrire seulement les plans les plus représentatifs de la vidéo. L'approche décrite dans [Uch99] consiste d'abord à assigner un groupe à chaque image, puis à réunir les deux groupes les plus similaires de manière itérative. L'algorithme s'arrête quand la similarité entre les groupes est inférieure à un seuil. Pour chaque groupe créé, les images adjacentes sont déterminées et constituent un segment. Une mesure d'importance est alors calculée pour chaque segment suivant sa longueur et sa rareté. Tous les segments de petite taille sont alors supprimés et seuls les segments ayant une mesure d'importance suffisante sont conservés. Puis, l'image la plus proche du centre de gravité de chaque segment est extraite comme image clé. Le résumé est ensuite présenté en ajustant la taille des images clés suivant leur importance. Un algorithme pour organiser les images clés selon leur taille a été proposé dans [Gir03] afin d'optimiser l'affichage du résumé. Sun et al. [Sun00a] ont proposé une méthode de résumé où la vidéo est divisée uniformément en segments. A chaque segment est associée une mesure de changement qui est égale à la distance entre les première et dernière images. Les différentes mesures sont alors ordonnées puis séparées en 2 groupes : un groupe de petit changement et un groupe de grand changement. Pour le groupe de petit changement, les première et dernière images du segment sont conservées alors que pour le groupe de grand changement, toutes les images sont gardées. Finalement si le nombre d'images clés désiré est atteint, l'algorithme s'arrête sinon les images conservées constituent une nouvelle vidéo et celle-ci est à nouveau divisée en segments jusqu'à atteindre le nombre désiré d'images clés. Gong et al. [Gon00] utilisent une décomposition en valeur singulière (SVD) pour créer le résumé de vidéo. Ils créent une matrice où chaque colonne contient un vecteur caractéristique associé à chaque image de la vidéo. Une SVD est ensuite réalisée pour réduire l'espace des caractéristiques et elle est à l'origine d'une mesure sur le changement visuel d'un groupe d'images. Ils l'emploient dans un algorithme de regroupement des vecteurs caractéristiques dans l'espace réduit afin de former des groupes d'images avec des contenus similaires.

Finalement, les plans les plus longs de chaque groupe sont conservés et l'image la plus proche du centre du groupe est choisie comme image clé. Dans [Kop04], une mesure de pertinence est associée à chaque plan suivant différentes caractéristiques (détection des visages, identification d'objets en mouvement,...) et les plans sont sélectionnés dans l'ordre décroissant de cette mesure jusqu'à atteindre la taille du résumé souhaitée.

L'inconvénient des approches qui travaillent sur la totalité de la vidéo est qu'elles peuvent s'avérer très contraignantes par le temps de calcul et la mémoire sollicitée.

Pour s'adapter à la demande des utilisateurs, des méthodes ont été élaborées pour créer un résumé de vidéo hiérarchique. Par exemple, Ferman et al. [Fer03] proposent un résumé hiérarchique avec deux niveaux de résolution. Le premier est réalisé en appliquant un algorithme de regroupement (fuzzy c-means) pour chaque plan de la vidéo afin d'extraire un ensemble d'images clés non redondantes qui représente les différents plans de la vidéo. Le second consiste à calculer une matrice de similarité entre les images clés et à regrouper les similaires suivant le nombre d'images demandé par l'utilisateur. De manière générale, différents algorithmes de regroupement par similarité ont été utilisés et parfois adaptés pour créer un résumé hiérarchique (algorithme des k-moyens [Zho96, Far02], algorithme « fuzzy c-means » [Yu04], algorithme de classification hiérarchique ascendant [Ane04, Ben06] ou algorithme de l'arbre couvrant de poids minimum ou « Minimum Spanning Tree » [Kir02, Che03b])

L'inconvénient de ces approches est de générer un résumé hiérarchique qui ne prend pas en compte la position temporelle des images.

Des travaux ont suggéré de créer le résumé hiérarchique en tenant compte de la composante temporelle. La mesure de similarité dépend alors du contenu et de la position temporelle. Dans [Rui99], chaque plan est représenté par les première et dernière images. Puis un algorithme de regroupement par similarité avec contrainte temporelle est appliqué pour construire des groupes et obtenir les scènes de la vidéo. La représentation hiérarchique de la vidéo a également été étudiée dans [Ven02]. Des mesures de similarité avec contrainte temporelle déjà proposées dans la littérature sont rappelées et de nouvelles sont suggérées. Elles sont ensuite utilisées dans l'algorithme de classification hiérarchique ascendant pour construire le résumé. Dans [Zhu04], une structure est proposée pour regrouper des plans similaires en super-groupes. Puis, des groupes sont créés en réunissant les plans proches temporellement de chaque super-groupe afin de fournir un résumé hiérarchique de vidéo.

Un des avantages des approches avec différents niveaux est la possibilité de contrôler le niveau de détail à atteindre dans la sélection des images clés en offrant une décomposition pyramidale de la vidéo. La prise en compte de la composante temporelle permet au résumé d'être plus cohérent puisque le regroupement d'images ne peut avoir lieu que si les images sont proches temporellement.

2.2.1.4 Autres méthodes

Des méthodes de résumé ont été proposées et s'appuient sur diverses informations (image, son, texte ou annotation manuelle). La méthode décrite dans [Yu03] permet la création de résumé qui dépend des demandes de l'utilisateur. A partir de caractéristiques extraites automatiquement (détection d'objets et de visages) ainsi que des annotations manuelles, les images de la vidéo qui répondent à la requête de l'utilisateur sont retournées pour le résumé. Parsin et al. [Par04] ont développé un système qui permet de spécifier les préférences de l'utilisateur

concernant le contenu désiré du résumé en imposant des contraintes sur les caractéristiques des segments de vidéo. De multiples caractéristiques au niveau des pixels et sémantiques, extraites de la vidéo et de la bande sonore, sont combinées afin de sélectionner les segments de la vidéo pour le résumé. Sundaram et al. [Sun00b] proposent la segmentation en scènes de la vidéo en combinant des informations visuelles et auditives. Dans [Tok00], une méthode de résumé est proposée et repose sur les sous-titres de la vidéo pour la segmenter en sujets.

Le problème du résumé a parfois été traité en le formulant sous une forme mathématique. Yahiaoui et al. [Yah01a] ont proposé l'expérience suivante. Un sujet voit un résumé d'une vidéo suivi d'un extrait de la vidéo. Il doit alors dire si l'extrait appartient ou non à la vidéo. La réponse attendue est positive si au moins une image de l'extrait est similaire au résumé. En revanche, s'il a un doute, aucune réponse est fournie. Un sujet virtuel est alors considéré pour simuler le comportement d'un véritable sujet. Cette expérience virtuelle est alors traduite sous la forme d'une mesure mathématique de performance du résumé afin de le construire en optimisant cette mesure. Dans [Li05], une métrique est proposée pour quantifier la distorsion entre un résumé et la vidéo originale. Celle-ci est optimisée pour fournir le résumé avec un minimum de distorsion.

Lorsque les vidéos proviennent d'une même série, la méthode décrite dans [Yah01b] permet la création de résumés de meilleure qualité. En effet, si les résumés sont créés indépendamment les uns des autres alors des informations redondantes peuvent apparaître entre les résumés. Les similarités et les différences entre les vidéos sont donc identifiées pour une meilleure représentation du contenu de chaque vidéo.

2.2.2 Résumé dynamique

Comparé au résumé statique, le résumé dynamique conserve les propriétés dynamiques de la vidéo et par conséquent il est plus agréable à regarder que le résumé statique. De plus l'information audiovisuelle est préservée et donc fournit une représentation plus proche de la vidéo originale. Dans [Li01], deux types de résumé dynamique sont distingués : résumé suivant des extraits clés (highlight) et résumé donnant une vue d'ensemble de la vidéo (summary sequence). Ce dernier est utilisé pour donner à l'utilisateur une impression globale du contenu de la vidéo alors que le résumé par extraits clés contient seulement les passages les plus intéressants de la vidéo, comme une bande annonce d'un film qui montre les scènes les plus attrayantes sans révéler la fin de l'histoire.

La création de résumé suivant des extraits clés (highlight) est une tâche difficile sans connaissance *a priori* sur la nature de la vidéo. Dans [Pfe96], une méthode a été conçue pour retenir certaines scènes importantes de la vidéo (scènes contenant des objets ou des personnes, scènes d'action, les scènes où un dialogue est reconnu, suppression des scènes à la fin du film...). Toutes les scènes qui ont été sélectionnées forment la bande annonce du film. Cependant, la création de résumé suivant des extraits clés (highlight) repose en général sur des hypothèses trop élémentaires pour être véritablement efficaces. Une amélioration possible pour détecter les événements importants est la recherche des explosions, des fusillades ou des gros plans qui peuvent être considérés comme des extraits clé.

Beaucoup d'approches essaient de résoudre le problème du résumé suivant des extraits clés en exploitant des connaissances *a priori* sur le type de la vidéo, comme les vidéos de sport ou des vidéos avec des caractéristiques spécifiques. Pour les vidéos de sport, les travaux sont basés sur la détection d'événements comme dans des vidéos de football [Eki03], de baseball [Rui00] ou de basketball [Zha01]. Par exemple, dans [Sun03], un modèle de chaînes de Markov, entraîné

par un algorithme EM (Espérance-Maximisation), est utilisé pour détecter les phases de jeu et de pause dans une vidéo de football. Ils développent également une technique pour détecter les lancers du baseball en construisant une hiérarchie suivant les phases du lancer. Le résumé dynamique est alors conçu en sélectionnant les phases importantes. Pour les journaux télévisés, un système est également proposé dans [Yan03].

La plupart des approches qui élaborent un résumé dynamique (ou statique) reposent sur des caractéristiques de bas niveau. Elles ne garantissent pas que les résumés créés contiennent tout le contenu de la vidéo, et donc que ces résumés aient un sens pour l'utilisateur. Des informations sémantiques ont alors besoin d'être ajoutées pour améliorer la performance du résumé. Mais peu de travaux ont été réalisés pour annoter automatiquement le contenu des vidéos dans un cadre général. Pour collecter ce genre d'informations, le contenu de la vidéo est parfois annoté manuellement. Une méthode sur l'annotation semi-automatique peut être trouvée dans [Wu04]. Néanmoins, les résumés suivant des extraits clés restent un procédé subjectif dont une extraction automatique et efficace est encore loin d'être réalisée.

La plupart des recherches sur le résumé dynamique se focalisent sur le résumé donnant une vue d'ensemble de la vidéo (summary sequence). Dans [Nam99], la vidéo est découpée en sous-plans où un index sur l'information de mouvement est calculé. La sélection d'images clés s'effectue sur ces sous-plans et le résumé dynamique est ensuite créé par interpolation des images clés. Le système « CueVideo » décrit dans [Pon99] consiste à changer la vitesse de lecture de la vidéo. Il délivre des passages plus rapides en présence de scènes statiques et moins rapides pour les scènes dynamiques. Néanmoins, ce type d'approches ne permet pas d'avoir des compressions élevées de la vidéo.

Comme la compréhension automatique du contenu est encore loin d'être atteinte, des approches essaient de diminuer la redondance visuelle de la vidéo. Une métrique basée sur l'entropie est définie dans [Gon01] pour mesurer la redondance du contenu de la vidéo et un résumé est obtenu en optimisant cette mesure. L'approche développée dans [Lu04, Lu05] consiste à créer un résumé basé sur un graphe orienté. Ils construisent un graphe orienté où les noeuds sont les plans et les arrêtes symbolisent la distance entre les plans. Le chemin le plus long forme alors le résumé dynamique. Dans [Ma02], les auteurs essaient de modéliser l'attention portée par un observateur. Plusieurs caractéristiques sont extraites comme le contraste, le mouvement, le son et le texte, elles sont utilisées pour définir des mesures d'attention. Elles sont ensuite combinées pour obtenir une courbe temporelle d'attention. Le résumé dynamique est alors créé en sélectionnant les portions de la vidéo où la courbe d'attention est maximale. Un résumé statique peut également être obtenu et une méthode [Ma05b] a été développée pour optimiser la visualisation du résumé. Ngo et al. [Ngo03] construisent une méthode de résumé qui repose sur des graphes pour remonter aux scènes de la vidéo et sur des mesures d'attention pour sélectionner dans les scènes, les sous-plans appropriés pour le résumé.

Des travaux ont été effectués sur le résumé dynamique en s'appuyant sur des informations sémantiques. Un système est discuté dans [Nag02] pour annoter facilement une vidéo et ensuite créer un résumé de vidéo. Leur outil d'annotation contient une transcription de la voix, une description des scènes et des objets ainsi qu'une segmentation automatique de la vidéo. Ils emploient une mesure d'importance des mots et des phrases dans la vidéo ainsi que la détection des scènes pour créer le résumé. En effet, la transcription comportant des mots et des phrases importantes permet de fournir une collection de scènes significatives pour décrire la vidéo.

Par la suite, nous allons nous intéresser uniquement au résumé statique pour les raisons

suivantes. Nous avons constaté que les méthodes les plus sophistiquées sur le résumé dynamique reposent en général sur la transcription de la bande son. Par ailleurs, les méthodes qui utilisent des caractéristiques de bas niveau et qui n'ont pas de connaissance *a priori* sur la vidéo réalisent souvent un résumé dynamique à partir d'un résumé statique. Comme nous travaillons uniquement avec le contenu visuel des vidéos et que nous n'avons aucune connaissance *a priori* sur les vidéos, nous allons étudier le résumé statique. De plus, les applications comme l'indexation, la recherche et la navigation dans les vidéos sont principalement obtenues à partir des résumés statiques. Dans le manuscrit, le terme « résumé » fera donc référence à un résumé statique.

2.3 Passage des caractéristiques de bas niveau à une description sémantique

Pour représenter correctement le contenu d'une vidéo et faciliter la création du résumé, les caractéristiques de haut niveau devraient être utilisées. Pourtant, la plupart des travaux actuels choisissent des critères de bas niveau pour décrire les vidéos. L'extraction automatique d'index sémantiques est encore loin d'être réalisée et seuls les travaux qui ont une connaissance *a priori* sur le contenu de la vidéo sont susceptibles de fournir une description de plus haut niveau sémantique. Comme déjà signalé, la détection d'événements est par exemple réalisée sur des séquences de sports [Rui00, Zha01] ou des journaux télévisés [Yan03]. En l'absence d'information *a priori* sur la vidéo, le contenu des images est en général représenté par des descripteurs de bas niveau pouvant traduire des informations de couleur, de mouvement ou de texture.

La majorité des travaux qui proposent des résumés de vidéo caractérise le contenu de l'image par un histogramme couleur [Uch99, Gon00, Fer03, Che03b, Lu04]. Néanmoins, il existe de nombreuses variantes pour le choix de l'histogramme couleur qui dépend de l'espace couleur utilisé et de la manière dont les bins sont construits. Par exemple, des travaux [Yu03, Ben06] créent un histogramme couleur directement à partir des données compressées MPEG. Comme observé auparavant, l'avantage de l'histogramme couleur est qu'il est robuste aux bruits et insensible aux orientations.

Cependant des travaux ont été menés par le groupe MPEG pour concevoir des descripteurs efficaces et compacts permettant la recherche de similarité dans une base de vidéos. La norme MPEG7 [Cal02] est un standard pour la description du contenu multimédia et permet de rechercher un contenu ou de naviguer dans la vidéo plus efficacement que les systèmes à base de mots clés. Cette description fournit des informations sur la vidéo (titre, réalisateur, acteurs...), des informations sémantiques (qui, quoi, quand, où) et des informations de bas niveau (histogramme couleur, mesure de l'activité...). Ce standard permet de décrire les caractéristiques du contenu multimédia mais ne normalise pas leur extraction et ne spécifie pas le système de recherche.

Le groupe MPEG a donc défini différents descripteurs de couleur, de mouvement et de texture. Dans [Man01], quatre descripteurs couleur sont répertoriés, dont chacun possède des propriétés spécifiques : descripteur SCD (Scalable Color Descriptor), descripteur CSH (Color Structure Histogram), descripteur DCD (Dominant Color Descriptor) et descripteur CLD (Color Layout Descriptor). Par exemple, le descripteur CSH représente la structure locale de la couleur dans une image alors que le descripteur DCD permet d'obtenir les couleurs dominantes sur une partie de l'image ou sur toute l'image. De la même manière, trois descripteurs de tex-

ture [Man01] sont considérés dans MPEG7 : descripteur TBD(Texture Browsing Descriptor), descripteur HTD (Homogeneous Texture Descriptor) et le descripteur LEHD (Local Edge Histogram Descriptor). Par exemple, le descripteur TBD caractérise la régularité et la direction de la texture. Des descripteurs de mouvement ont également été définis dans MPEG7 [Jea01] : descripteur d'activité de mouvement (Motion Activity), descripteur du mouvement de caméra (Camera Motion), descripteur WPD (Warping Parameter Descriptor), descripteur de la trajectoire du mouvement (Motion Trajectory), descripteur du mouvement paramétrique (Parametric Motion).

Le nombre croissant de descripteurs peut nécessiter la fusion des distances ou des similarités entre les descripteurs pour améliorer l'analyse du contenu des vidéos et donc créer un résumé de meilleure qualité ou retrouver plus facilement des extraits de vidéo par une requête. Peu de travaux ont étudié la combinaison des index dans le cadre du résumé et celle-ci repose généralement sur des considérations simples comme la combinaison linéaire des distances [Zhu03c, Zhu03b, Par04] ou la multiplication des différentes distances entre elles [Cio05].

La première méthode de résumé que nous allons présenter s'appuie sur des caractéristiques de bas niveau (couleur, orientation et mouvement). Nous étudierons plus particulièrement la manière de les combiner pour créer le résumé. En revanche, les deux autres méthodes de résumé que nous allons exposer reposent sur des caractéristiques de plus haut niveau (mouvement de caméra et attention visuelle).

Chapitre 3

Création de résumé de vidéo à partir d'index bas niveau

Nous allons, dans ce chapitre, présenter une méthode de résumé hiérarchique de vidéo à partir d'index bas niveau. Trois nouveaux descripteurs flous et compacts sont extraits (couleur, orientation et mouvement) et sont utilisés pour construire le résumé. Nous avons plus particulièrement étudié la manière de les combiner pour définir une mesure de similarité entre images. La méthode de résumé que nous proposons utilise une combinaison entre les index qui repose sur la logique floue.

Sommaire

3.1 Introduction

Dans ce chapitre, nous allons décrire une méthode de résumé de vidéo qui pourra être exploitée pour retrouver rapidement un extrait dans une vidéo ou naviguer facilement dans une base de vidéos. L'objectif du résumé est d'extraire les passages importants de la vidéo pour pouvoir les présenter à un utilisateur et donc lui fournir un aperçu rapide. La création d'un résumé demande de caractériser le contenu sémantique de la vidéo. Néanmoins l'extraction automatique du contenu sémantique est encore loin d'être réalisée malgré les avancées dans le traitement des images et dans les algorithmes de reconnaissance. Nous allons ainsi utiliser des caractéristiques de bas niveau pour analyser les vidéos et en générer des résumés.

Un problème important dans la conception des résumés est le choix des caractéristiques de bas niveau qui représentent le contenu des images. Comme signalé dans le chapitre 2, de nombreux descripteurs ont été développés dans la littérature. Cependant, il est souvent difficile de leur donner un sens au regard des descripteurs. Nous avons choisi d'extraire de nouveaux descripteurs (couleur, orientation et mouvement) qui soient facilement interprétables. De plus, ils présentent l'avantage d'être compacts et reposent sur les ensembles flous, adaptés pour représenter des données imprécises.

Les méthodes de résumé de vidéo sont en général conçues à partir de l'extraction d'un seul descripteur. Or, l'utilisation de plusieurs descripteurs est peu étudiée dans les méthodes de résumé et repose souvent sur une combinaison linéaire entre les distances des descripteurs. Nous allons donc construire une combinaison, s'appuyant sur la logique floue, pour comparer le contenu des images selon différents descripteurs. La logique floue permet de modéliser le « langage naturel » et de représenter les connaissances que nous avons sur le système à étudier. Son usage dans les systèmes experts d'aide à la décision a permis, à partir des observations et des règles d'inférence floues, d'aboutir à une conclusion sur l'état du système. Le premier système expert flou a été introduit par Mamdani en 1975 [Mam75] et a été développé pour contrôler un moteur à vapeur. Dans les procédés de classification [Liu02], un système flou peut également être mis en place pour décrire, sous forme de règles, les opérations effectuées par un expert humain. Nous allons présenter une méthode de résumé de vidéo qui utilise les systèmes flous pour comparer les similarités entre plusieurs descripteurs.

Naturellement, une des contraintes principales d'un résumé est qu'il doit posséder une taille aussi petite que possible tout en conservant le maximum d'information. Afin de répondre à cette exigence, notre méthode de résumé est conçue de manière hiérarchique et se décompose en deux étapes : micro-segmentation et macro-segmentation. A partir des descripteurs extraits, la vidéo est partitionnée en segments homogènes suivant une mesure de similarité donnée pour former le niveau le plus fin du résumé et obtenir une micro-segmentation de la vidéo. Puis les segments proches temporellement sont regroupés pour réduire leur nombre. Un algorithme de regroupement par similarité avec contrainte temporelle a été mis au point pour créer les différents niveaux de résolution du résumé et produire une macro-segmentation de la vidéo. L'algorithme que nous présentons est une adaptation de l'algorithme de classification hiérarchique ascendant (CHA). L'avantage d'un résumé hiérarchique est que la granularité du

résumé peut être adaptée en fonction de la demande de l'utilisateur. Une application telle que la recherche par l'exemple va permettre de vérifier l'efficacité de la méthode de résumé de vidéo. Elle consiste à effectuer une requête et à rechercher dans des vidéos des segments similaires.

Nous allons présenter successivement les différents descripteurs (couleur, orientation et mouvement), la méthode de résumé et l'application à la recherche par l'exemple.

3.2 Extraction des descripteurs

Beaucoup de travaux ont été réalisés sur l'extraction de descripteurs afin de représenter le contenu des vidéos. Comme déjà signalé (Section 2.3), ceux-ci sont souvent des descripteurs de bas niveau auxquels il est souvent difficile de donner une interprétation. Il s'agit alors, de développer des descripteurs qui permettent d'exprimer les caractéristiques de bas niveau des images. Bien que la norme MPEG7 propose des descripteurs, nous avons choisi de créer de nouveaux descripteurs flous qui peuvent être interprétés facilement. Par exemple, le descripteur mouvement va informer sur le degré d'activité d'une paire d'images ou le descripteur couleur va révéler la proportion de quelques couleurs présentes dans l'image. De plus, les descripteurs que nous avons proposés sont compacts afin de simplifier la représentation des images et faciliter leur stockage.

3.2.1 Couleur

La couleur est peut-être la caractéristique visuelle la plus utilisée pour représenter les images et elle est souvent décrite par l'intermédiaire d'un histogramme couleur. Nous souhaitons définir un histogramme comportant peu de bins dont chacun caractérise une couleur donnée.

Un espace couleur doit alors être choisi pour construire l'histogramme. Il existe de nombreux espaces couleur (RGB, HSV,...) qui ne présentent pas les mêmes propriétés. L'espace couleur RGB (Red, Green, Blue, ou en français RVB) décrit les couleurs à partir de la combinaison de trois couleurs primaires : rouge, vert et bleu. Néanmoins, la perception des couleurs dans l'espace RGB n'est pas uniforme et dépend des conditions d'éclairage. L'espace HSV (Hue, Saturation, Value) lui est souvent préféré parce qu'il traduit directement les notions intuitives des couleurs. La transformation de l'espace RGB à l'espace HSV est non-linéaire mais inversible. La teinte H qui varie de 0 à 2π représente la nuance de couleur et où celle-ci se situe dans le spectre des couleurs. La saturation S décrit la pureté de la couleur et varie de 0 à 1. Enfin la valeur V correspond à l'intensité lumineuse de la couleur et elle est représentée par des valeurs de 0 à 255. L'espace HSV est souvent quantifié uniformément suivant chacune de ces composantes pour réduire le nombre de couleurs dans l'image. L'inconvénient de ce procédé est qu'il attribue le même poids aux pixels proches du centre d'un bin de ceux qui sont situés aux bords du même bin. L'utilisation des ensembles flous peut résoudre ce genre de problème en associant aux pixels un degré d'appartenance suivant chaque bin considéré.

De plus, Sural et al. [Sur02] ont observé que la saturation détermine la transition entre les couleurs et les niveaux de gris. Lorsque S vaut zéro et V augmente, le pixel va du noir au blanc en passant par tous les niveaux de gris. En revanche, si la saturation S augmente pour une valeur L et une teinte H données, le pixel passe du niveau de gris à la couleur pure indiquée par H. Par conséquent, pour de faibles valeurs de S, le pixel peut être représenté par

un niveau de gris (V) tandis que pour des valeurs élevées de S, le pixel peut être apparenté à la teinte (H).

Suite à ces observations, nous avons cherché à associer à chaque pixel soit une couleur soit un niveau de gris. Nous avons ainsi visualisé les couleurs dans le plan saturation-luminosité (Saturation-Value) et nous avons défini empiriquement trois zones : zone couleur, zone niveau de gris et zone intermédiaire (Fig. 3.1). Les courbes y_1 et y_2 qui permettent de les distinguer sont données par l'équation 3.1. Chaque pixel de l'image est donc projeté dans le plan saturation-luminosité et associé à une de ces trois zones.

Si le pixel projeté appartient à la zone intermédiaire, nous déterminons un degré d'appartenance à la zone couleur et un à la zone niveau de gris afin de lui attribuer une pondération entre ces deux zones (couleur et niveau de gris). Les distances entre le pixel projeté et les courbes y_1 et y_2 sont calculées. On note d_1 (resp d_2) la distance minimum entre (v, s) et y_1 (resp (v, s) et y_2). A partir de l'équation 3.2, le degré d'appartenance w_1 à la zone couleur et celui w_2 à la zone niveau de gris sont obtenus en calculant l'inverse des distances d_1 et d_2. Dans la figure 3.1, le pixel projeté (v, s) appartient à la zone intermédiaire, la distance d_1 (resp d_2) correspond à la distance entre (v, s) et (v_1, s_1) (resp (v, s) et (v_2, s_2)) et les degrés d'appartenance obtenus sont $w_1 = 0.44$ et $w_2 = 0.56$.

$$\begin{aligned} y_1 &= \begin{cases} 21.25 \cdot (v - 0.3)^2 + 0.15 & \text{si } v < 0.3 \\ 0.15 & \text{si } v \geq 0.3 \end{cases} \\ y_2 &= \begin{cases} 18.75 \cdot (v - 0.5)^2 + 0.25 & \text{si } v < 0.5 \\ 0.25 & \text{si } v \geq 0.5 \end{cases} \end{aligned} \tag{3.1}$$

$$w_1 = \frac{\frac{1}{d_1}}{\frac{1}{d_1} + \frac{1}{d_2}} \quad \text{et} \quad w_2 = \frac{\frac{1}{d_2}}{\frac{1}{d_1} + \frac{1}{d_2}} \tag{3.2}$$

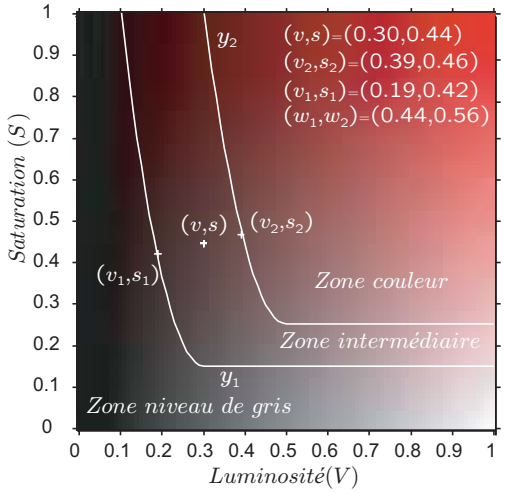


FIG. 3.1 – Visualisation dans le plan saturation-luminosité des trois zones définies : zone couleur, zone niveau de gris et zone intermédiaire. Chaque pixel projeté est associé à une de ces trois zones.

Si le pixel se situe dans la zone couleur, il est associé à une de ces 6 couleurs : rouge, jaune,

vert, cyan, bleu, magenta. Les fonctions d'appartenance de ces 6 couleurs ont été fixées de manière empirique à partir de la distribution des couleurs, comme montré dans la figure 3.2.

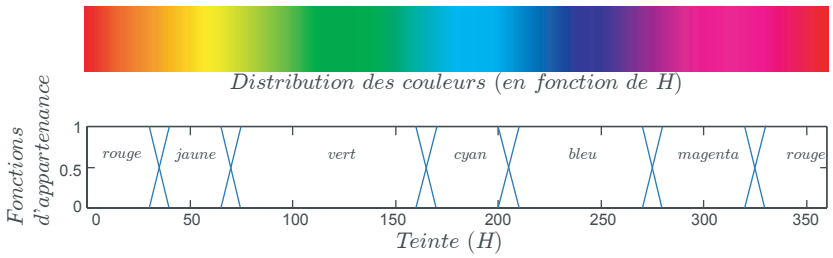


FIG. 3.2 – Fonctions d'appartenance des 6 couleurs : rouge, jaune, vert, cyan, bleu, magenta.

En revanche, si le pixel est dans la zone niveau de gris, il est associé à un de ces 5 niveaux : noir, gris-foncé, gris, gris-clair et blanc, selon les fonctions d'appartenance de la figure 3.3.

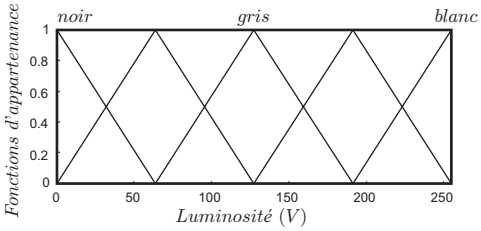


FIG. 3.3 – Fonctions d'appartenance des 5 niveaux de gris : noir, gris-foncé, gris, gris-clair et blanc.

Enfin, si le pixel se situe dans la zone intermédiaire, il est supposé appartenir à la fois aux zones couleur et niveau de gris. Simplement les degrés d'appartenance sont pondérés par w_1 pour la zone niveau de gris et par w_2 pour la zone couleur. Finalement, chaque pixel est représenté par 11 bins (6 pour la couleur et 5 pour les niveaux de gris). Si le pixel n'est pas dans la zone intermédiaire, il y a 2 degrés d'appartenance au maximum non nuls. En revanche, s'il se trouve dans la zone intermédiaire, 4 degrés d'appartenance au maximum sont différents de zéro sur les 11 bins. La figure 3.4 donne un exemple de descripteurs couleur. Chaque pixel est associé à la couleur dont le bin est maximum.

En résumé, chaque pixel est quantifié sur 11 bins (6 couleurs et 5 niveaux de gris) et la somme sur ces bins vaut un. Nous calculons l'histogramme couleur de l'image sur ces 11 bins. Il est ensuite normalisé par la taille de l'image et décrit le contenu de l'image en terme de couleur. L'avantage de ce descripteur est que les 11 bins ont un sens. En effet, chaque bin représente une couleur ou un niveau de gris.

3.2.2 Orientation

Les orientations dans les images [Agh03] ont également été très étudiées. Cet index apparaît être particulièrement performant pour distinguer les scènes (par exemple des scènes d'intérieur/extérieur). Dans [Guy02], des filtres de Gabor sont employés pour classer les scènes naturelles. Dans [Wan04], les orientations extraites par un détecteur de Canny sont utilisées pour afficher automatiquement l'image dans le bon sens.

FIG. 3.4 – Exemple de descripteurs couleur. En haut : Images de la séquence « The Avengers ». En bas : Descripteurs couleur où chaque pixel est associé à la couleur dont le bin est maximum.

Nous souhaitons définir un descripteur d'orientation des contours. Similairement aux travaux de Kosecka et al. [Kos03], nous réalisons l'extraction des contours par un simple calcul de gradient en chaque pixel de l'image. Seuls les pixels ayant la norme du gradient supérieure à un seuil donné δ sont considérés appartenir aux contours de l'image. Les orientations des gradients de ces pixels sont ensuite déterminées et permettent de construire un histogramme flou. Dans notre expérimentation, nous utilisons 5 bins pour représenter les orientations et le seuil δ est fixé à 10. La figure 3.5 montre les fonctions d'appartenance des bins pour 0°, 90°, 180° et 270°. Le dernier bin représente la quantité de pixels qui n'appartiennent pas à un contour. Finalement, l'histogramme est normalisé par la taille de l'image et décrit les orientations des contours de l'image. Celui-ci permet de renseigner sur la présence d'horizontales et/ou de verticales dans l'image.

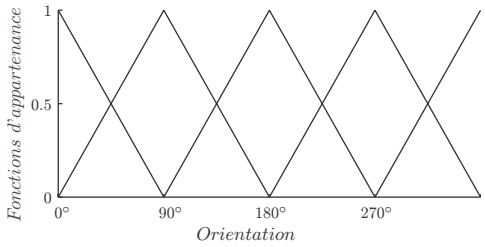


FIG. 3.5 – Fonctions d'appartenance des 4 orientations : 0 °, 90 °, 180 ° et 270 °.

La figure 3.6 montre un exemple de 2 images où les contours de l'image sont utilisés pour déterminer l'histogramme des orientations.

3.2.3 Mouvement

Pour décrire le contenu dynamique des vidéos, il est essentiel d'étudier le mouvement dans les vidéos et plus particulièrement le degré d'activité. En général, les films d'action ont beaucoup de segments avec une activité élevée alors que les journaux TV sont plutôt caractérisés par une activité faible. C'est pourquoi nous définissons un descripteur d'activité très compact qui capture des notions intuitives sur l'intensité du mouvement. La construction

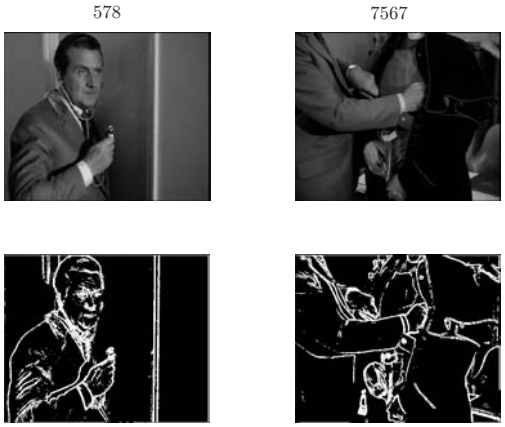


FIG. 3.6 – Exemple de descripteurs d'orientation. En haut : Images de la séquence « The Avengers ». En bas : Extraction des contours de l'image afin d'obtenir le descripteur d'orientation.

de ce descripteur repose sur l'estimation du mouvement entre deux images successives.

3.2.3.1 Estimation du flot optique

Cette section décrit une méthode d'estimation du flot optique qui conduira au nouveau descripteur d'activité. La détermination du flot optique est une étape importante et conditionnera les performances du descripteur. La méthode d'estimation que nous avons choisie [Bru01a, Bru01b] a été développée dans notre laboratoire et permet une représentation compacte et multi-échelle du mouvement entre deux images successives.

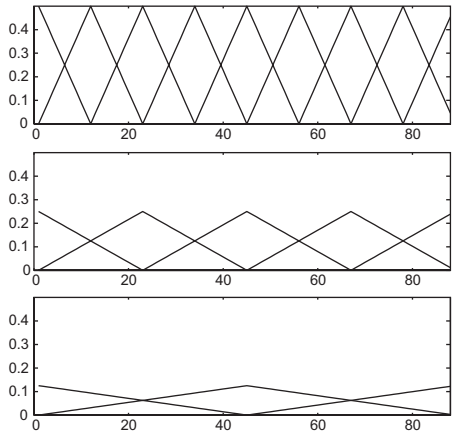
Soit $V_j(p_i, t)$ l'estimation du flot optique ou vecteur vitesse au pixel $p_i = (x_i, y_i)$ entre les images $I(p_i, t)$ et $I(p_i, t+1)$ au niveau de résolution j. Le flot optique est modélisé par une combinaison linéaire de fonctions d'échelle :

$$V_j(p_i, t) = \sum_{k1,k2=0}^{2^{j-1}} \theta_{j,k1,k2} \cdot \Phi_{j,k1,k2}(p_i) \tag{3.3}$$

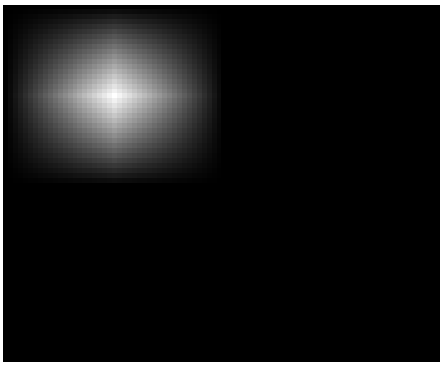
$\Phi_{j,k1,k2}$ représente la fonction d'échelle, dans notre cas, fonction B-Spline de degré 1 avec 3 niveaux de la résolution (Fig. 3.7). Les indices $j, k1, k2$ représentent respectivement le niveau d'échelle, le décalage horizontal et vertical. En supposant la conservation de luminosité entre deux images successives, l'algorithme réalise de manière itérative et robuste l'estimation des coefficients d'échelle $\theta_{j,k1,k2} = (\theta_{x,j,k1,k2}, \theta_{y,j,k1,k2})$ en minimisant la fonction objective suivante :

$$E = \sum_{p_i} \rho(I(p_i + V_j(p_i, t+1) - I(p_i, t), \sigma)) \tag{3.4}$$

La fonction $\rho(\cdot, \sigma)$ est une fonction qui pondère les données selon l'erreur (M-estimateur de Geman-McClure). La minimisation de la fonction objective (Eq. 3.4) est décrite dans [Bru01b]. De plus, connaissant la décomposition du champ vitesse à un niveau de résolution j, nous pouvons déterminer les coefficients d'échelle à un niveau de résolution plus bas c'est-à-dire inférieur à j. La connaissance fine du mouvement n'est pas nécessaire pour l'indexation de vidéos et l'estimation est réalisée sur des images sous-échantillonnées (72x88 pixels) pour



(a) Fonctions B-Spline de degré 1 avec trois niveaux de résolution (à une dimension). En haut, 9 fonctions d'échelle ($j = 2$), au milieu 5 fonctions d'échelle ($j = 1$) et en bas 3 fonctions d'échelle $j = 0$.



(b) Fonction B-Spline de degré 1 au niveau de résolution $j = 1$ (à deux dimensions) sur une image de taille 72 par 88.

FIG. 3.7 – Exemple de fonctions d'échelle B-Spline.

accélérer les traitements (temps de calcul inférieur à une seconde). La méthode offre une signature de mouvement avec 162 coefficients (81 coefficients pour chaque composante, abscisse et ordonnée).

Comme les coefficients d'échelle caractérisent le mouvement sur des régions de l'image, ils offrent une représentation locale et compacte du mouvement. La figure 3.8 montre un exemple d'estimation du mouvement. Il correspond au mouvement de caméra puisque les voitures restent au milieu de l'image et leur mouvement apparent est nul.

3.2.3.2 Descripteur d'activité

La notion d'activité est introduite pour représenter la manière dont la perception humaine capture l'intensité d'action ou le rythme dans un segment de vidéo. Par conséquent, le degré d'activité dépend de l'amplitude du mouvement et donc de l'amplitude des coefficients d'échelle. Le principe de construction du descripteur d'activité est illustré dans la figure 3.10. L'amplitude du mouvement est obtenue à partir des coefficients d'échelle selon l'équation 3.5.

$$A_{j,k1,k2} = \sqrt{{\theta_{x,j,k1,k2}}^2 + {\theta_{y,j,k1,k2}}^2} \tag{3.5}$$

L'estimation des coefficients d'échelle a été réalisée avec 81 fonctions d'échelle, ce qui représente 81 régions de l'image. Ne souhaitant pas obtenir une description de l'activité trop fine, la décomposition en ondelettes permet facilement le passage à des niveaux de résolution plus grossier. Connaissant la décomposition du flot optique au niveau de la résolution $j = 2$, les coefficients d'échelle ont été obtenus au niveau de la résolution $j = 1$. Puis, à partir des coefficients d'échelle $\theta_{x,1,k1,k2}$ et $\theta_{y,1,k1,k2}$ de chaque paire d'images, les amplitudes $A_{k1,k2} = A_{1,k1,k2}$ sont calculées et stockées sous la forme d'une grille 5 par 5 suivant les décalages

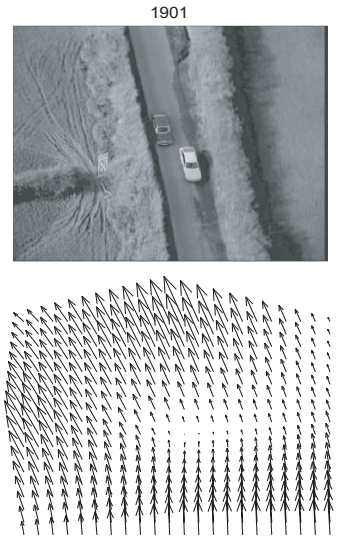


FIG. 3.8 – Exemple d'estimation du mouvement. En haut : Image de la séquence « The Avengers ». En bas : Estimation du flot optique avec un niveau de résolution $j = 2$. Le champ de vitesse correspond au mouvement de caméra et le mouvement apparent des voitures est nul puisque les voitures restent au centre de l'image.

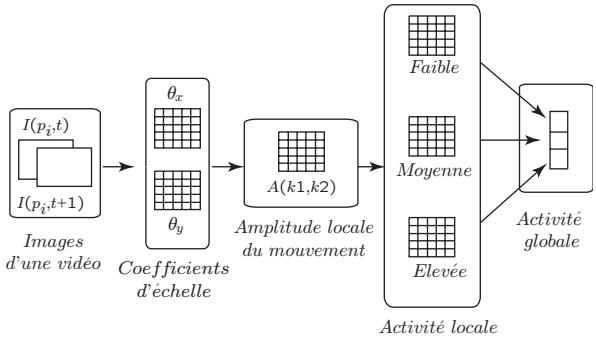


FIG. 3.9 – Principe de construction du descripteur d'activité.

horizontal $k1$ et vertical $k2$ (Fig. 3.9). Cette représentation permet de montrer où les différentes amplitudes sont localisées dans l'image.

Néanmoins, les amplitudes $A_{k1,k2}$ sont des valeurs numériques qui ne permettent pas de décrire de manière intuitive l'activité. En effet, il paraît plus intéressant de savoir si une image a une forte ou faible activité que de connaître exactement les valeurs numériques de l'amplitude. Les ensembles flous sont adaptés pour obtenir une description symbolique de l'activité. Ainsi l'amplitude $A_{k1,k2}$ est transformée en valeurs symboliques : faible, moyenne et élevée. Les fonctions d'appartenance sont représentées sur la figure 3.10.

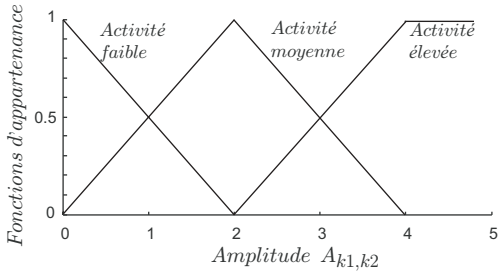


FIG. 3.10 – Fonctions d'appartenance de l'activité. Trois ensembles flous sont définis : activité faible, moyenne et élevée.

A partir de la grille des amplitudes $A_{k1,k2}$, nous obtenons trois grilles (Fig. 3.9) où chacune décrit l'activité suivant les qualificatifs symboliques : faible, moyenne et élevée.

Afin de visualiser l'activité entre deux images, nous utilisons la méthode du centre de gravité (Fig. 3.11). Elle consiste à calculer la moyenne des valeurs modales des 3 ensembles flous, pondérés par leur degré d'appartenance. Une valeur modale d'un ensemble flou est par définition une valeur x telle que le degré d'appartenance vaut $\mu(x) = 1$. Dans notre implémentation, les valeurs modales des 3 ensembles flous (activité faible, moyenne et élevée) sont respectivement de 0, 2 et 4.

A partir de la description locale, nous caractérisons l'activité globale avec seulement 3 composantes en calculant la moyenne sur chacun des trois ensembles flous (Fig. 3.10). L'ensemble ayant la plus grande moyenne informe sur le niveau d'activité entre deux images. La figure 3.11 donne un exemple de descripteurs locaux et globaux d'activité. Le blanc correspond à une activité élevée et le noir à une activité faible. Nous pouvons observer que le descripteur local d'activité représente le déplacement d'objet lorsque la caméra est fixe.

FIG. 3.11 – Exemple de descripteurs d'activité. En haut : Images de la séquence « The Avengers ». Au milieu : Descripteurs locaux d'activité. Le blanc correspond à une activité forte et le noir à une activité faible. En bas : Descripteurs globaux d'activité. Les cases de gauche, du centre et de droite représentent respectivement une activité faible, moyenne et élevée.

Finalement, la méthode d'estimation du mouvement, développée au laboratoire, peut être utilisée pour décrire l'activité de manière soit globale soit locale. La représentation locale peut être intéressante si l'activité est recherchée dans les régions de l'image. Cependant, dans l'optique de regrouper des images suivant l'activité, la description locale n'est pas adéquate. En effet, si une personne se déplace dans la scène ou si le mouvement de caméra change au cours du temps, ces images auront un descripteur d'activité local différent. C'est pourquoi nous avons choisi le descripteur global d'activité avec 3 composantes (activité faible, moyenne et élevée).

3.3 Similarité entre les descripteurs

A partir des descripteurs extraits, la différence de contenu entre les images peut être examinée et repose généralement sur la comparaison des descripteurs. Afin de caractériser la ressemblance entre les images, une mesure de similarité a été spécifiée pour chaque descripteur. Néanmoins, la similarité suivant plusieurs descripteurs est une tâche difficile qui nécessite souvent la combinaison des similarités de chacun des descripteurs. En général, la fusion des similarités s'effectue en deux étapes : normalisation et combinaison des similarités. Comme la dynamique des similarités peut être différente, l'étape de normalisation est indispensable. L'étape de combinaison est souvent réalisée de manière simple par une somme pondérée des similarités [Ang02, Ohb04]. Nous allons définir et étudier une nouvelle technique pour combiner les différents descripteurs. Celle-ci sera utilisée pour créer le résumé de vidéo en regroupant les images similaires et donc elle permettra de conserver uniquement les images les moins similaires comme images clés.

3.3.1 Similarité pour chacun des descripteurs

Une mesure de similarité pour chaque descripteur est classiquement réalisée par l'intermédiaire d'une distance. Parmi les nombreuses distances possibles, nous avons choisi la distance usuelle de Hamming pour comparer des vecteurs caractéristiques d'un même descripteur.

Soient f_i^p et f_i^q deux vecteurs caractéristiques extraits des images p et q selon un descripteur donné i. Par la manière dont les descripteurs ont été construits, la somme des composantes de chacun d'eux vaut un. Ainsi la distance de Hamming maximale entre deux vecteurs caractéristiques au sein d'un même descripteur est inférieure ou égale à 2. La distance entre les vecteurs f_i^p et f_i^q est alors définie par :

$$d_i(f_i^p, f_i^q) = \frac{1}{2} \sum_k |f_{i,k}^p - f_{i,k}^q| \tag{3.6}$$

où $f_{i,k}^p$ est la k^{th} composante du vecteur caractéristique i de l'image p et $d_i(f_i^p, f_i^q)$ est la distance entre les images p et q selon le descripteur i. La division par 2 permet simplement de recentrer la distance entre les bornes 0 et 1. Néanmoins, il est connu [San99] qu'il existe une non-linéarité entre les distances et le jugement donné par un sujet humain. Si A et B sont des représentants de deux stimuli (images) dans l'espace des caractéristiques alors $d(A, B)$ est la distance dans l'espace des caractéristiques et un sujet humain en fera un jugement du type $g(d(A, B))$ où g est une fonction croissante. Ainsi nous avons choisi de formaliser cette non-linéarité par la fonction racine carrée qui a l'avantage d'augmenter les distances élevées et donc de peu différencier les distances élevées des distances très élevées. Le passage de la

notion de distance d_i à la notion de similarité s_i s'effectue classiquement par la transformation suivante :

$$s_i(f_i^p, f_i^q) = 1 - (d_i(f_i^p, f_i^q))^{1/2} \tag{3.7}$$

Ainsi quel que soit le descripteur flou utilisé i, la similarité s_i est définie par l'équation 3.7. Il reste à définir une similarité entre images selon plusieurs descripteurs.

3.3.2 Similarité entre plusieurs descripteurs

Deux méthodes sont présentées afin de décrire les ressemblances des images suivant différents descripteurs. La première est réalisée par une simple combinaison linéaire des similarités tandis que la seconde, plus originale, repose sur la théorie de la logique floue.

3.3.2.1 Similarité classique

Nous supposons extrait un ensemble de descripteurs, noté $F^p = \{f_1^p, \ldots, f_i^p, \ldots, f_m^p\}$, où m est le nombre de descripteurs, dans notre cas $m = 3$ (couleur, orientation et mouvement) et f_i^p est un vecteur caractéristique représentant l'image d'indice p selon le descripteur i. Une mesure de similarité s peut être obtenue en combinant les similarités de plusieurs descripteurs pour représenter la ressemblance entre deux images selon les descripteurs considérés. Une méthode simple pour les combiner est de réaliser la somme pondérée des similarités de chacun des descripteurs, ce qui correspond à la combinaison linéaire des similarités suivant différents descripteurs. La similarité s entre deux vecteurs caractéristiques F^p et F^q pour les images d'indice p et q est donnée par l'équation suivante :

$$s(F^p, F^q) = \frac{1}{w} \sum_{i=1}^{m} w_i \cdot s_i(f_i^p, g_i^q) \tag{3.8}$$

où w_i est un coefficient de pondération tel que $w = \sum_i w_i$. Pour donner autant de poids à chacun des descripteurs, il faut que les coefficients w_i soient égaux. En revanche, si on souhaite privilégier la similarité d'un descripteur, celui-ci doit avoir un poids plus élevé que les autres.

L'inconvénient de cette combinaison linéaire est qu'elle ne prend pas en compte la dynamique des similarités. Bien que les similarités soient comprises entre 0 et 1, les valeurs des similarités issues des différents descripteurs ne sont pas forcément cohérentes. Considérons le cas où les similarités s_i ont la même valeur, leur combinaison s aura alors la même valeur ($\forall w_i$). Néanmoins, ce n'est pas parce que les similarités ont la même valeur que leurs similarités prises individuellement sont équivalentes. Par exemple, si la similarité s_i entre deux images est de 0.6 pour un descripteur i et si la similarité s_j entre ces deux mêmes images vaut 0.8 pour un autre descripteur j, la proximité de ces images au sens du descripteur i peut nous paraître plus proche que la proximité de ces images pour le descripteur j. Afin de prendre en compte ces remarques, une combinaison va être réalisée au sens de la logique floue. Un autre avantage de la logique floue est de pouvoir modéliser notre expertise grâce à l'introduction de règles floues. Prenons l'exemple où deux images sont très similaires pour tous les descripteurs considérés à l'exception d'un seul descripteur, alors les images peuvent être considérées comme similaires ou au contraire différentes suivant la définition des règles floues.

3.3.2.2 Similarité au sens de la logique floue

La similarité est déterminée à partir d'un système d'inférence floue. Nous allons présenter un bref résumé de cette approche, une description plus détaillée des systèmes d'inférence pourra être trouvée dans [Hak97, Che98, Blo03, Yac05].

Un système d'inférence se compose de trois phases : la fuzzification, l'inférence et la défuzzification. La fuzzification consiste à transformer les valeurs numériques en valeurs floues définies sur un espace de représentation. Soit s_i la similarité selon un descripteur i et $E = \{PS, MS, S\}$ l'espace symbolique de représentation de s_i où PS signifie « descripteurs pas similaires », MS « descripteurs moyennement similaires » et S « descripteurs similaires ». La figure 3.12 donne un exemple de représentation trapézoïdale de chacun des termes linguistiques. Chaque ensemble flou (PS, MS et S) est défini par une fonction d'appartenance (μ_{PS}, μ_{MS} et μ_S) qui décrit le degré avec lequel l'élément s_i^* appartient à cet ensemble. Par exemple, dans la figure 3.12, la fuzzification symbolique de s_i^* est donnée par $0.25/PS + 0.75/MS + 0/S$. Afin de réaliser une fuzzification de chacune des similarités, les paramètres a, b, c et d doivent être fixés (Fig. 3.12) et peuvent être différents d'un descripteur à un autre. Comme signalé précédemment, les similarités ne sont pas forcément cohérentes entre les descripteurs et l'intérêt de ces paramètres a, b, c et d est de pouvoir être ajustés en fonction de notre expertise.

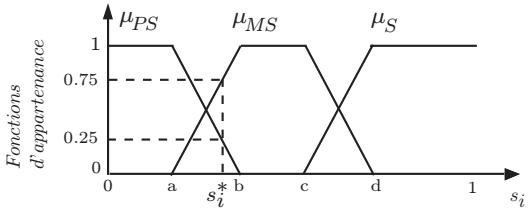


FIG. 3.12 – Exemple de fonctions d'appartenance trapézoïdales.

L'inférence est un mécanisme qui repose sur des règles floues définies généralement à partir de connaissances issues des observations. Bien qu'il existe plusieurs méthodes d'inférence, nous avons choisi celle de Mamdani [Mam75], méthode d'inférence la plus couramment utilisée pour fournir une description linguistique du système. Les règles permettent d'établir des relations entre les variables linguistiques. Par exemple, soient s_i et s_j deux entrées de similarité suivant les descripteurs i et j. Un ensemble de règles est alors défini pour les combiner et ainsi définir la similarité suivant ces deux descripteurs. Une règle possible est : si s_i est S et s_j est S alors s est S où s est la variable de sortie, c'est-à-dire la similarité suivant les descripteurs i et j. Cette règle se compose d'un prédicat (s_i est S et s_j est S) et d'une conclusion (s est S). Le prédicat (aussi appelé prémisse ou condition) et la conclusion sont des propositions floues ou des combinaisons de propositions floues. Le tableau 3.1 montre les règles floues que nous avons adoptées pour combiner les similarités s_i et s_j entre deux descripteurs i et j. Chacune d'elles possède ici un prédicat composé de deux propositions floues et d'une conclusion formée d'une proposition floue.

La *combinaison des propositions floues* nécessite la définition des opérateurs « ET » et « OU ». Nous avons employé ceux définis par Zadeh [Zad65] qui sont les plus répandus. L'opérateur « ET » (respectivement « OU ») est réalisé par le minimum (respectivement maximum) entre les degrés de vérité des propositions. Le degré de vérité d'une proposition « s_i est S » est égale à $\mu_S(s_i)$ où s_i est la similarité définie pour le descripteur i et μ_S représente la fonction d'appartenance associée à l'ensemble flou S (descripteurs similaires). L'*implication* permet

TAB. 3.1 – Exemple de règles floues pour la combinaison des similarités s_i et s_j pour les descripteurs i et j.

		s_j		
		PS	MS	S
s_i	PS	PS	PS	MS
	MS	PS	MS	S
	S	MS	S	S

ensuite de déterminer la conclusion de la règle. Elle est réalisée par le minimum entre le degré d'activation de la règle et la fonction d'appartenance de la conclusion. Le degré d'activation d'une règle représente le degré de vérité du prédicat obtenue par la combinaison des propositions floues du prédicat. Finalement les différentes règles sont liées par l'opérateur « OU ». Elle est effectuée en calculant le maximum des fonctions d'appartenance de sortie. L'*agrégation* des conclusions de chacune des règles forme l'ensemble flou de sortie de l'inférence.

La figure 3.13 montre un exemple de combinaison entre les similarités s_i^* et s_j^* de deux descripteurs. Il correspond à un système d'inférence à deux entrées (s_i^* et s_j^*) et à une sortie s, et les règles floues utilisées sont celles présentées dans le tableau 3.1. Comme la sortie est ici un ensemble flou, une étape de défuzzification est nécessaire pour obtenir une valeur réelle.

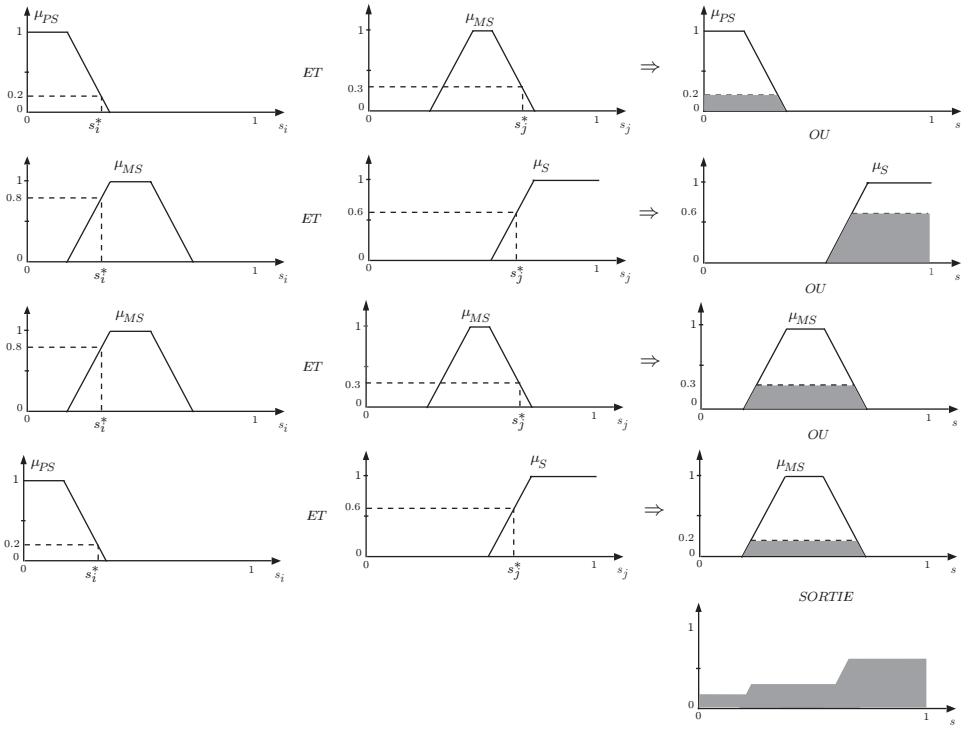


FIG. 3.13 – Exemple de combinaison des similarités s_i et s_j entre deux descripteurs i et j.

La défuzzication réalise donc la transformation de l'ensemble flou de sortie de l'inférence en valeur numérique. Nous utilisons la méthode fréquemment rencontrée du centre de gravité. Soit F l'ensemble flou de sortie du mécanisme d'inférence. La sortie numérique s^* est obtenue par :

$$s^* = \frac{\int \mu_F(s) \cdot s \; ds}{\int \mu_F(s) \; ds} \tag{3.9}$$

Nous avons choisi en sortie du système d'inférence (après défuzzication), une valeur numérique pour représenter la similarité suivant les différents descripteurs. Néanmoins, nous aurions pu aussi rester dans l'espace symbolique, (avant défuzzication) et prendre une décision à partir de ces ensembles flous. Il aurait cependant fallu définir un critère sur les degrés d'appartenance pour décider si les images sont effectivement similaires. Nous avons opté pour décider de la similarité entre deux images suivant la valeur numérique de sortie du système. Cela permet d'être dans les mêmes conditions que la combinaison linéaire des similarités de descripteurs. Par la suite, nous allons voir comment la similarité a été introduite pour construire le résumé.

3.4 Méthode de résumé hiérarchique de vidéo

Pour pouvoir naviguer dans une base de vidéos ou rechercher un extrait dans une séquence, nous proposons une méthode de construction de résumé de vidéo possédant plusieurs niveaux de résolution. Cette méthode est divisée en deux étapes : micro-segmentation de la vidéo et regroupement des segments pour former une macro-segmentation de la vidéo.

Chaque vidéo doit d'abord être structurée afin de faciliter sa visualisation. De nombreuses méthodes ont été proposées pour segmenter la vidéo en plans. Néanmoins, le découpage d'une vidéo en plans est une tâche difficile qui nécessite encore des recherches [Tar05, Che05a] pour améliorer la détection des transitions entre les plans. Une étude comparative a montré que la performance des méthodes de détection diminuait lorsque la transition entre les plans est progressive [Gar00]. C'est d'autant plus vrai en présence de forts mouvements. Pour être indépendant d'une étape de détection des changements de plan, nous découpons la vidéo en segments selon la similarité d'un ou plusieurs descripteurs. Chaque segment est ensuite représenté par une image clé. Cette étape constitue le niveau le plus fin de la résolution du résumé de vidéo et correspond à une micro-segmentation de la vidéo.

Pour réduire le nombre d'images clés, une hiérarchie est construite grâce à un algorithme de regroupement par similarité avec contrainte temporelle. Cette étape consiste à regrouper les segments selon leur similarité pour produire une macro-segmentation. Cette approche a pour objectif de fournir une idée générale et rapide du contenu de la vidéo à l'utilisateur.

De plus, comme notre approche ne dépend pas du type de descripteurs, nous l'expliquerons dans un cadre général.

3.4.1 Segmentation suivant la similarité d'un ou plusieurs descripteurs

Une méthode de segmentation est proposée pour découper la vidéo en segments homogènes suivant la similarité d'un ou plusieurs descripteurs. Une étude expérimentale est ensuite réalisée pour juger de la méthode de segmentation.

3.4.1.1 Méthode de segmentation des vidéos

A partir de la similarité définie entre les images, nous avons conçu une méthode de segmentation qui consiste à regrouper les images similaires et temporellement voisines. L'intérêt de la segmentation est de former des groupes d'images homogènes suivant la mesure de similarité considérée et donc de réduire le nombre d'images en ne conservant qu'une seule image par

groupe créé. Les images ainsi retenues constitueront le premier niveau de résolution (résolution fine) du résumé de vidéo.

Nous allons décrire la procédure de segmentation qui permet de découper la vidéo en segments homogènes. Elle consiste à traiter séquentiellement les images. La première image de la vidéo, qui forme le premier groupe, est comparée à l'image suivante. Si leurs descripteurs sont proches alors les deux images sont regroupées et la moyenne des descripteurs est utilisée pour représenter le centre de gravité de ce groupe. En revanche, si elles ne sont pas similaires, un nouveau groupe est créé. Ce processus est répété jusqu'à la dernière image de la vidéo. Les groupes obtenus forment alors des segments et ils correspondent à la micro-segmentation de la vidéo.

Cette procédure est proche de celle développée par Zhuang et al. [Zhu98]. Cependant, ils effectuent une segmentation à l'intérieur de chacun des plans de la vidéo en utilisant un algorithme de regroupement des images par similarité sans aucune contrainte temporelle. Un groupe ainsi créé peut contenir des images qui ne sont pas forcément contiguës entre elles. De plus, la similarité entre les images est simplement obtenue par comparaison d'histogrammes. L'avantage de notre méthode est qu'elle fonctionne sans connaissance *a priori* des plans de la vidéo. Il suffit de faire l'hypothèse que les similarités entre des images d'un même plan sont plus grandes que celles se trouvant aux transitions entre les plans. L'algorithme que nous proposons impose le regroupement d'images contiguës et a été étudié en combinant plusieurs descripteurs avec différentes mesures de similarité. La méthode de segmentation est récapitulée dans l'algorithme 3.1.

Algorithme 3.1 Algorithme de regroupement par similarité avec contrainte temporelle

Étape 1 (initialisation) : $p = 1$ indice de l'image courante, $n = 1$ nombre de segments créés, centre de gravité du segment courant noté C_n, est égal à $F^p = \{f_1^p, \ldots, f_m^p\}$ et $k = 1$ nombre d'images contenues dans le segment courant.
Étape 2 : $p = p + 1$, s'il n'y pas d'image, on arrête.
Étape 3 : Calcul de la similarité de l'image $p + 1$ avec le centre de gravité courant. Si la similarité est inférieure à un seuil δ alors on va à l'étape 4 sinon à l'étape 5.
Étape 4 : l'image $p+1$ crée un nouveau segment et son vecteur caractéristique est le centre de gravité, $n = n + 1$, $k = 1$ et on va l'étape 2.
Étape 5 : l'image $p+1$ est ajoutée au segment courant $C_n = \frac{k}{k+1} \cdot C_n + \frac{1}{k+1} \cdot F^{p+1}$, $k = k+1$ et on va l'étape 2.

La méthode de segmentation a un paramètre δ qui fournit un contrôle sur la répartition des segments. Si le seuil δ est élevé, beaucoup de segments seront créés en raison de la difficulté à trouver des images similaires. De même, si le seuil δ est trop faible, peu de segments seront formés et des images dissimilaires pourront être assemblées. Une étude doit donc être réalisée pour choisir la valeur du paramètre δ et évaluer la méthode de segmentation.

3.4.1.2 Mesures d'évaluation de la segmentation

La méthode de segmentation fournit un découpage de la vidéo en segments où chacun contient des images qui sont similaires suivant un ou plusieurs descripteurs. Cela implique que les transitions entre les segments doivent être incluses dans les changements de plan. Afin de déterminer la valeur du seuil δ, notre méthode de segmentation est comparée à la segmentation de vidéos dont la partition en plans est connue. L'objectif n'est pas de retrouver les plans de la

vidéo mais de vérifier que les changements de plan contiennent une transition entre segments.

Afin d'évaluer la performance de la segmentation, nous utilisons le rappel et la précision. Le rappel R évalue la capacité de la méthode à retrouver des transitions entre segments dans des changements de plan parmi tous les changements de plan.

$$R = N_{CT}/N_C$$

où N_C désigne le nombre de changements de plan (nombre de transitions correctes) et N_{CT} le nombre de transitions trouvées par la segmentation et contenues dans un changement de plan (nombre de transitions correctes et trouvées). La précision P évalue la capacité de la méthode à retrouver uniquement des transitions entre segments correspondant à des changements de plan.

$$P = N_{CT}/N_T$$

où N_T désigne le nombre de transitions retrouvées par la segmentation (nombre de transitions trouvées) et N_{CT} le nombre de transitions trouvées par la segmentation et contenues dans un changement de plan (nombre de transitions correctes et trouvées). Comme le changement entre deux plans peut posséder une ou plusieurs images, nous le considérons trouvé s'il contient au moins une transition entre deux segments.

Le rappel pourrait être utilisé comme critère d'évaluation de la segmentation. Pour éviter la multiplication des segments à l'intérieur des plans, il faut aussi prendre en compte la précision. Cependant, il est attendu que le rappel soit élevé, proche de 1 au détriment de la précision. En effet, un plan n'est pas nécessairement homogène selon un descripteur considéré. Par exemple, dans un plan, le décor peut changer par un mouvement de caméra. Ainsi plusieurs segments peuvent être créés à l'intérieur d'un même plan.

Nous avons choisi d'évaluer la segmentation de la vidéo selon un ou plusieurs descripteurs en utilisant la mesure F_β. Celle-ci représente une moyenne harmonique pondérée entre la précision et le rappel. Soit P la précision et R le rappel alors la mesure F_β est obtenue par :

$$F_\beta = \frac{(1+\beta^2).P.R}{\beta^2.P+R} \tag{3.10}$$

Si $\beta = 1$, la précision et le rappel sont combinés avec des poids égaux. Si $\beta = 1/2$ alors la précision a deux fois plus de poids que le rappel. En revanche, si $\beta = 2$ c'est le rappel qui a deux fois plus de poids que la précision. Comme notre méthode de segmentation fournit une micro-segmentation, la résolution obtenue est donc plus fine que la résolution des plans. Afin d'en tenir compte, nous utilisons la mesure F_2 pour évaluer la segmentation et déterminer la valeur du seuil δ. Comme celui-ci dépend du choix des descripteurs et de la mesure de similarité utilisée, nous considérons la mesure F_2 comme le critère à optimiser pour obtenir la valeur de δ et cela pour chacune des combinaisons possibles (par exemple, descripteurs couleur et orientation avec similarité linéaire, descripteurs couleur et mouvement avec similarité au sens de la logique floue...).

Des vidéos avec des contenus différents ont été choisies pour étudier la méthode de segmentation. Ainsi quatre vidéos ont été sélectionnées : un documentaire sportif sur le saut à ski avec 20 plans et 3272 images (13 transitions instantanées et 6 transitions progressives), un extrait de la série « The Avengers » avec 28 plans et 2496 images (27 transitions instantanées et 0 transition progressive), un journal télévisé avec 42 plans et 6872 images (36 transitions instantanées et 5 transitions progressives) et le générique de la série « The Avengers » avec 44

plans et 3137 images (42 transitions instantanées et 1 transition progressive). Par la suite, le documentaire sur le saut à ski, l'extrait de la série, le journal télévisé et le générique seront respectivement appelés « Documentaire », « Série », « Journal » et « Générique ».

3.4.1.3 Résultats en considérant individuellement les descripteurs

Les résultats des segmentations selon les descripteurs couleur d'une part et orientation d'autre part sont présentés dans le tableau 3.2. Le rappel et la mesure F_2 sont calculés sur l'ensemble des quatre vidéos exposées dans la section précédente. Nous pouvons observer que le nombre de segments dépend du seuil choisi. Nous pouvons également remarquer que le seuil optimal selon la mesure F_2 pour le descripteur couleur est de 0.60 alors que pour le descripteur orientation, le seuil vaut 0.75. Le descripteur couleur fournit une meilleure segmentation que le descripteur orientation avec un $F_2 = 77\%$ contre $F_2 = 61\%$. Enfin, en ce qui concerne le descripteur mouvement, nous ne l'avons pas étudié parce que ce descripteur ne peut que comparer des paires d'images. Néanmoins, celui-ci peut être associé avec un descripteur (couleur ou orientation) pour étudier la segmentation des vidéos. En effet, seule la première image d'un nouveau segment ne pourra pas être comparée à l'image suivante selon le descripteur mouvement.

Tab. 3.2 – Résultats de la segmentation selon les descripteurs couleur ou orientation.

	Descripteur couleur		Descripteur orientation	
Seuil δ	Rappel (%)	F_2(%)	Rappel (%)	F_2 (%)
0.5	70	67.8	8.4	10.3
0.6	**90.7**	**77.2**	28.4	32
0.65	96.1	74.4	43	45.8
0.7	99.2	66.8	57.6	57.4
0.75	100	53.2	**67.6**	**61.7**
0.8	100	36.1	83.8	57.9
0.85	100	19.1	93	39.4
0.9	100	7.9	98.4	18.7
0.95	100	4.2	100	7

3.4.1.4 Résultats en combinant les descripteurs couleur et orientation

La méthode de segmentation peut être effectuée suivant deux types de combinaison (linéaire ou floue). La combinaison linéaire correspond à une somme pondérée des similarités de chaque descripteur où les poids sont choisis égaux. En revanche, la combinaison floue est effectuée comme sur la figure 3.14. Les paramètres $[a, b, c, d]$ qui permettent de définir l'espace symbolique de représentation PS (pas similaire), MS (moyennement similaire) et S (similaire) ont été fixés de manière empirique et sont différents suivant les descripteurs étudiés. La similarité suivant le descripteur couleur s_c est modélisée avec les paramètres $[a, b, c, d] = [0.5, 0.6, 0.6, 0.7]$ alors que la similarité suivant le descripteur orientation s_o a les paramètres suivants $[a, b, c, d] = [0.4, 0.6, 0.6, 0.8]$. L'espace de sortie est également représenté de la même façon avec l'espace symbolique PS, MS et S avec les paramètres suivants $[a, b, c, d] = [0.2, 0.4, 0.6, 0.8]$.

Différentes règles peuvent être utilisées pour combiner les similarités des différents descripteurs. Les règles que nous avons adoptées sont celles qui ont été présentées dans le tableau 3.1.

Ce jeu de règles (Règles 1) donne le même poids aux similarités des différents descripteurs qui sont combinés. En effet, quels que soient les descripteurs employés, il y a interchangeabilité des règles. Par exemple, lors de la combinaison entre les descripteurs couleur et orientation, nous avons la règle suivante : si s_c est similaire et si s_o est moyennement similaire alors la combinaison s est similaire. Réciproquement, nous avons aussi la règle suivante : si s_c est moyennement similaire et si s_o est similaire alors la combinaison s est similaire.

La défuzzification dans cet exemple (Fig. 3.14) est calculée par la méthode du centre de gravité sur la courbe d. La similarité vaut finalement $s = 0.3$ et elle est le résultat de la combinaison entre les similarités couleur et orientation.

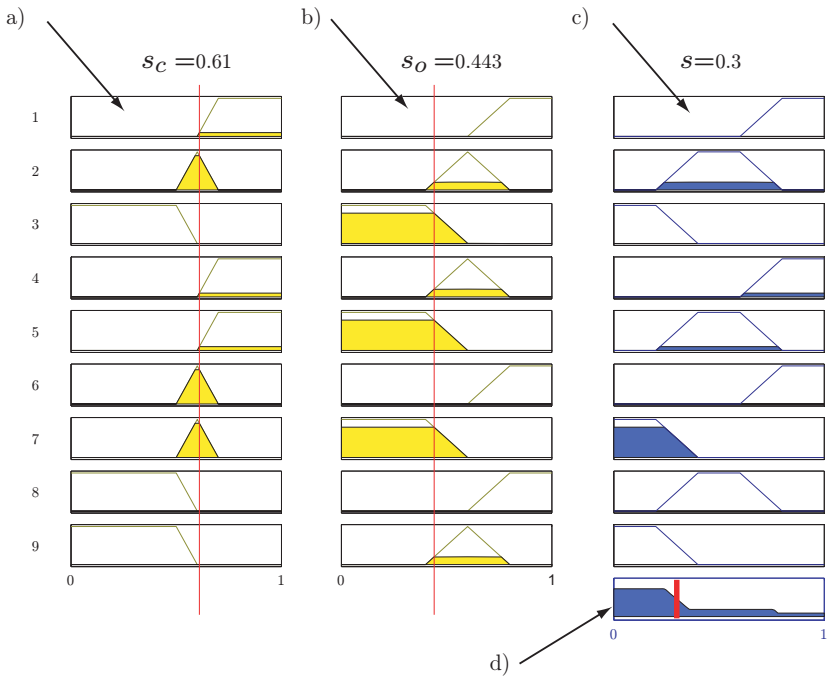


FIG. 3.14 – Exemple de combinaison selon le premier jeu de règles (Règles 1) entre les similarités couleur $s_c = 0.61$ et orientation $s_o = 0.443$. La combinaison résultante s est de 0.3. Les numéros à gauche (de 1 à 9) correspondent aux 9 règles. La règle 1 peut ainsi être déduite : si s_c est similaire (courbe a) et si s_o est similaire (courbe b) alors la combinaison s (courbe c) est similaire. La défuzzification est réalisée par la méthode du centre de gravité sur la courbe d.

Afin de prendre en compte les performances des descripteurs pris séparément, nous développons un second jeu de règles (Règles 2) pour combiner deux descripteurs. Le descripteur couleur fournit de meilleurs résultats que les deux autres. De même, le descripteur orientation a de meilleurs résultats que le descripteur mouvement. Ces remarques sont cohérentes avec le fait que le descripteur couleur a un vecteur caractéristique de dimension plus grande que les deux autres. De plus, le descripteur mouvement informe sur le niveau d'activité entre deux images successives. De nombreuses images peuvent néanmoins avoir le même niveau d'activité et donc ne pas permettre une bonne segmentation. Ainsi, suivant la similarité s_i du descripteur i qui est considéré comme le plus performant, les règles suivantes de combinaison entre les deux descripteurs i et j ont été construites :

- Si s_i est similaire alors s est similaire.
- Si s_i est moyennement similaire et si s_j est similaire alors s est similaire.

– Si s_i est moyennement similaire et si s_j est moyennement similaire alors s est moyennement similaire.
– Si s_i est moyennement similaire et si s_j n'est pas similaire alors s n'est pas similaire.
– Si s_i n'est pas similaire alors s n'est pas similaire.

Ce jeu de règles peut également se retrouver sur la figure 3.15. Chaque ligne (de 1 à 5) correspond à une règle. Par exemple, pour la ligne 4, nous avons la règle suivante : Si s_c est moyennement similaire et si s_o est similaire alors la combinaison résultante s est similaire.

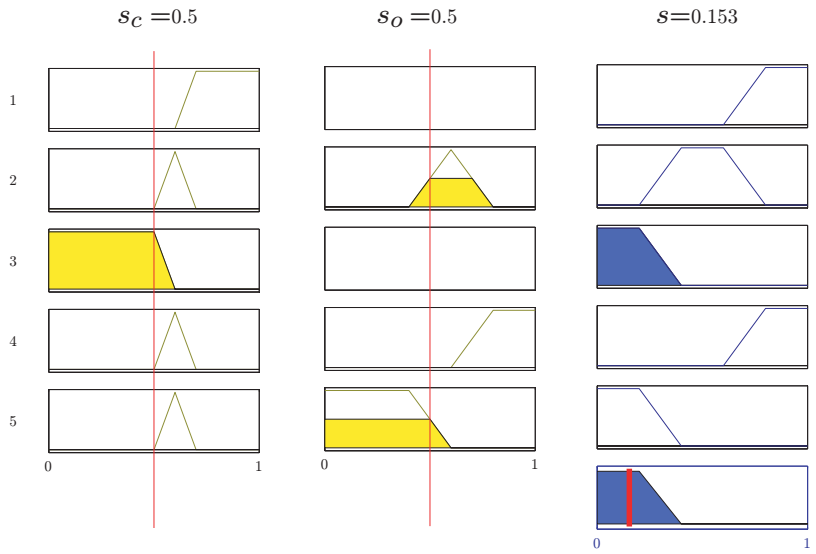


FIG. 3.15 – Exemple de combinaison selon le deuxième jeu de règles (Règles 2) entre les similarités couleur $s_c = 0.5$ et orientation $s_o = 0.5$. La combinaison résultante s est de 0.153. La première ligne correspond à la règle suivante : si s_c est similaire (première colonne) alors la combinaison s (troisième colonne) est similaire.

Les résultats de la méthode de segmentation suivant les descripteurs couleur et orientation sont présentés dans le tableau 3.3. Nous pouvons constater que la combinaison floue selon les règles 1 ou 2 a un F_2 supérieur à celui de la combinaison linéaire. La combinaison floue a un F_2 maximum de 81.7% (pour le seuil 0.75 avec le premier jeu de règles et pour le seuil 0.7 avec le second jeu de règles) alors que la combinaison linéaire présente un F_2 maximum de 81.4% pour le seuil 0.7. De plus, pour la combinaison floue (Règles 1), le rappel est de 97.7% au seuil de 0.75 alors que le rappel est de 96.2% au seuil de 0.7 pour la combinaison linéaire. Enfin, quelle que soit la combinaison (linéaire ou floue), le F_2 obtenu est meilleur que le F_2 pris séparément (couleur ou orientation). La segmentation des vidéos selon les descripteurs couleur et orientation présente de meilleurs résultats (au sens de la mesure F_2) que la segmentation selon le descripteur couleur ou orientation. Néanmoins, le découpage selon plusieurs descripteurs doit normalement augmenter le nombre de segments dans une vidéo puisque les images contenues dans un segment sont similaires suivant les descripteurs étudiés. Cela signifie que la segmentation des vidéos avec les descripteurs couleur et orientation améliore la détection des changements de plan et de ce fait compense l'augmentation du nombre de segments. Cependant, si le nombre de descripteurs augmente, le critère de similarité entre les images augmentera le nombre de segments et entraînera la diminution de la mesure F_2.

TAB. 3.3 – Résultats de la segmentation en combinant les descripteurs couleur et orientation.

	Combinaison linéaire		Combinaison floue (Règles 1)		Combinaison floue (Règles 2)	
Seuil δ	Rappel (%)	F_2 (%)	Rappel	F_2 (%)	Rappel (%)	F_2 (%)
0.6	63.8	63.2	88.5	79.2	89.2	79
0.65	86.9	80.6	90	78.5	92.3	79.7
0.7	**96.2**	**81.4**	95.4	81.3	**96.2**	**81.7**
0.75	99.2	73.5	**97.7**	**81.7**	96.9	80.5
0.8	100	55.6	97.7	79.4	96.9	78.4

3.4.1.5 Résultats en combinant les descripteurs couleur et mouvement

Nous pouvons reprendre l'étude en combinant la couleur et le mouvement. Les résultats de la combinaison entre les descripteurs couleur et mouvement sont présentés dans le tableau 3.4. Nous pouvons observer que les résultats de la combinaison linéaire sont inférieurs à ceux de la combinaison floue (Règles 1 ou 2). Cela signifie que les descripteurs couleur et mouvement ne doivent pas être combinés de façon linéaire. Les figures 3.16 et 3.17 montrent comment les similarités des descripteurs sont modélisées ainsi que la combinaison par les règles (Règles 1 ou 2). Comme signalé précédemment, la combinaison floue des descripteurs (couleur et mouvement) n'améliore pas forcément la segmentation. En effet, les mesures F_2 des combinaisons floues sont du même ordre de grandeur (au alentour de 77%) que celles obtenues pour la segmentation selon le descripteur couleur. Mais la segmentation obtenue avec les descripteurs couleur et mouvement présente l'avantage de former des segments homogènes suivant ces deux descripteurs.

TAB. 3.4 – Résultats de la segmentation en combinant le descripteur couleur et mouvement.

	Combinaison linéaire		Combinaison floue (Règles 1)		Combinaison floue (Règles 2)	
Seuil δ	Rappel (%)	F_2 (%)	Rappel	F_2 (%)	Rappel (%)	F_2 (%)
0.6	**92.3**	**38.9**	85.4	74.8	88.5	77.8
0.65	92.3	27.4	89.2	75.1	90	77.7
0.7	88.5	18.5	**93.8**	**76.2**	**93.8**	**79.4**
0.75	87.7	12.6	96.2	75.3	95.4	78.8
0.8	96.2	10.3	97.7	73.7	96.2	78.3

L'étude sur la combinaison des descripteurs orientation et mouvement a aussi été réalisée. Cependant, comme les résultats sont moins bons que la combinaison des descripteurs couleur et orientation, et couleur et mouvement, nous ne les avons pas présentés. On pouvait observer de manière similaire à la combinaison des descripteurs couleur et mouvement que les résultats de la combinaison linéaire sont inférieurs à ceux de la combinaison floue (Règles 1 ou 2).

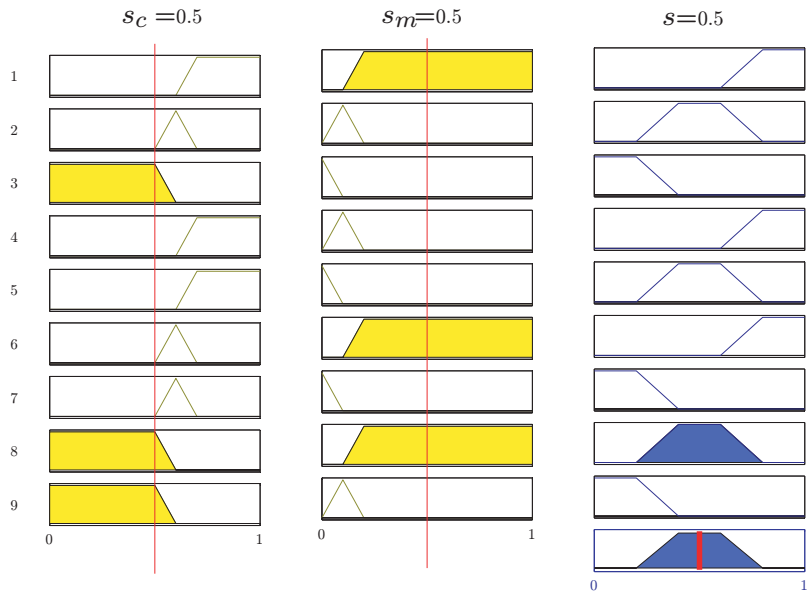


FIG. 3.16 – Exemple de combinaison selon le premier jeu de règles (Règles 1) entre les similarités couleur $s_c = 0.5$ et mouvement $s_m = 0.5$. La combinaison résultante s est de 0.5.

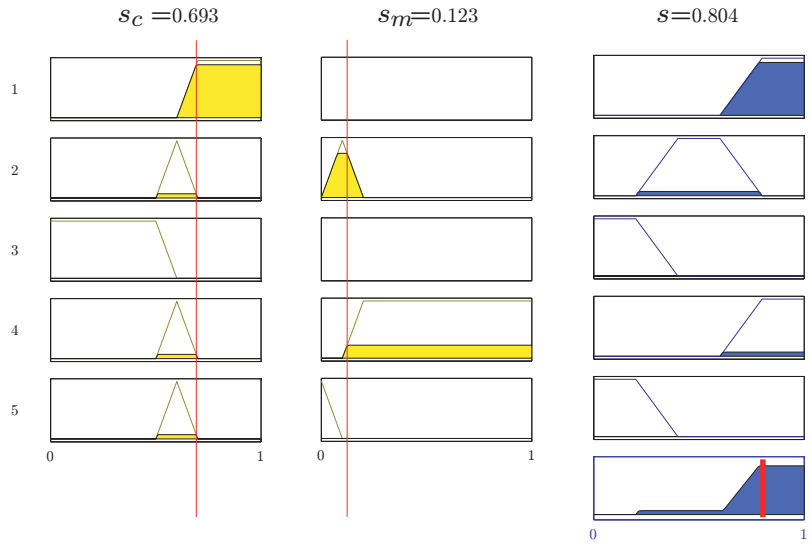


FIG. 3.17 – Exemple de combinaison selon le deuxième jeu de règles (Règles 2) entre les similarités couleur $s_c = 0.693$ et mouvement $s_m = 0.123$. La combinaison résultante s est de 0.804.

3.4.1.6 Résultats en combinant les descripteurs couleur, orientation et mouvement

La combinaison selon les trois descripteurs est présentée et nécessite la définition de nouvelles règles. Tout d'abord, nous avons élaboré des règles (Règles 1) de même poids quels que soient les descripteurs employés. Si les similarités d'au moins deux descripteurs sur les trois sont similaires (respectivement pas similaires) alors la combinaison résultante est similaire (respectivement pas similaire). Sinon la combinaison résultante est considérée comme moyennement similaire. Ces règles peuvent se retrouver sur la figure 3.18.

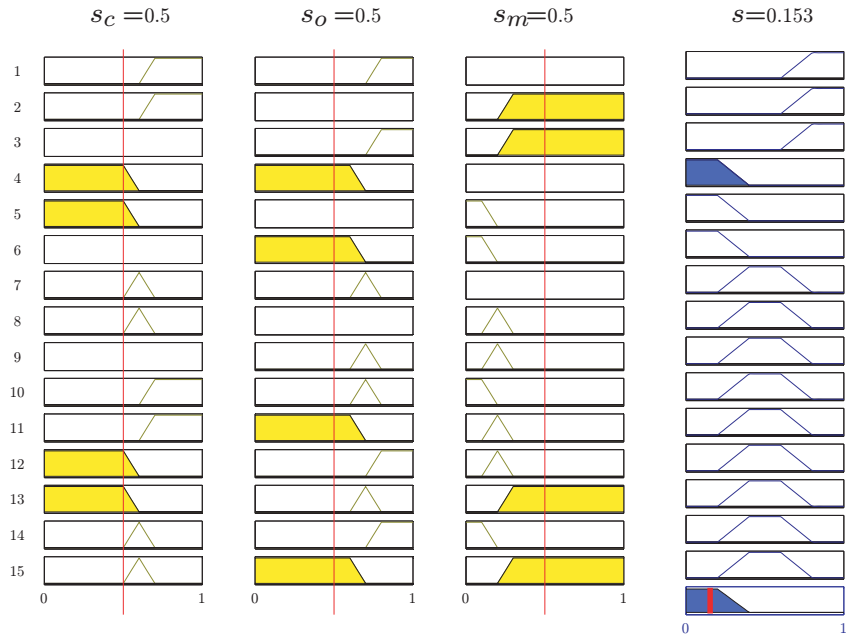


FIG. 3.18 – Exemple de combinaison selon le premier jeu de règles (Règles 1) entre les similarités couleur $s_c = 0.5$, orientation $s_o = 0.5$ et mouvement $s_m = 0.5$. La combinaison résultante s est de 0.123.

Comme la combinaison des similarités de deux descripteurs, nous développons un second jeu de règles qui prend en compte les performances des descripteurs. Les règles suivantes de combinaison entre les similarités couleur s_c, orientation s_o et mouvement s_m sont construites :

- Si s_c est similaire alors s est similaire.
- Si s_c est moyennement similaire et si s_o est similaire alors s est similaire.
- Si s_c est moyennement similaire et si s_o n'est pas similaire alors s n'est pas similaire.
- Si s_c est moyennement similaire et si s_o est moyennement similaire et si s_m est similaire alors s est similaire.
- Si s_c est moyennement similaire et si s_o est moyennement similaire et si s_m n'est pas similaire alors s n'est pas similaire.
- Si s_c est moyennement similaire et si s_o est similaire et si s_m est moyennement similaire alors s est moyennement similaire.
- Si s_c n'est pas similaire alors s n'est pas similaire.

Ce jeu de règles (Règles 2) peut également se retrouver sur la figure 3.19.

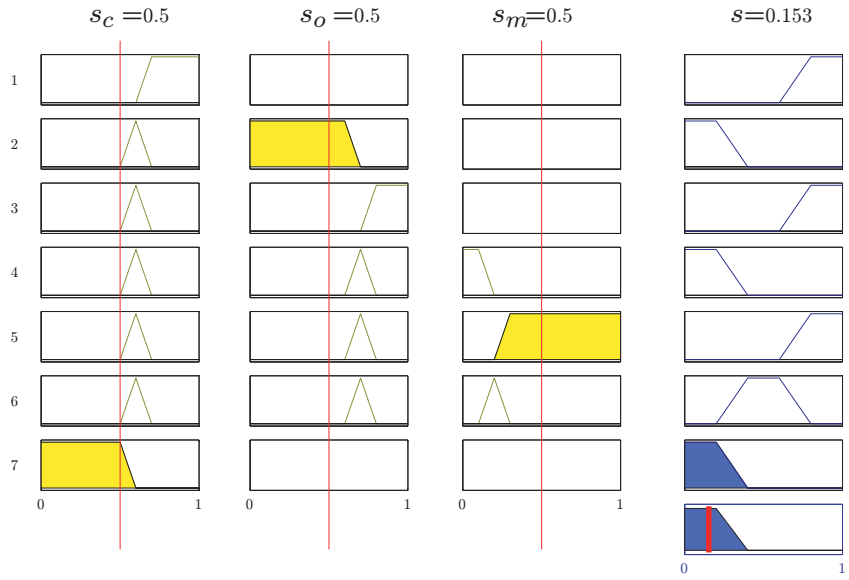


FIG. 3.19 – Exemple de combinaison selon le deuxième jeu de règles (Règles 2) entre les similarités couleur $s_c = 0.5$, orientation $s_o = 0.5$ et mouvement $s_m = 0.5$. La combinaison résultante s est de 0.123.

Le tableau 3.5 montre les résultats de la segmentation en combinant les trois descripteurs (couleur, orientation et mouvement). La combinaison linéaire fournit une segmentation moins bonne que celles obtenues avec les combinaisons floues (Règles 1 ou 2). Nous pouvons également remarquer que la combinaison floue suivant les règles 2 qui privilégie les performances des descripteurs fournit le meilleur F_2 avec 81.1%. Comme déjà signalé auparavant, plus le nombre de descripteurs augmente plus la similarité entre les images est difficile et donc plus le nombre de segments augmente. Cela explique que la mesure F_2 (74.8%) de la combinaison floue suivant les règles 1 pour les trois descripteurs soit inférieure à la mesure F_2 (81.7%) de la combinaison floue (Règles 1 ou 2) pour les descripteurs couleur et orientation. Par contre, la mesure F_2 (81.1%) de la combinaison floue suivant les règles 2 pour les trois descripteurs est du même ordre de grandeur que la mesure F_2 (81.7%) de la combinaison floue (Règles 1 ou 2) pour les descripteurs couleur et orientation.

TAB. 3.5 – Résultats de la segmentation en combinant les descripteurs couleur, orientation et mouvement.

Combinaison	linéaire		floue (Règles 1)		floue (Règles 2)	
Seuil δ	Rappel (%)	F_2 (%)	Rappel (%)	F_2 (%)	Rappel (%)	F_2 (%)
0.6	91.5	67.8	85.4	74.8	90.8	79.6
0.65	94.6	55.5	**86.9**	**74.8**	93.8	81
0.7	96.9	40.9	89.2	74.1	94.6	81.1
0.75	96.2	25.4	90	72	**96.2**	**81.1**
0.8	97.7	15.8	90.8	70.9	96.2	79.6

3.4.1.7 Création du premier niveau du résumé

La segmentation de vidéos est à l'origine de la méthode de création du premier niveau de résolution du résumé (résumé fin). Ainsi, pour chaque combinaison possible de descripteurs, la segmentation ayant la valeur F_2 maximale est choisie comme segmentation de la vidéo. Le seuil δ de la méthode de segmentation est donc fixé à partir du F_2 et ne dépend que du choix de la combinaison et des descripteurs. La vidéo utilisée n'est par contre pas liée au seuil δ puisque la mesure F_2 est calculée sur l'ensemble des quatre vidéos.

De plus, chaque segment qui contient moins de 5 images est regroupé avec l'un de ses deux segments voisins. La similarité entre vecteurs caractéristiques permet de déterminer avec lequel il est regroupé. Comme la segmentation conduit à regrouper des images suivant leur similarité (selon un ou plusieurs descripteurs), les images clés sont déterminées en fonction des segments obtenus. Ainsi les images contenues dans un segment sont similaires et sont donc représentées par l'image la plus proche du centre de gravité du segment. Finalement les images obtenues sont définies comme les images clés et sont utilisées pour résumer la vidéo.

La figure 3.20 montre un exemple de segmentation selon le descripteur couleur sur un extrait de la vidéo « Générique ». Chaque segment est représenté par 3 images : image du début, image clé et image de fin du segment. Par exemple, le premier segment va de l'image 1332 à l'image 1393 et l'image 1358 est l'image clé qui résume ce segment. Nous pouvons remarquer que de nombreux segments sont inclus dans des plans et que les changements de plan sont situés à la transition entre deux segments. Par exemple, la fin du premier segment (image 1393) et le début du segment suivant (image 1394) correspond à un changement de plan. Nous pouvons aussi constater que certain plan possède plusieurs segments. Par exemple, le segment de 1428 à 1476 et celui de 1477 à 1486 appartiennent au même plan. Dans cette figure 3.20, deux changements de plan n'ont pas été correctement identifiés. Il s'agit du segment de 1531 à 1567 et celui de 1717 à 1782 qui contiennent un changement de plan. Cette segmentation suivant le descripteur couleur ne permet pas de les détecter en raison de la similarité en terme de couleur.

La figure 3.21 présente la segmentation obtenue en combinant les descripteurs couleur et orientation avec les règles 1. Nous pouvons observer que tous les segments présentés sont inclus dans un plan. Par exemple, la transition entre le segment des images 1717 à 1781 et celui des images 1782 à 1815 est correctement positionnée et correspond à un changement de plan. Ces deux exemples (Fig. 3.21 et Fig. 3.20) montrent également les images clés résumant chaque segment. Nous pouvons observer la similarité entre les images clés et les segments.

La sélection d'images au niveau de la segmentation de la vidéo correspond au niveau le plus fin du résumé de vidéo. Cependant, comme le nombre de segments obtenus est élevé, le nombre d'images clés est trop conséquent pour fournir un aperçu rapide à l'utilisateur. Une étape de regroupement des segments par similarité s'avère donc nécessaire.

FIG. 3.20 – Exemple de segmentation selon le descripteur couleur sur un extrait de la vidéo « Générique ». Un segment est représenté par 3 images. Pour le premier segment, l'image 1332 est l'image de début du segment, l'image 1358 est l'image clé du segment et enfin l'image 1393 est l'image de fin du segment.

FIG. 3.21 – Exemple de segmentation (Règles 1) en combinant le descripteur couleur et orientation sur un extrait de la vidéo « Générique ».

3.4.2 Regroupement des segments par similarité avec contrainte temporelle

Afin de créer un résumé de vidéo de taille variable, nous construisons un résumé hiérarchique possédant différents niveaux de résolution. Comme signalé dans le chapitre 2, de nombreux algorithmes (comme l'algorithme des k-moyens [Zho96, Far02], l'algorithme de classification hiérarchique ascendant [Ben06] ou l'algorithme de l'arbre couvrant de poids minimum [Kir02]) ont été appliqués pour regrouper les segments similaires. L'inconvénient de ces approches est que la composante temporelle n'est pas exploitée pour effectuer le regroupement. Ainsi des segments éloignés temporellement peuvent être réunis.

Des travaux ont introduit la composante temporelle dans les algorithmes de regroupement. Yeung et al. [Yeu96] proposent une méthode hiérarchique ascendante où la mesure de similarité est contrainte temporellement. Dans [Gir00], un algorithme de classification hiérarchique avec contrainte temporelle est appliqué. De même, Rui et al. [Rui98] utilisent l'algorithme des k-moyens avec une mesure de similarité ayant une contrainte temporelle. Une étude approfondie décrite dans [Ven02] a été menée en distinguant différentes mesures de similarité et a abouti à une macro-segmentation des vidéos en employant l'algorithme de classification hiérarchique avec contrainte temporelle.

Nous allons présenter un nouvel algorithme de regroupement par similarité avec contrainte temporelle. Afin de préserver une cohérence temporelle entre les images clés de la vidéo, une contrainte temporelle est imposée pour le regroupement des segments. Elle a pour but d'empêcher le regroupement de segments trop éloignés. L'algorithme consiste à regrouper des segments similaires et proches temporellement (mais pas nécessairement adjacents) selon un ou plusieurs descripteurs. Cette approche est similaire à celle de l'algorithme de classification hiérarchique avec contrainte temporelle. Elle consiste à regrouper les deux éléments (segments) les plus similaires et à recommencer jusqu'à atteindre un nombre souhaité d'éléments. La différence de cette approche par rapport à celle que nous proposons est que notre algorithme possède deux paramètres (seuil de similarité et largeur de la fenêtre temporelle) qui permettent, si on le souhaite, de forcer le regroupement de segments moyennement similaires mais proches temporellement lors de la première itération. Puis par itérations successives, les segments de plus en plus éloignés mais similaires sont regroupés. L'avantage de la méthode est qu'elle permet par exemple de regrouper au premier niveau de résolution seulement des segments adjacents puis au dernier niveau de regrouper des segments sans contrainte temporelle.

On suppose que la vidéo a été partitionnée en segments homogènes comme présenté dans la section précédente. Chaque segment homogène possède un vecteur caractéristique C_i et l'image du segment la plus proche du vecteur C_i est définie comme image clé. Une distance temporelle d_t et une similarité temporelle s_t entre les segments sont alors définies par l'équation 3.11. La distance temporelle d_t correspond à l'éloignement temporel entre deux segments et la similarité temporelle s_t est une fonction quadratique qui permettra le regroupement de segments ayant un éloignement temporel inférieur à la largeur d'une fenêtre temporelle donnée.

$$d_t(i,j) = |i - j - 1|$$

$$s_t(i,j) = \begin{cases} 1 - \left(\frac{d_t(i,j)}{\mathrm{w}}\right)^2 & \text{si } d_t < \mathrm{w} \\ 0 & \text{si } d_t \geq \mathrm{w} \end{cases} \tag{3.11}$$

où i, j sont les positions des segments dans la vidéo et w est la largeur de la fenêtre temporelle.

La similarité pondérée s_w entre deux segments est alors obtenue en multipliant la similarité temporelle s_t par la similarité des descripteurs s. Ainsi les segments dont la similarité s_w est plus grande qu'un seuil T_h sont regroupés. Afin d'avoir plusieurs niveaux de hiérarchie, la largeur de la fenêtre w doit augmenter et le seuil T_h doit diminuer entre deux niveaux de hiérarchie. Ainsi les regroupements sont d'abord réalisés localement puis en montant dans la hiérarchie, ils sont effectués de manière plus globale. Cette méthode de regroupement par similarité avec contrainte temporelle permet la création d'un résumé hiérarchique. L'algorithme 3.2 détaille la procédure de construction de la hiérarchie avec contrainte temporelle. Nous pouvons noter que l'algorithme commence avec une fenêtre w de petite taille et se poursuit avec une fenêtre de taille de plus en plus grande. Lors des premières itérations, les segments similaires et proches temporellement sont regroupés. Puis, par itérations successives, les segments éloignés et moins similaires sont regroupés. Au dernier niveau de la hiérarchie, le nombre d'images clés est donc réduit.

Algorithme 3.2 Algorithme de regroupement par similarité avec contrainte temporelle

INITIALISATION : Etant donné un seuil T_h, la largeur de la fenêtre temporelle w et N le nombre de segments.

Etape 1 : Calcul de la similarité entre les segments

$\forall i, j \in 1 \cdots N$, calcul de $s(i,j)$, $s_t(i,j)$ et $s_w(i,j) = s_t(i,j) \cdot s(i,j)$

Etape 2 : Recherche de segments à regrouper

$\forall i \in 1 \cdots N,\ fusion\{i\} = \{i\}$

Pour $i = 1 \cdots N$

  Pour $j = i \cdots i + \mathrm{w}$

    Si $s_w(i,j) > T_h$

      $fusion\{i\} = fusion\{i\} \cup \{j\}$

    Fin Si

  Fin Pour

Fin Pour

% Mise à jour

Pour $i = 1 \cdots N$

  $\Omega = \{x \in fusion\{i\}\ tel\ que\ x > i\}$

  $k = min(\Omega)$

  Si $NONVIDE(k)$

    $fusion\{k\} = fusion\{i\}$

    $fusion\{i\} = \emptyset$

  Fin Si

Fin Pour

Etape 3 : Regroupement des segments

- Calcul de la moyenne des vecteurs caractéristiques pondérés par le nombre d'images contenues dans chacun des segments à regrouper.
- Chaque regroupement est ensuite positionné à l'endroit où le segment qu'il contient a le plus d'images et il est représenté par l'image clé de ce segment.
- Mise à jour de N, $T_h = T_h - 0.05$, $\mathrm{w} = 2 \cdot \mathrm{w}$
- Retourner à l'étape 1 pour créer un nouveau niveau de hiérarchie.

Dans cet algorithme 3.2, la mise à jour des paramètres T_h et w est primordiale. En effet, les regroupements s'effectuent seulement avec des segments similaires (supérieurs à T_h)

et éloignés au maximum de la taille de la fenêtre. Les itérations successives augmentent la largeur de cette fenêtre ($w = 2 \cdot w$) donc permet le regroupement de segments de plus en plus éloignés et de moins en moins similaires ($T_h = T_h - 0.05$). D'autres mises à jour sont ici possibles et modifient singulièrement le regroupement. Supposons que $T_h = T_h - 0.05$ soit remplacé par $T_h = T_h + 0.05$ et que l'initialisation de T_h soit plus faible que dans le cas précédent, alors les segments moyennement similaires et proches temporellement seront regroupés lors des premières itérations puis les segments très similaires et pas forcément proches temporellement seront regroupés lors des dernières itérations. De la même manière, l'évolution du paramètre w joue un rôle important. Par exemple, si on souhaite un regroupement sans contrainte temporelle alors la largeur de la fenêtre doit être fixée à $w = \infty$. La figure 3.22 présente une illustration de l'étape 2 de l'algorithme 3.2 de regroupement par similarité avec contrainte temporelle. Trois éléments sont distribués sur un axe temporel. Les regroupements (notés c_i) sont obtenus avec une fenêtre rectangulaire de taille (a) $w = 2$ et (b) $w = 4$. Nous pouvons remarquer que les regroupements ne dépendent pas seulement de la similarité entre les données mais aussi de la distribution temporelle des données et donc de la largeur de la fenêtre.

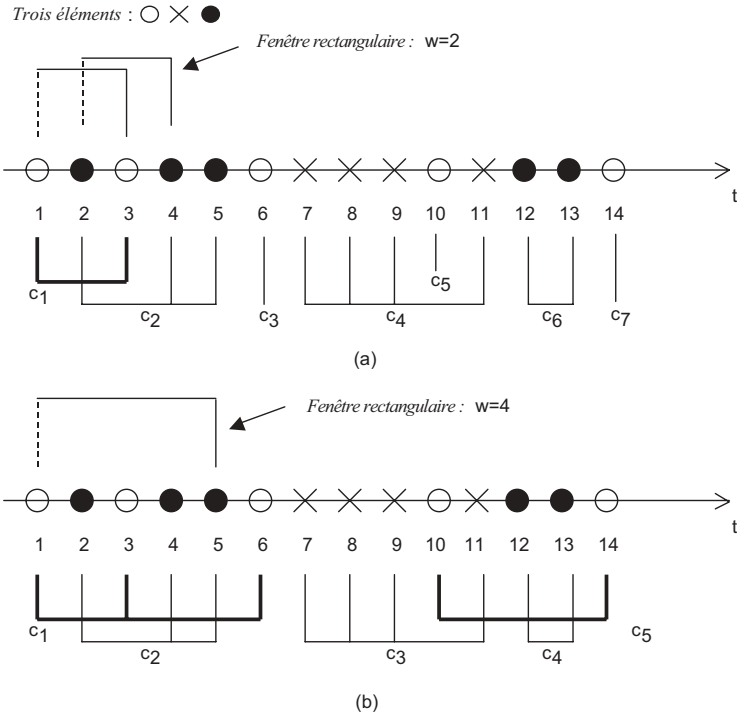


FIG. 3.22 – Illustration de l'étape 2 de l'algorithme 3.2 de regroupement par similarité avec contrainte temporelle. Trois éléments distribués sur un axe temporel sont regroupés avec une fenêtre rectangulaire de taille (a) $w = 2$ et (b) $w = 4$.

La figure 3.23 présente un exemple de l'algorithme de regroupement sur un extrait de la vidéo « Documentaire » où la segmentation est obtenue avec les descripteurs couleur et orientation, et les règles 1 de combinaison. Nous pouvons remarquer que les images qui sont regroupées ont des contenus similaires. La hiérarchie permet le regroupement d'images de plus en plus éloignées temporellement et similaires. La figure 3.24 présente le résumé hiérarchique

de la vidéo « Documentaire ». La combinaison est effectuée avec les trois descripteurs selon les règles 2. Le premier niveau (Fig. 3.24(a)) correspond à la première étape de la méthode de « micro-segmentation » de la vidéo où les images clés sont extraites sur chacun des segments. L'algorithme de regroupement par similarité avec contrainte temporelle permet de passer du premier niveau au troisième niveau. Le résumé possède 43 images pour le premier niveau, 26 images pour le deuxième et enfin 4 images pour le troisième. Cet algorithme permet la réduction du résumé en regroupant les images les plus similaires et proches temporellement. Le dernier niveau de résolution regroupe les images similaires et éloignées dans le temps pour ne conserver que les images dissimilaires. Si un utilisateur souhaite un nombre donné d'images pour le résumé alors l'algorithme de regroupement par similarité s'arrête au niveau de résolution qui précède celui où le nombre d'images clés obtenues est inférieur au nombre demandé par l'utilisateur. Puis les deux segments les plus similaires sont regroupés jusqu'à atteindre le nombre demandé.

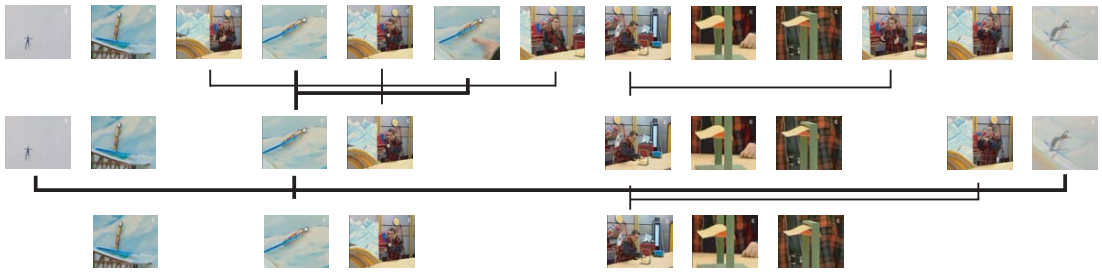
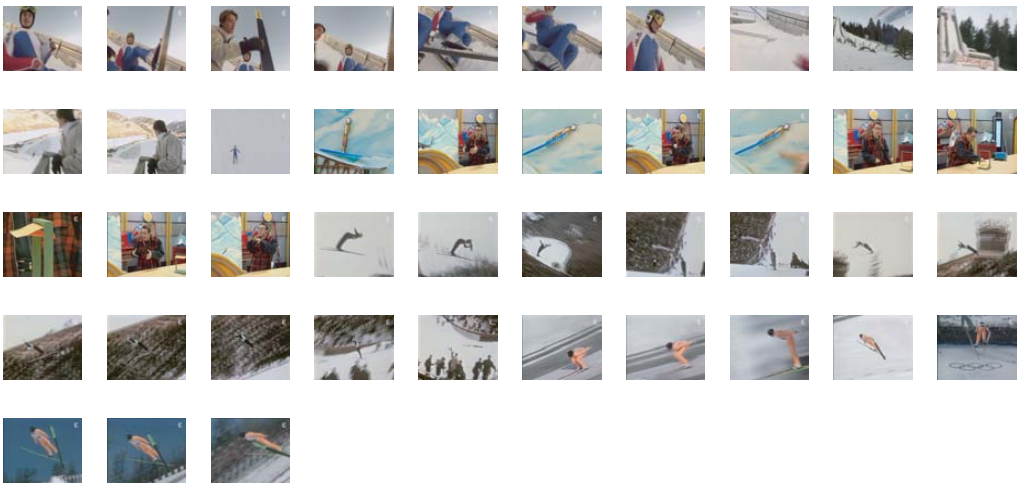


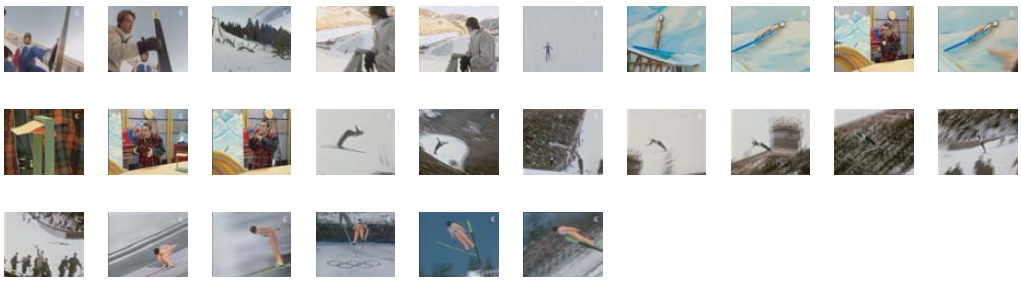
FIG. 3.23 – Exemple de hiérarchie avec les descripteurs couleur et orientation selon les règles 1 de combinaison sur un extrait de la vidéo « Documentaire ». Les paramètres de l'algorithme sont fixés à $T_h = 0.75$ et w = 10.

En définitive, nous avons présenté une méthode de résumé hiérarchique qui a l'avantage de s'adapter à la demande de l'utilisateur. Un algorithme de regroupement par similarité avec contrainte temporelle a été mis au point et permet le regroupement de segments suivant la similarité des descripteurs et leurs positions dans la vidéo. Cet algorithme possède deux paramètres qui réalisent les regroupements suivant nos attentes. Par exemple, l'algorithme peut empêcher le regroupement de segments similaires mais éloignés temporellement lors des premières itérations, puis lors des dernières itérations, le regroupement peut avoir lieu avec des segments similaires et quelles que soient leurs positions dans la vidéo. De la même manière, l'algorithme peut forcer le regroupement de segments moyennement similaires mais proches temporellement lors des premières itérations, puis autoriser le regroupement de segments très similaires quelles que soient leurs positions lors des dernières itérations.

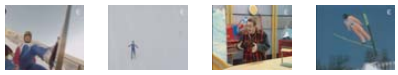
Nous avons également étudié la manière de combiner la similarité entre les descripteurs. Une approche classique qui repose sur la combinaison linéaire des similarités a été comparée à celle fournissant une combinaison au sens de la logique floue. Nous avons observé que les combinaisons linéaire et floue donnent des résultats voisins pour les descripteurs couleur et orientation. En revanche, dès que le mouvement est utilisé, les résultats sont meilleurs avec les combinaisons floues. En effet, le mouvement est moins discriminant que les deux autres descripteurs et une combinaison linéaire avec les mêmes poids est moins efficace que celle avec la logique floue.



(a) Premier niveau de résolution.



(b) Deuxième niveau de résolution.



(c) Troisième niveau de résolution.

FIG. 3.24 – Exemple de résumé hiérarchique sur la vidéo « Documentaire ». La combinaison est réalisée avec les descripteurs couleur, orientation et mouvement selon les règles 2. Les paramètres de l'algorithme sont fixés à $T_h = 0.8$ et w = 10. Le résumé possède 43 images clés au premier niveau, 26 au deuxième et 4 au troisième.

3.5 Application à la recherche par l'exemple

Une application telle que la recherche par exemple permet de juger de l'efficacité de la méthode proposée pour créer des résumés de vidéos. Comme notre méthode est divisée en 2 étapes : micro-segmentation et macro-segmentation, nous effectuons d'abord une étude de recherche par l'exemple au niveau de la segmentation des vidéos (première étape), puis nous procédons à une étude de recherche par l'exemple en utilisant la hiérarchie des vidéos (deuxième étape).

3.5.1 Au niveau de la segmentation

Il s'agit de vérifier que la méthode de segmentation des vidéos (premier niveau de la hiérarchie) est efficace pour retrouver les segments auxquels les images clés appartiennent. La recherche par l'exemple consiste à comparer une image requête (ou une paire d'images si le mouvement est utilisé) avec les segments des différentes vidéos. Plus précisément, la similarité entre le ou les descripteurs de l'image requête et celui (ou ceux) du centre de gravité d'un segment est déterminée. Le segment qui a la similarité la plus grande (selon un ou plusieurs descripteurs) avec l'image requête est considéré comme le plus proche de l'image requête. De plus, les segments sont triés dans l'ordre décroissant de similarité. En cas d'égalité des similarités, ce sont les similarités du descripteur couleur (ou orientation si la couleur n'est pas utilisée) qui départagent ces segments.

Cette méthode a été testée sur les 4 vidéos où la segmentation a été étudiée. Nous avons effectué un tirage aléatoire de 100 images dans chaque vidéo et nous avons vérifié que les segments auxquels elles appartiennent sont bien retrouvés. Nous employons une mesure définie dans [Man01] qui détermine le nombre de segments retrouvés sur les α premiers segments retournés.

Le tableau 3.6 présente les résultats de la recherche par l'exemple en combinant les descripteurs couleur et orientation au niveau du premier niveau de la hiérarchie du résumé de vidéo. Cependant, comme les segmentations sont différentes suivant la combinaison utilisée, les résultats ne peuvent pas être comparés entre les différentes combinaisons. Nous pouvons seulement évaluer l'apport de la combinaison sur la recherche par l'exemple par rapport à celle effectuée par les descripteurs pris séparément. Pour les 4 vidéos, la recherche par l'exemple fournit de bons résultats avec plus de 80% de segments retrouvés quelle que soit la combinaison. Nous pouvons également observer que la combinaison améliore les résultats de la recherche par l'exemple. Comme l'illustre la segmentation obtenue avec les règles 1 de combinaison, la recherche par l'exemple suivant le descripteur couleur offre de bons résultats avec 81.7% de segments retrouvés pour $\alpha = 1$ et pour les 4 vidéos. En revanche, la combinaison des descripteurs (couleur et orientation) est plus efficace avec 85% de segments retrouvés. Nous pouvons aussi remarquer que les différentes combinaisons présentent des résultats similaires. Cependant, bien que les images requêtes soient les mêmes pour les différentes combinaisons, la recherche des segments s'appuie sur une segmentation différente et dépend de la combinaison. Le tableau 3.7 montre les résultats de la recherche par l'exemple en combinant les descripteurs couleur, orientation et mouvement. Nous pouvons constater que seules les combinaisons floues améliorent la recherche par l'exemple alors que la combinaison linéaire a des résultats moins bons que ceux fournis par la couleur. Ceci s'explique par une combinaison donnant les mêmes poids aux différents descripteurs.

TAB. 3.6 – Résultats de la recherche par l'exemple en combinant les descripteurs couleur et orientation.

Combinaison		floue en % (Règles1)			floue en % (Règles2)			linéaire en %		
Vidéo	α	comb.	couleur	orient.	comb.	couleur	orient.	comb.	couleur	orient.
Documentaire	1	80	79	29	75	77	27	80	78	27
	2	94	95	43	95	94	40	96	96	42
	3	100	100	54	98	99	50	98	99	52
Journal	1	93	91	27	92	91	27	94	90	25
	2	99	97	35	99	97	36	99	97	37
	3	100	98	48	99	99	48	99	98	46
Série	1	81	77	38	75	72	35	81	79	36
	2	88	85	48	88	81	44	90	86	52
	3	93	92	54	94	90	49	94	91	54
Générique	1	86	80	36	82	79	37	89	81	39
	2	96	92	43	94	91	41	98	92	45
	3	98	96	49	97	96	50	99	96	48
Total	1	85	81.7	32.5	81	79.7	31.5	86	82	31.7
	2	94.2	92.2	42.2	94	90.7	40.2	95.7	92.7	44
	3	97.7	96.5	51.2	97	96	49.2	97.5	96	50

TAB. 3.7 – Résultats de la recherche par l'exemple en combinant les descripteurs couleur, orientation et mouvement.

Combinaison		floue en % (Règles1)				floue en % (Règles2)				linéaire en %			
Vidéo	α	comb.	coul.	orient.	mouv.	comb.	coul.	orient.	mouv.	comb.	coul.	orient.	mouv.
Documentaire	1	76	72	24	3	78	77	28	1	69	81	35	0
	2	89	89	39	4	93	92	41	5	84	93	50	0
	3	91	95	46	5	99	100	51	9	91	96	62	2
Journal	1	94	93	31	3	92	91	28	4	75	91	36	4
	2	96	97	38	7	99	97	35	5	88	97	47	6
	3	96	97	49	10	99	98	49	12	91	100	50	9
Série	1	81	79	31	2	75	74	33	1	70	80	34	2
	2	92	88	48	3	89	88	49	3	79	89	52	2
	3	94	93	50	5	93	93	52	4	82	91	54	6
Générique	1	81	77	29	4	82	81	34	5	70	85	41	6
	2	89	84	38	5	95	91	41	6	77	94	57	6
	3	95	90	45	7	97	95	45	7	81	95	62	7
Total	1	83	80.2	28.7	3	81.7	80.7	30.7	2.7	71	84.2	36.5	3
	2	91,5	89.5	40.7	4.7	94	92	41.5	4.7	82	93.2	51.5	3.5
	3	94	93.7	47.5	6.7	97	96.5	49.2	8	86.2	95.5	57	6

3.5.2 Au niveau de la hiérarchie

La construction d'une hiérarchie doit normalement diminuer les performances de la recherche par l'exemple. En effet, le regroupement de segments et le remplacement par le vecteur moyen des segments doit abaisser les performances de la recherche par l'exemple. Par contre, la hiérarchie doit accélérer le résultat de la recherche par l'exemple puisque les comparaisons entre l'image requête et les segments doivent diminuer. Afin de s'assurer de la robustesse de la hiérarchie, nous effectuons une recherche par l'exemple en comparant la requête au dernier niveau de la hiérarchie (ici 4 niveaux). Les trois segments du niveau 4 les plus proches de la requête sont conservés, puis les segments du niveau inférieur (niveau 3) qui ont engendré les trois segments du niveau 4 sont déterminés. Parmi ces segments, les trois plus proches sont alors déterminés et le procédé est répété jusqu'au premier niveau.

Nous avons testé cette approche sur la hiérarchie en utilisant d'une part les descripteurs couleur et orientation et d'autre part les descripteurs couleur, orientation et mouvement. L'algorithme de regroupement par similarité a été appliqué avec les paramètres initiaux $T_h = 0.75$ et w = 10 pour les descripteurs couleur et orientation, et $T_h = 0.85$ et w = 10 pour les descripteurs couleur, orientation et mouvement. Le tableau 3.8 présente les résultats de la recherche par l'exemple et permet de comparer la recherche effectuée au premier niveau de résolution de celle effectuée au dernier niveau. Nous avons également fourni le nombre d'images (ou segments) des différents niveaux pour avoir une idée de la réduction du résumé.

Finalement, nous pouvons constater que la recherche peut être effectuée en utilisant la hiérarchie. En effet, les résultats restent relativement stables que la recherche soit réalisée au premier ou quatrième niveau. Nous pouvons aussi remarquer que plus la réduction est élevée entre les niveaux de hiérarchie, plus il est difficile de retrouver les segments. Par exemple, pour la combinaison floue (Règles 1) des descripteurs couleur, orientation et mouvement sur la vidéo « Série », nous obtenons 94% de segments retrouvés au premier niveau de résolution pour un $\alpha = 3$ alors que nous avons seulement 81% des segments retrouvés au quatrième niveau. Ceci s'explique par le nombre d'images qui diminuent de 44 à 7 images entre ces deux niveaux.

3.6 Conclusion

Nous avons présenté une nouvelle méthode de résumé hiérarchique selon un ou plusieurs descripteurs. Trois nouveaux descripteurs flous ont été élaborés et ont la propriété d'être très compacts (11 composantes pour la couleur, 5 pour l'orientation et 3 pour le mouvement). Une mesure de similitude entre ces descripteurs a été définie à partir d'un système d'inférence floue et a permis la comparaison des images. La méthode de création de résumé de vidéo est basée sur deux étapes : segmentation des vidéos (micro-segmentation) et regroupement des segments par similarité avec contrainte temporelle (macro-segmentation). Notre méthode effectue une segmentation homogène à partir d'un ou plusieurs descripteurs. Cette segmentation constitue le niveau le plus fin de la hiérarchie. Puis, un algorithme de regroupement par similarité avec contrainte temporelle est appliqué pour réduire la taille du résumé de vidéo et fournir une visualisation rapide à l'utilisateur. Cet algorithme présente l'avantage de regrouper des segments similaires et proches temporellement puis par les itérations successives, de regrouper des segments de plus en plus éloignés et de moins en moins similaires. Ce procédé permet d'adapter la longueur du résumé en fonction de la demande de l'utilisateur.

Une application du résumé de vidéo, la recherche par l'exemple, est réalisée pour vérifier

TAB. 3.8 – Recherche par l'exemple sur chacune des 4 vidéos suivant deux niveaux de résolution : premier et quatrième niveaux de hiérarchie de résumé de vidéo.

Descripteur		couleur et orientation				couleur, orientation et mouvement)			
Combinaison		floue (Règles1)		floue (Règles2)		floue (Règles1)		floue (Règles2)	
	α	niv. 1	niv. 4	niv. 1	niv. 4	niv. 1	niv. 4	niv. 1	niv. 4
Documentaire	1	80	80	75	75	76	72	78	78
	2	95	95	94	94	89	82	94	94
	3	99	99	98	98	91	83	99	98
Nombre d'images		51	30	51	30	43	5	48	29
Journal	1	95	95	94	94	95	84	93	93
	2	98	98	99	99	97	84	100	100
	3	99	99	99	99	97	85	100	100
Nombre d'images		72	44	67	45	73	11	70	46
Série	1	83	81	78	77	81	72	77	77
	2	89	87	88	87	92	80	91	91
	3	94	92	95	93	94	81	93	93
Nombre d'images		46	18	42	18	44	7	45	20
Générique	1	88	86	85	80	81	71	83	83
	2	97	96	95	91	89	78	95	94
	3	100	98	99	95	95	82	98	97
Nombre d'images		66	34	66	35	64	8	66	29

la performance du résumé proposé. Cette étude est effectuée au premier niveau de résolution du résumé (segmentation des vidéos). Nous avons montré que les différents segments auxquels appartiennent les images requêtes sont correctement retrouvés. Afin d'accélérer la recherche par l'exemple (en diminuant les comparaisons), nous avons également vérifié que le résumé hiérarchique est efficace pour retrouver les segments.

La combinaison des descripteurs permet d'améliorer les résultats de la segmentation et de la recherche par l'exemple. L'intérêt de pouvoir utiliser un système d'inférence floue est la conception de règles qui permettent de combiner les similarités suivant le souhait de l'expert humain. Nous l'avons comparé à la combinaison linéaire et nous avons montré que, si les trois descripteurs sont employés, alors les résultats de la segmentation et de la recherche par l'exemple sont plus satisfaisants avec la combinaison floue qu'avec la combinaison linéaire. Le descripteur mouvement est par exemple moins discriminant que les deux autres (couleur et orientation) et donc ne doit pas être pondéré de la même façon lors de la combinaison linéaire. En revanche, lorsque les descripteurs couleur et orientation sont utilisés, la combinaison linéaire ou floue améliore les résultats par rapport à ceux fournis par les descripteurs pris séparément. Les deux combinaisons (linéaire et floue) ont des résultats relativement proches. Par contre, lorsque les descripteurs couleur et mouvement sont utilisés, les résultats obtenus montrent l'intérêt de la combinaison floue.

La recherche par l'exemple pourrait aussi être améliorée. Au lieu de fournir une image requête, l'utilisateur pourrait donner des informations sur les segments attendus. Par exemple, une requête peut être faite en renseignant uniquement sur les proportions des couleurs élémentaires souhaitées, sur le degré d'activité voulu et sur les orientations des contours à privilégier (plutôt des verticales ou horizontales). Ce genre de requête peut facilement être traduit avec les descripteurs que nous avons développés.

Cependant, la méthode de résumé proposée ici repose sur des descripteurs de bas niveau. Dans la suite de ce manuscrit, nous allons étudier et fournir une description de plus haut niveau du contenu des vidéos pour créer le résumé.

Chapitre 4

Classification des mouvements de caméra

Pour obtenir une description plus sémantique du contenu des vidéos, nous proposons ici d'étudier une nouvelle caractéristique, le mouvement de caméra. Nous développons une méthode de classification des mouvements de caméra basée sur le Modèle des Croyances Transférables (MCT). Elle sera à l'origine d'une nouvelle méthode de résumé détaillée dans le chapitre suivant.

Sommaire

4.1 Introduction

Dans le chapitre 3, nous avons introduit une méthode de résumé à partir d'index bas niveau. L'analyse du contenu a été effectuée par des caractéristiques de bas niveau telles que la couleur, la texture ou le mouvement auxquels il a été délicat d'attribuer un sens. Devant les limites de la description bas niveau des vidéos, nous avons étudié une nouvelle caractéristique, le mouvement de caméra, pour indexer de manière plus efficace le contenu des vidéos. L'extraction et l'identification des mouvements de caméra sont une étape essentielle pour analyser le contenu et fournir une meilleure représentation des vidéos.

A partir du mouvement de caméra, de nombreuses informations sémantiques peuvent être déduites comme l'activité d'une scène. Par exemple, un film d'action comporte de nombreuses scènes avec de forts mouvements de caméra pour donner du rythme. De même, par la manière de filmer une scène, notre regard peut aussi être dirigé. Un zoom avant va focaliser notre attention sur une zone bien précise de la scène. La connaissance du mouvement de caméra peut aussi être exploitée pour séparer les objets en mouvement du fond et peut être utilisée dans les algorithmes de segmentation. Cet index est ainsi un outil puissant pour extraire le contexte sémantique de la scène.

Le mouvement de caméra a été employé dans diverses applications. En utilisant des données *a priori* sur le contenu des vidéos ainsi que la classification des mouvements de caméra, des scènes de sport ont pu être étiquetées comme les divers coups du cricket [Laz02], les différentes phases du football américain [Laz03] et les séquences de basketball [Tan00]. De la même manière, en analysant la statistique des mouvements de caméra, Takagi et al. [Tak03] démontrent que le mouvement de caméra est une signature suffisante pour différencier les activités sportives et les classer. Le mouvement de caméra peut aussi être utilisé pour la segmentation de la vidéo en plans [Qi03a, Bou99], la création de résumé de vidéo [Por03, Fau04] ainsi que dans des modèles d'attention visuelle [Che05b]. De plus, les demandes pour l'archivage de vidéos sont devenues si fortes que les expérimentations de TREC Video [Tre05] comprennent en 2005 une nouvelle tâche concernant la classification des mouvements de caméra, sachant que l'objectif de TREC Video est de promouvoir le progrès dans la recherche basée sur le contenu.

La plupart des approches supposent que le mouvement dominant provient du mouvement de caméra. Un modèle paramétrique est souvent utilisé pour le représenter et différents algorithmes ont été proposés pour estimer les paramètres soit dans le domaine compressé [Kim04, Tan00, Gil04] soit dans le domaine non compressé [Por03, Bou99, Fau04]. D'autres méthodes obtiennent le mouvement de caméra directement en analysant les vecteurs du flux MPEG [Che04, Zhu05, Lee02, Dua04]. Néanmoins, beaucoup de ces approches attribuent un type de mouvement de caméra à partir des paramètres extraits localement (soit entre deux images successives soit à partir des images prédites par le flux MPEG) en utilisant un algorithme d'apprentissage [Che04, Dua04], une stratégie de seuillage [Kim04, Zhu05] ou un algorithme de regroupement suivant des prototypes [Lee02] (template-matching algorithm). Une étape de filtrage est parfois ajoutée pour obtenir des mouvements plus consistants [Zhu05]. De plus, très peu de ces méthodes quantifient le mouvement identifié. Par exemple, un zoom est détecté mais l'agrandissement n'est pas défini, ce qui peut être un inconvénient dans certaines applications.

Comme alternative aux diverses approches présentées, nous proposons une méthode originale de classification des mouvements de caméra basée sur le Modèle des Croyances Transférables (MCT). Cette théorie, très utilisée en fusion de données, est une structure adaptée pour traiter des données imprécises et incertaines, pour combiner différentes sources d'information et pour gérer le conflit entre les sources. Elle sera exploitée pour combiner les paramètres d'un modèle de mouvement. L'objectif de la classification est de pouvoir étiqueter de manière robuste une vidéo suivant les trois grandes familles de mouvement de caméra que sont la translation (pan et/ou tilt), le zoom et l'absence de mouvement. La translation correspond soit à la rotation de la caméra autour de l'axe vertical et/ou horizontal soit au suivi de la caméra le long de l'axe vertical et/ou horizontal. Le zoom conduit à l'agrandissement ou à la réduction d'une partie de l'image. Enfin, l'absence de mouvement appelée par simplification, « statique » est une scène tournée à caméra fixe où les objets y figurant peuvent être mobiles. Ces trois mouvements de caméra sont similaires à ceux de TREC Video 2005 avec la détermination de la translation horizontale, de la translation verticale et du zoom. A partir d'un modèle paramétrique du mouvement, notre approche consiste à estimer le mouvement de caméra au niveau de chaque paire d'images, puis à analyser le mouvement de caméra au niveau de segment (i.e. sur un ensemble d'images). Bien que l'estimation du mouvement utilisé ici soit réalisée dans le domaine non compressé, notre méthode peut être adaptée dans le domaine compressé comme dans [Gil04, Kim04, Tan00]. En effet, les paramètres du modèle qui sont manipulés peuvent être indifféremment estimés dans le domaine compressé ou non compressé. Les contributions principales de ce travail résident dans l'identification des mouvements de caméra qui repose sur un certain nombre de règles : combinaison conçue pour éviter d'identifier les mouvements de caméra avec de faibles amplitudes, filtrage selon le MCT pour s'assurer de la cohérence temporelle des mouvements, et analyse au niveau des segments pour conserver les mouvements avec une durée et une amplitude conséquentes.

La suite du chapitre est organisée de la façon suivante. La section 4.2 présente l'architecture de la méthode de classification et de quantification des mouvements de caméra. L'estimation du modèle paramétrique est exposée dans la section 4.3. Après un bref rappel sur le Modèle des Croyances Transférables dans la section 4.4, nous détaillons notre méthode de classification des mouvements de caméra dans la section 4.5. Nous exposons dans la section 4.6 la manière de quantifier les mouvements détectés. La section 4.7 présente les résultats de la méthode de classification. Nous terminons avec nos conclusions dans la section 4.8.

4.2 Architecture du système

L'architecture du système est présentée dans la figure 4.1 et se compose de trois phases : extraction des paramètres du mouvement, classification des mouvements de caméra et quantification des mouvements de caméra. Le coeur du travail est la phase de classification qui est divisée en trois étapes. La première étape est conçue pour convertir les paramètres du modèle affine de mouvement en valeurs symboliques. Cette représentation a pour but de faciliter la définition de règles pour combiner les données et fournir des fonctions de masse au niveau de chaque image suivant les différents mouvements de caméra. Un filtrage des fonctions de masse selon le MCT est réalisé et contribue à améliorer la cohérence temporelle des masses de croyance. La deuxième phase effectue une séparation entre les images statiques et dynamiques (zoom, translation). Enfin, dans la troisième étape, l'intégration temporelle des mouvements est réalisée et permet d'étudier le mouvement au niveau du segment (en regroupant des images

ayant une certaine croyance en un mouvement donné). L'avantage de cette analyse est de préserver seulement les mouvements avec des amplitudes et des durées significatives. La phase de quantification est ensuite réalisée en extrayant différentes caractéristiques sur chaque segment de la vidéo contenant un type de mouvement de caméra identifié. Par exemple, un mouvement de zoom est caractérisé par un coefficient d'agrandissement.

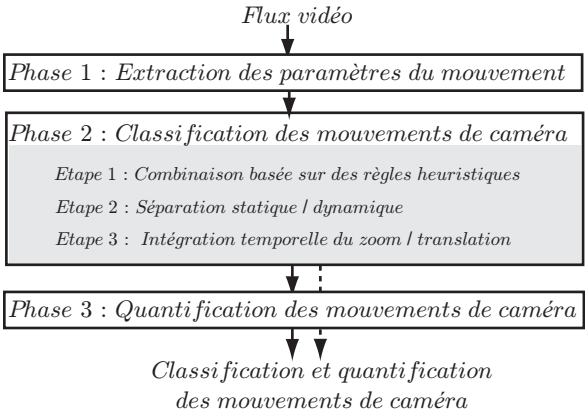


FIG. 4.1 – Architecture de la méthode de classification et de quantification des mouvements de caméra.

4.3 Modèle affine du mouvement de caméra

Le mouvement dominant, supposé provenir du mouvement de caméra, est estimé entre deux images successives par un modèle paramétrique. Il caractérise le champ de vecteurs vitesse reliant les pixels d'une image à l'image suivante. Le modèle considéré ici est un modèle affine permettant de décrire 5 types classiques de mouvement de caméra : zoom, rotation, translation horizontale, translation verticale et statique. Le champ de vecteurs vitesse s'exprime en fonction de la position du pixel $p_i = (x_i, y_i)$ de l'image $I(p_i, t)$ à l'instant t selon l'expression suivante :

$$v_x(p_i) = c_1 + a_1 \cdot x_i + a_2 \cdot y_i$$
$$v_y(p_i) = c_2 + a_3 \cdot x_i + a_4 \cdot y_i$$

où $\theta_t = (c_1, c_2, a_1, a_2, a_3, a_4)$ sont les paramètres à estimer. Le pixel à la position $p_i = (x_i, y_i)$ dans l'image $I(p_i, t)$ se déplace à la position $p_i^{'} = (x_i^{'}, y_i^{'})$ dans l'image suivante $I(p_i^{'}, t+1)$ selon la relation suivante :

$$x_i^{'} = x_i + v_x(p_i)$$
$$y_i^{'} = y_i + v_y(p_i)$$

Le vecteur de p_i à $p_i^{'}$ est égal au vecteur vitesse puisque l'intervalle de temps entre deux images successives vaut par convention un. La détermination des coefficients du modèle est effectuée par le logiciel Motion2D [Odo95], outil développé à l'IRISA/INRIA Rennes par l'équipe Vista. Il réalise une estimation robuste, multi-échelle du mouvement dominant à partir des gradients spatio-temporels de l'intensité des images. Néanmoins, cette estimation du mouvement n'est pas une identification exacte de celui-ci. Les erreurs d'estimation ne devront

pas perturber la classification des mouvements. A ces limites s'ajoute une imprécision liée au tournage (par exemple, caméra à l'épaule occasionnant des vibrations) susceptible de générer des mouvements de caméra non souhaités et peu visibles appelés mouvements secondaires. Ces mouvements se superposent aux mouvements perçus et possèdent un ordre de grandeur pouvant être localement similaire aux mouvements recherchés, mais ils ne devront pas être identifiés.

Les paramètres en sortie $\theta_t = (c_1, c_2, a_1, a_2, a_3, a_4)$ de l'estimateur décrivent un champ de mouvement qui n'est pas en relation directe avec les mouvements recherchés. Il est ainsi nécessaire d'identifier les classes de mouvement en fonction des 6 paramètres de l'estimateur. Différents champs de mouvement sont présentés sur la figure 4.2 où seuls certains paramètres sont différents de zéro :

- Champs de translation avec $(c_1, c_2) \neq 0$ (Fig. 4.2.a-b)
- Champ divergent avec $\frac{1}{2}(a_1 + a_4) \neq 0$ (Fig. 4.2.c)
- Champ rotationnel avec $\frac{1}{2}(a_2 - a_3) \neq 0$ (Fig. 4.2.d)

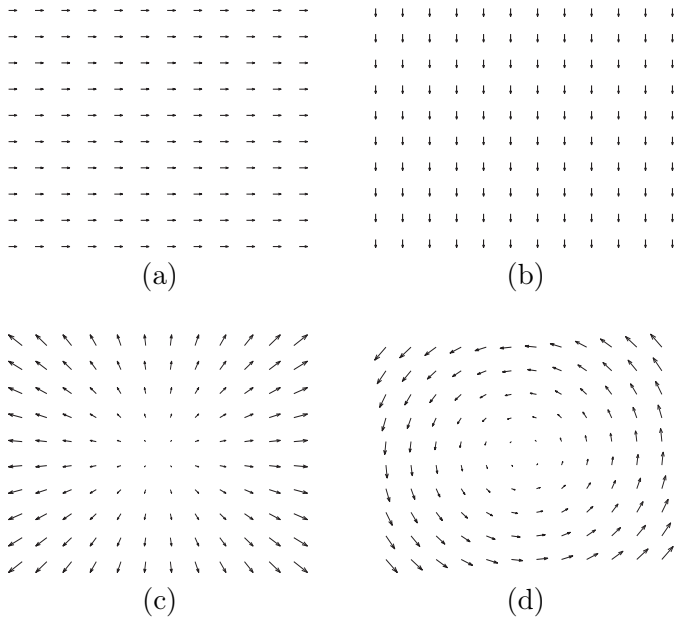


FIG. 4.2 – Exemple de champs de vecteurs vitesse : (a) et (b) champs de translation uniforme (horizontal ou vertical), (c) champ divergent correspondant à un zoom et (d) champ rotationnel.

Comme le montre la figure 4.2, l'information fournie par certains paramètres du modèle est spécifique à un mouvement et elle est utilisée pour détecter les mouvements de caméra. Néanmoins, avant d'utiliser ces coefficients, nous réalisons un filtrage moyenneur de taille L_1 sur les paramètres θ_t pour réduire le bruit et les erreurs d'estimation. Un exemple de l'estimation des paramètres est montré sur la figure 4.3 où la séquence contient un zoom avant. Nous pouvons observer qu'effectivement les paramètres sont bruités et parfois erronés (fortes impulsions entre les images 400 et 450).

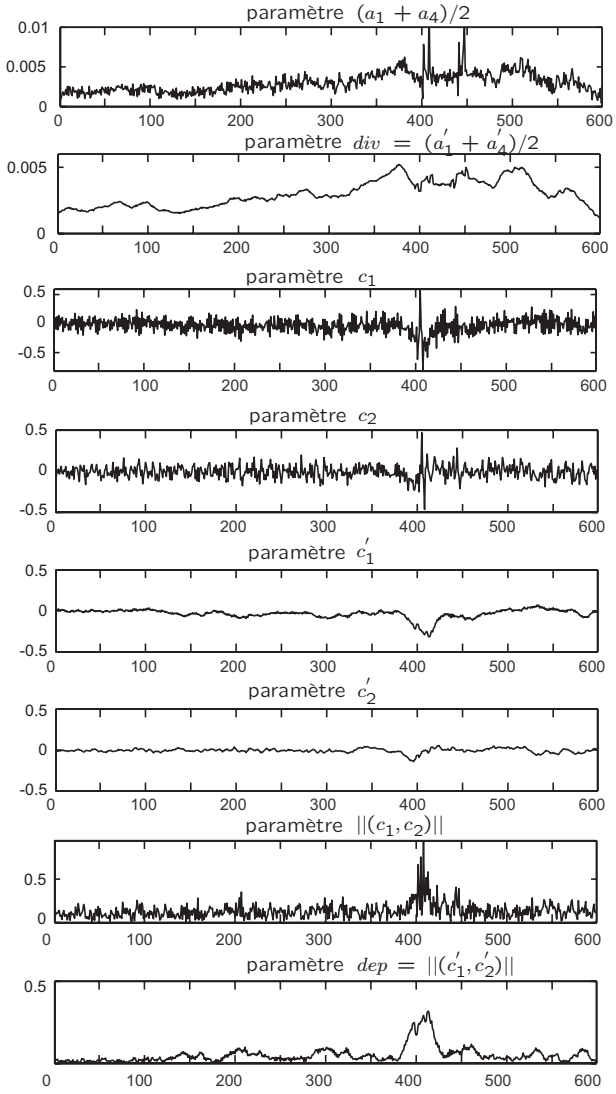


FIG. 4.3 – Evolution au cours du temps des paramètres estimés sur une séquence d'images contenant un zoom avant. Pour ce qui est du mouvement de zoom, le seul paramètre qui doit être plus grand que 0 est $(a_1 + a_4)/2$ et les paramètres c_1 et c_2 sont sensés être nuls. Pour atténuer le bruit et les erreurs d'estimation (fortes impulsions sur les courbes), les paramètres sont filtrés sur une fenêtre de taille $L_1 = 13$ (environ une demi-seconde).

A partir des paramètres filtrés $\theta_t^{'} = (c_1^{'}, c_2^{'}, a_1^{'}, a_2^{'}, a_3^{'}, a_4^{'})$, le déplacement de la caméra $dep(t)$ et le divergent $div(t)$ entre deux images successives $I(p_i, t)$ et $I(p_i, t+1)$ sont définis ainsi que le déplacement total $dep(t_o, t_f)$ et le chemin parcouru $ch(t_o, t_f)$ entre deux instants t_o et t_f :

$$\begin{aligned}
\overrightarrow{dep}(t) &= (c_1^{'}(t), c_2^{'}(t)) \\
dep(t) &= \left\| \overrightarrow{dep}(t) \right\| \\
div(t) &= \tfrac{1}{2}(a_1^{'}(t) + a_4^{'}(t)) \\
dep(t_o, t_f) &= \left\| \sum_{j=t_o}^{t_f - 1} \overrightarrow{dep}(j) \right\| \\
ch(t_o, t_f) &= \sum_{j=t_o}^{t_f - 1} \left\| \overrightarrow{dep}(j) \right\|
\end{aligned}$$

La figure 4.4 illustre la distinction entre le déplacement total et le chemin parcouru. Le déplacement total correspond au déplacement en ligne droite de la position initiale à la position finale alors que le chemin parcouru est le chemin original et correspond à l'intégration de tous les déplacements entre les instants d'échantillonnage.

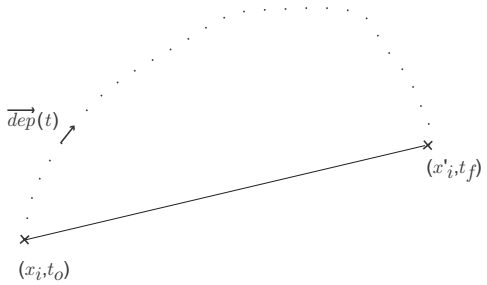


FIG. 4.4 – Exemple du déplacement et du chemin parcouru entre la position initiale (x_i, t_0) et la position finale (x'_i, t_f). La ligne en pointillée est le chemin parcouru alors que la ligne droite représente le déplacement total entre les instants t_0 et t_f.

Selon les amplitudes des variables *div* et *dep*, les différents mouvements de caméra peuvent être extraits. Une translation (respectivement un zoom) est détectée si le déplacement (respectivement le divergent) est élevé. Quand un léger zoom et une forte translation se produisent simultanément, le zoom n'est pas ou peu visible et donc il ne devra pas être signalé. De la même manière, seul le zoom est conservé en présence d'un fort zoom et d'une faible translation. Afin de satisfaire à ces règles, les variables ont besoin d'être converties en valeurs linguistiques pour être combinées.

Avant de décrire notre approche de classification des mouvements, la section suivante rappellera les fondements du Modèle des Croyances Transférables.

4.4 Modèle des Croyances Transférables

La théorie de l'évidence (également appelée théorie des fonctions de croyance ou théorie de Dempster-Shafer) a été introduite par A. Dempster et G. Shafer [Sha76]. Elle a ensuite été formalisée par P. Smets sous le nom de Modèle des Croyances Transférables (MCT).

Cette théorie a été utilisée dans de nombreuses applications. Par exemple, elle a été employée dans l'analyse du comportement et des actions de personnes : reconnaissance de sauts d'athlètes dans des vidéos [Ram06], reconnaissance des postures du corps humain [Gir05], reconnaissance des expressions faciales [Ham05]. Elle a également été appliquée en imagerie médicale pour détecter un cancer de la peau (Mélanome) [Van99b, Van99a]. Dans [Cap02, Cap04], un algorithme de segmentation est développé pour localiser des tumeurs ou des lésions au niveau du cerveau dans les images à résonance magnétique. Megherbi et al. [Meg05] ont aussi utilisé cette théorie pour associer des données dans un contexte de suivi de cibles multiples en traitant le problème d'apparition et de disparition des cibles. Nous allons utiliser la théorie de l'évidence à la reconnaissance des mouvements de caméra dans les vidéos.

Nous allons présenter un bref résumé de ce modèle, une description plus détaillée pourra être trouvée dans [Sme94].

4.4.1 Cadre de discernement

L'ensemble des N solutions à un problème donné est appelé cadre de discernement $\Omega = \{H_1, \cdots, H_N\}$. Ces solutions ou hypothèses doivent être exhaustives et exclusives entre elles et l'une d'elles est la solution au problème posé.

A partir du cadre de discernement, l'ensemble des sous-ensembles de Ω noté 2^Ω est défini. Il se compose de l'ensemble des hypothèses singletons et de toutes les combinaisons possibles entre ces hypothèses singletons.

$$2^\Omega = \{A/A \subseteq \Omega\} = \{\emptyset, H_1, \cdots, H_N, H_1 \cup H_2, \cdots, \Omega\}$$

Notre démarche s'inscrit dans un contexte de monde fermé. Ceci implique que la solution est obligatoirement dans le cadre de discernement. En revanche, dans un contexte de monde ouvert, la solution n'est pas forcément contenue dans le cadre de discernement.

4.4.2 Fonction de masse

A chaque sous-ensemble de Ω peut être affecté une mesure de confiance qui représente la croyance en cette proposition. Une fonction de masse (Basic Belief Assignment (BBA)) se définit par :

$$\begin{array}{rccl} m & : & 2^\Omega & \rightarrow & [0,1] \\ & & A_i & \rightarrow & m(A_i) \end{array}$$

où $m(A_i)$ représente le degré de croyance qui est attribué exactement à la proposition A_i. La détermination d'une BBA par un capteur ou une source d'information dans un contexte de monde fermé est contrainte par les règles suivantes :

$$\begin{array}{l} m(\emptyset) = 0 \\ \sum_{A \in 2^\Omega} m(A) = 1 \end{array}$$

Les sous-ensembles $A \subseteq \Omega$ où $m(A) > 0$ sont appelés éléments focaux de m.

4.4.3 Combinaison

Soient m_1 et m_2 deux BBA définies sur le même cadre de discernement et attribuées respectivement par une source 1 et une source 2. Suivant les applications, deux combinaisons sont possibles : la combinaison conjonctive $m_1 \textcircled{\cap} m_2(A_i)$ et disjonctive $m_1 \textcircled{\cup} m_2(A_i)$.

$$m_1 \textcircled{\cap} m_2(A_i) = \sum_{A_j \cap A_k = A_i} m_1(A_j) \cdot m_2(A_k)$$
$$m_1 \textcircled{\cup} m_2(A_i) = \sum_{A_j \cup A_k = A_i} m_1(A_j) \cdot m_2(A_k)$$

La combinaison conjonctive (respectivement disjonctive) s'interprète comme un « et » (respectivement « ou ») logique. L'intérêt de ces combinaisons est de pouvoir les utiliser dans des règles logiques.

4.4.4 Décision

A partir d'une fonction de masse, une transformation a été proposée par P. Smets [Sme94] pour obtenir une fonction de probabilité sur l'ensemble Ω appelée fonction de probabilité pignistique et définie par :

$$BetP^{\Omega}(A) = \sum_{B \subseteq \Omega} \frac{m^{\Omega}(B)}{1 - m^{\Omega}(\emptyset)} \frac{|A \cap B|}{|A|}, \ \forall A \subseteq \Omega \tag{4.1}$$

où $|A|$ est le cardinal de $A \subseteq \Omega$. Cette fonction peut être utilisée pour la prise de décision.

4.4.5 Produit Cartésien

Soient deux cadres de discernement distincts et disjoints Ω_1 et Ω_2, une fonction de masse peut être également définie en effectuant une combinaison conjonctive sur $\Omega = \Omega_1 \times \Omega_2$ comme indiquée ci-dessous :

$$m_1 \textcircled{\cap} m_2(A \times B) = m_1(A) \cdot m_2(B) \ \forall A \subseteq \Omega_1, \ \forall B \subseteq \Omega_2 \tag{4.2}$$

L'intérêt du produit cartésien est de pouvoir exploiter le MCT même lorsque les cadres de discernement sont disjoints et donc non compatibles. Supposons un exemple où le *sexe* des personnes {masculin, féminin} et la *couleur des yeux* {marron, bleu} sont recherchés. Les hypothèses *sexe* et *couleur des yeux* ne sont pas exclusives entre elles. Le produit cartésien sur ces deux espaces peut être introduit et un raisonnement peut alors être appliqué.

4.5 Classification des mouvements de caméra

La classification des mouvements de caméra consiste à segmenter une vidéo suivant les mouvements de caméra. Notre méthode, qui repose sur le MCT, doit identifier les trois types de mouvement de caméra que sont la translation, le zoom et l'absence de mouvement. Il s'agit de reconnaître aussi bien les mouvements courts et de fortes amplitudes que les mouvements très longs et de faibles amplitudes, et d'éviter les fausses détections dues à une mauvaise estimation. Le principe de la classification des mouvements de caméra est présenté sur la figure 4.5. Cette méthode se divise en trois étapes : *combinaison basée sur des règles heuristiques*, *séparation statique/dynamique* et *intégration temporelle zoom-translation*. La suite de cette section décrit chacune des étapes de la méthode.

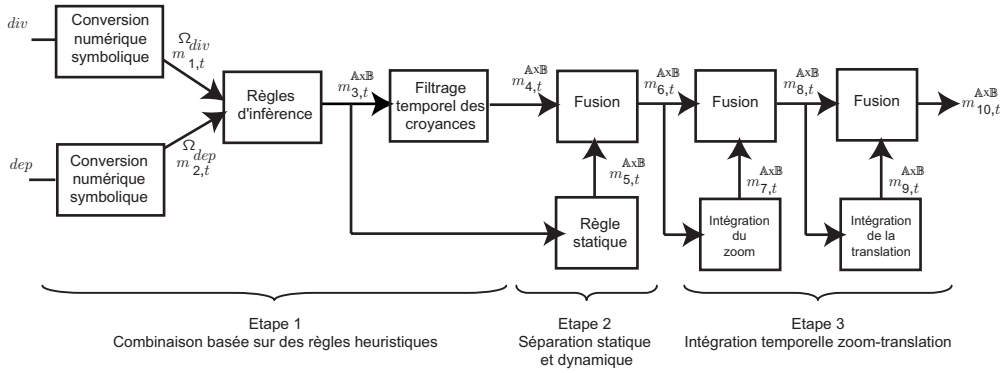


FIG. 4.5 – Principe de la phase de classification des mouvements de caméra.

4.5.1 Combinaison basée sur des règles heuristiques

La première étape consiste à convertir les paramètres du modèle affine en valeurs symboliques décrivant les mouvements recherchés. A partir de ces variables, nous établissons des règles heuristiques pour déterminer au niveau de chaque image des fonctions de masse sur les différents mouvements de caméra. Puis un filtrage temporel de ces fonctions de masse est effectué dans le but d'assurer la cohérence temporelle des masses de croyance sur un voisinage.

4.5.1.1 Conversion numérique-symbolique

Les variables numériques *div* et *dep* sont transformées en valeurs symboliques : faible (F), moyen (M), grand (G) et très grand (TG). Cette représentation symbolique est utilisée pour élaborer facilement des règles sur les mouvements de caméra. A chaque terme linguistique ou groupe (faible, moyen, grand et très grand) est associée une masse de croyance. Après une étude expérimentale, les fonctions d'appartenance qui permettent d'attribuer les masses ont été fixées comme indiqué sur la figure 4.6. En ce qui concerne la description symbolique du divergent, elle est effectuée à partir de la valeur absolue du divergent (4.6.a). En effet, la valeur absolue informe sur l'amplitude du zoom alors que le sens du zoom est obtenu par le signe du divergent. Nous pouvons aussi remarquer que la dynamique de $|div|$ est plus réduite que celle de *dep*. Finalement, les fonctions de masse (ou BBA) pour les variables *div* et *dep* sont définies respectivement sur le cadre de discernement $\Omega_{div} = \{F, M, G, TG\}$ et $\Omega_{dep} = \{F, M, G, TG\}$. La combinaison de ces fonctions de masse conduira à la détection des mouvements de caméra.

4.5.1.2 Règles d'inférence

La classification des mouvements de caméra est effectuée à partir de règles heuristiques. La construction des règles repose sur le Modèle des Croyances Transférables (MCT) qui est un outil adapté pour intégrer des mécanismes d'inférence.

Soient $\mathbb{A} = \{T, \overline{T}\}$ le cadre de discernement du mouvement de translation et $\mathbb{B} = \{Z, \overline{Z}\}$ le cadre de discernement du mouvement de zoom où T (resp Z) est une hypothèse sur la présence de la translation (resp zoom) et $\overline{T}$ (resp $\overline{Z}$) est une hypothèse sur l'absence de la translation (resp absence de zoom). L'identification du mouvement peut être réalisée en effectuant le produit cartésien des espaces $\mathbb{A} \times \mathbb{B}$. Par exemple, si une image appartient à la classe $\{(\overline{T}, \overline{Z})\}$ alors elle sera considérée comme statique. En revanche, si l'image appartient à

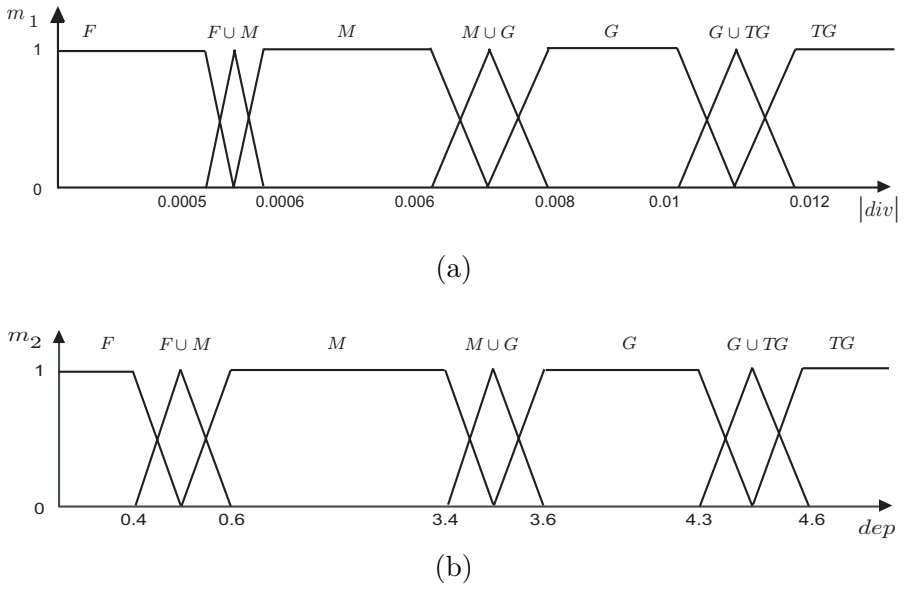


FIG. 4.6 – Définition des BBA (a) pour le divergent et (b) pour le déplacement.

la classe $\{(T, Z)\}$ alors elle aura un mouvement de caméra à la fois de translation et de zoom.

L'attribution de la croyance sur l'espace produit $\mathbb{A} \times \mathbb{B}$ est effectuée à partir des règles que nous avons définies :

- Si *div* est faible et *dep* est faible alors le mouvement de caméra est statique $\{(\overline{T}, \overline{Z})\}$.
- Si *div* est moyen et *dep* est grand alors le mouvement est la translation $\{(T, \overline{Z})\}$.
  ...
- Si *div* est très grand et *dep* est très grand alors le mouvement est la translation et le zoom $\{(T, Z)\}$.

L'ensemble des règles Ⓡ que nous avons défini pour la classification des mouvements de caméra est résumé dans le tableau 4.1. Par exemple, si *div* est grand et *dep* est moyen alors le mouvement détecté est le zoom et donc la masse de croyance sur l'ensemble produit $\mathbb{A} \times \mathbb{B}$ est assignée à $\{(\overline{T}, Z)\}$. Nous pouvons remarquer des propositions sur plusieurs mouvements comme par exemple $\{(\overline{T}, \overline{Z}), (\overline{T}, Z)\}$ qui signifie l'absence de translation et l'ignorance sur la présence du zoom, ce qui correspond à une proposition sur « statique ou zoom ». De même, $\{(\overline{T}, \overline{Z}), (T, \overline{Z})\}$ signifie l'absence de zoom et l'ignorance sur la présence de la translation, ce qui correspond à une proposition sur « statique ou translation » alors que $\mathbb{A} \times \mathbb{B}$ est une ignorance totale sur le mouvement présent. Ainsi les deux BBA $m_1^{\Omega_{div}}$ et $m_2^{\Omega_{dep}}$ définies sur les cadres de discernement Ω_{div} et Ω_{dep} sont combinées suivant l'opérateur Ⓡ qui décrit les règles mentionnées dans le tableau 4.1. Cette combinaison aboutit à la définition d'une nouvelle BBA $m_3^{\mathbb{A}\times\mathbb{B}} = m_1^{\Omega_{div}} Ⓡ m_2^{\Omega_{dep}}$ sur l'espace produit $\mathbb{A} \times \mathbb{B}$ qui caractérise directement la croyance sur les mouvements de caméra. Nous pouvons également constater que les règles sont conçues pour éviter dans la mesure du possible les mouvements secondaires. Par exemple, si le déplacement est très grand et le divergent est grand alors le mouvement de zoom sera négligé et seul le mouvement de translation sera pris en considération.

Afin de détailler la méthode, nous proposons de traiter l'exemple dans lequel pour une image donnée, les masses non nulles de $m_1^{\Omega_{div}}$ et $m_2^{\Omega_{dep}}$ sont les suivantes :

TAB. 4.1 – Attribution d'une BBA en fonction des valeurs du divergent et du déplacement selon les règles ®

		div			
		F	M	G	TG
dep	F	$\{(\overline{T},\overline{Z})\}$	$\{(\overline{T},\overline{Z}),(\overline{T},Z)\}$	$\{(\overline{T},Z)\}$	$\{(\overline{T},Z)\}$
	M	$\{(\overline{T},\overline{Z}),(T,\overline{Z})\}$	$\mathbb{A}\times\mathbb{B}$	$\{(\overline{T},Z)\}$	$\{(\overline{T},Z)\}$
	G	$\{(T,\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(T,Z)\}$	$\{(\overline{T},Z)\}$
	TG	$\{(T,\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(T,Z)\}$

$$\begin{array}{lcllcl} m_1^{\Omega_{div}}(F) & = & 0.7 & m_2^{\Omega_{dep}}(M\cup G) & = & 0.2 \\ m_1^{\Omega_{div}}(F\cup M) & = & 0.3 & m_2^{\Omega_{dep}}(G) & = & 0.8 \end{array}$$

La BBA $m_3^{\mathbb{A}\times\mathbb{B}}$ a alors 3 éléments focaux :

$$\begin{array}{lcl} m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(\overline{T},\overline{Z}),\ (T,\overline{Z})\right\}\right) & = & m_1^{\Omega_{div}}(F)\cdot m_2^{\Omega_{dep}}(M\cup G)=0.14 \\ m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(T,\overline{Z})\right\}\right) & = & m_1^{div}(F)\cdot m_2^{\Omega_{dep}}(G)+m_1^{\Omega_{div}}(F\cup M)\cdot m_2^{\Omega_{dep}}(G) \\ & = & 0.80 \\ m_3^{\mathbb{A}\times\mathbb{B}}\left(\mathbb{A}\times\mathbb{B}\right) & = & m_1^{\Omega_{div}}(F\cup M)\cdot m_2^{\Omega_{dep}}(M\cup G)=0.06 \end{array}$$

Finalement, à chaque image d'indice t de la vidéo correspond une BBA $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ sur l'espace $\mathbb{A}\times\mathbb{B}$. Les figures 4.7, 4.8, 4.9 présentent respectivement des exemples de séquences filmées à caméra fixe, ayant un mouvement de zoom arrière et de translation. Ces figures permettent de mettre en évidence les combinaisons des paramètres *div* et *dep*. Par exemple, bien que la séquence de la figure 4.7 soit filmée à caméra fixe, le paramètre *div* n'a pas toujours une amplitude faible et comporte parfois une amplitude moyenne. Comme le paramètre *dep* est toujours d'amplitude faible, deux fonctions de masse ont une valeur non nulle et représentent une croyance sur deux ensembles distincts de mouvement de caméra. Lorsque le divergent *div* est faible, une masse de croyance est portée sur l'ensemble « statique » alors qu'une masse est attribuée à l'ensemble « statique ou zoom » lorsque le divergent *div* devient moyen.

4.5.1.3 Filtrage temporel des fonctions de masse

Le filtrage temporel des fonctions de masse a été introduit et repose sur l'hypothèse que le mouvement de caméra ne peut pas être très différent d'une image à l'autre. Si le cas apparaît, alors on considère que les mouvements éventuels sont possibles, sans pouvoir privilégier l'un plutôt que l'autre. Ce filtrage selon le MCT, a pour effet de renforcer ou d'atténuer la présence d'un mouvement de caméra suivant la cohérence des croyances autour du voisinage de l'image traitée. Il s'agit d'ajouter du doute en réallouant la croyance sur l'union des mouvements dans le cas où les sources temporellement voisines délivrent des informations différentes. Par exemple, si les fonctions de masse attribuent une forte croyance sur deux mouvements différents pour un voisinage donné alors le filtrage redistribue cette croyance sur l'union des mouvements. Le filtre est réalisé par la combinaison disjonctive des sources $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ sur une fenêtre temporelle de taille L_2. Une nouvelle BBA est alors obtenue :

$$m_{4,t}^{\mathbb{A}\times\mathbb{B}} = m_{3,t-(L_2-1)/2}^{\mathbb{A}\times\mathbb{B}} \textcircled{\cup} \ \ldots \ \textcircled{\cup}\ m_{3,t+(L_2-1)/2}^{\mathbb{A}\times\mathbb{B}} \tag{4.3}$$

L'intérêt de cette combinaison est d'accroître les plages de mouvement détectées par réaffectation des croyances sur un voisinage et donc de conserver la présence d'un mouvement en comblant les trous éventuels pouvant être générés par des erreurs d'estimation. De plus, la différence entre ce type de filtrage et le filtre moyenneur est que le filtre moyenneur est utilisé pour atténuer le bruit sur les paramètres du modèle affine alors que le filtrage des fonctions de croyance permet de remettre en cause la décision sur le mouvement potentiel d'une image en fonction de ces voisines, ce qui a aussi pour effet de rajouter du doute au niveau des transitions entre deux mouvements successifs.

Les figures 4.7, 4.8 et 4.9 illustrent le filtrage temporel des fonctions de croyance avec une fenêtre de taille $L_2 = L_1 = 13$. Dans la figure 4.8, nous pouvons remarquer sur la courbe $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ que le filtrage permet d'ajouter du doute quand les croyances passent d'une proposition de mouvement à une autre (image 192). Ce changement (image 192) peut être interprété comme une transition entre deux mouvements de caméra. Il est traité en redistribuant des masses sur l'union des ensembles (image 186 de $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$). L'exemple du filtrage sur la figure 4.9 est encore plus manifeste. Comme le paramètre *div* alterne entre faible et moyen et le paramètre *dep* est moyen, nous pouvons observer une succession de masses de croyance sur des ensembles différents. Pour éviter la détection de mouvements impulsifs, le filtrage reporte les masses sur l'ensemble $\mathbb{A} \times \mathbb{B}$ en raison du manque de cohérence des masses de croyance. Les deux étapes suivantes de la classification permettront de retrouver le mouvement de caméra.

Globalement, les règles et le filtrage correspondent à un traitement très « prudent » des informations laissant une large place au doute plutôt que d'imposer un avis tranché sur le mouvement de caméra.

4.5.2 Séparation statique/dynamique

Cette deuxième étape consiste à séparer les images statiques des images dynamiques (zoom, translation) en tenant compte du voisinage temporel des croyances attribuées localement par les règles heuristiques. En l'absence de mouvement de caméra, les paramètres estimés du modèle affine ont une amplitude faible voire nulle. Cette propriété n'est pas toujours vérifiée en raison du bruit provenant de l'estimateur ou de vibration de la caméra. Néanmoins celle-ci est au moins conservée sur un voisinage autour de l'image traitée. Ainsi une nouvelle BBA est définie et repose sur la règle suivante : si un certain nombre d'images autour de l'image étudiée ont une croyance sur l'hypothèse statique (respectivement dynamique) alors une masse sera attribuée à l'hypothèse statique (respectivement dynamique) pour l'image étudiée.

Soit le cadre de discernement $\Omega = \{S, D\}$ où S et D désignent respectivement une hypothèse statique et une hypothèse dynamique. L'ensemble Ω est un grossissement de $\mathbb{A} \times \mathbb{B}$ et réciproquement $\mathbb{A} \times \mathbb{B}$ est un raffinement de Ω.

Pour savoir si l'image t est plutôt statique ou plutôt dynamique, chaque BBA $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$ est transformée en une BBA $m_{3,t}^{\Omega}$ comme indiqué ci-dessous :

$$\begin{aligned} m_{3,t}^{\Omega}(S) &= m_{3,t}^{\mathbb{A}\times\mathbb{B}}(\{(\overline{T}, \overline{Z})\}) \\ m_{3,t}^{\Omega}(D) &= \textstyle\sum_{K\subseteq\mathbb{A}\times\mathbb{B}\setminus(\overline{T},\overline{Z})} m_{3,t}^{\mathbb{A}\times\mathbb{B}}(K) \\ m_{3,t}^{\Omega}(\Omega) &= 1 - m_{3,t}^{\Omega}(S) - m_{3,t}^{\Omega}(D) \end{aligned}$$

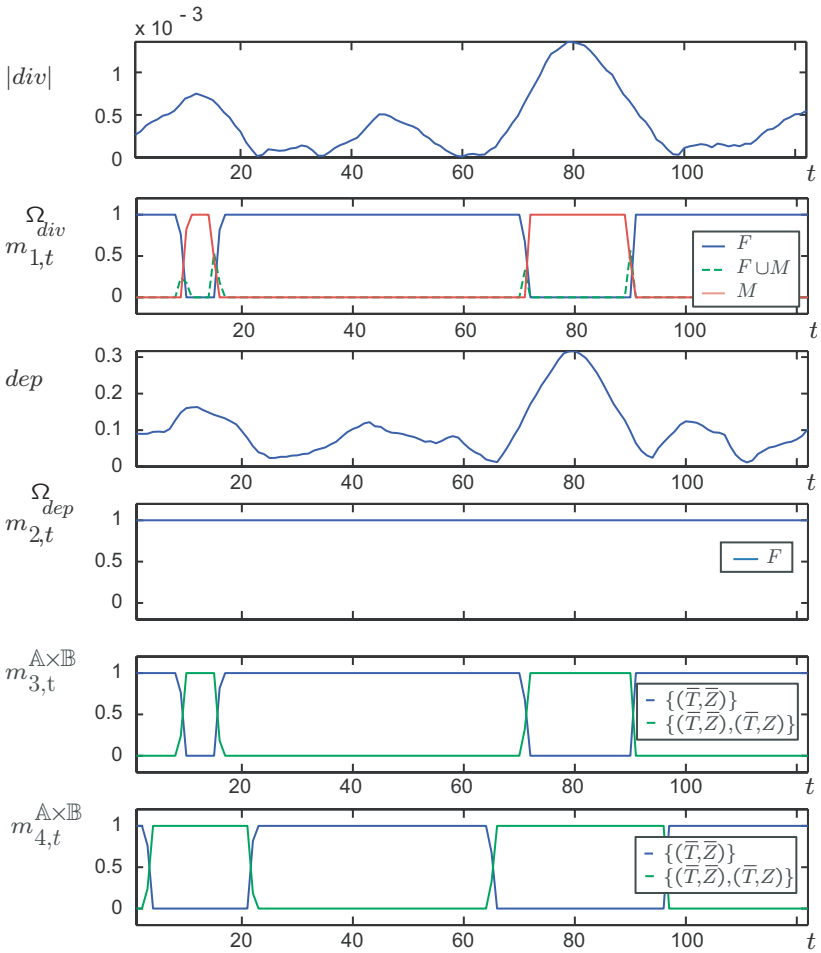


FIG. 4.7 – Etape 1 : Illustration des combinaisons au moyen de règles linguistiques sur une séquence filmée à caméra fixe. Le mouvement attendu est $\{(\overline{T},\overline{Z})\}$. Or parfois le divergent est moyen et le déplacement est grand, le zoom devient alors possible $\{(\overline{T},\overline{Z}),(\overline{T},Z)\}$ sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{3,t}$. Le filtrage (courbe $m^{\mathbb{A}\times\mathbb{B}}_{4,t}$) permet d'amplifier la zone de $\{(\overline{T},\overline{Z}),(\overline{T},Z)\}$ en rajoutant du doute.

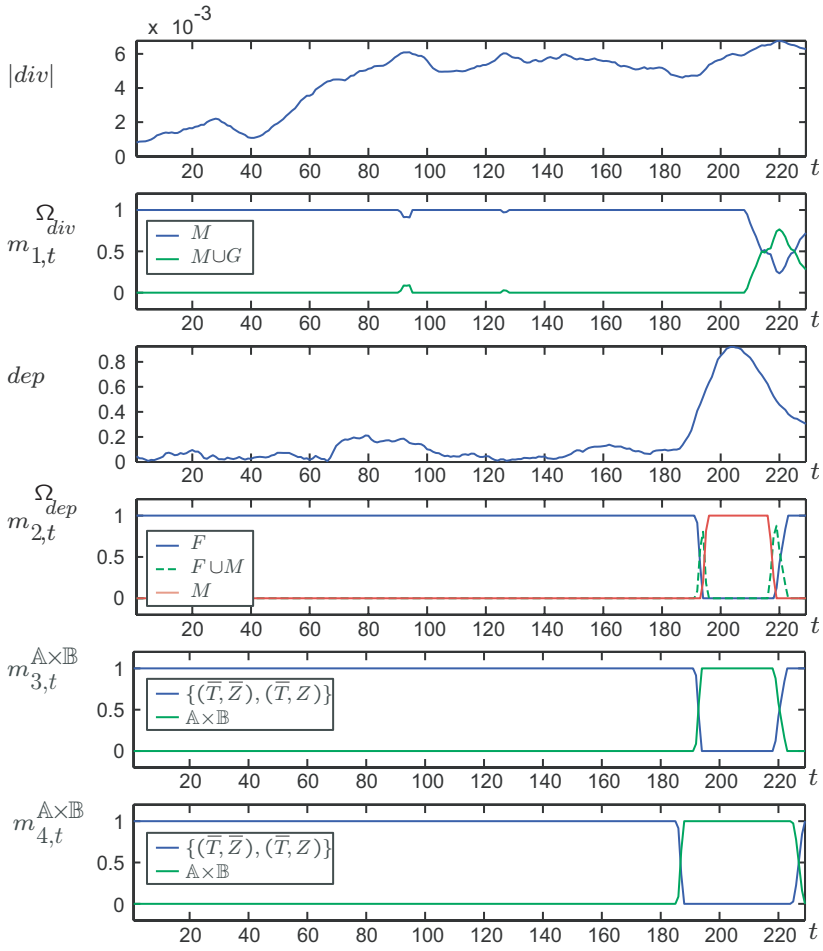


FIG. 4.8 – Etape 1 : Illustration des combinaisons au moyen de règles linguistiques sur une séquence ayant un mouvement de caméra de zoom arrière. Le mouvement attendu est $\{(\overline{T}, Z)\}$, néanmoins cet état n'est jamais obtenu sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{3,t}$. Comme le divergent est parfois moyen et le déplacement est faible, le mouvement est considéré être du statique ou du zoom $\{(\overline{T}, \overline{Z}), (\overline{T}, Z)\}$. Quand le déplacement devient moyen et le divergent est moyen, la masse est alors attribuée au doute total $\mathbb{A} \times \mathbb{B}$. Le filtrage (courbe $m^{\mathbb{A}\times\mathbb{B}}_{4,t}$) permet d'amplifier la zone de $\mathbb{A} \times \mathbb{B}$ en rajoutant du doute.

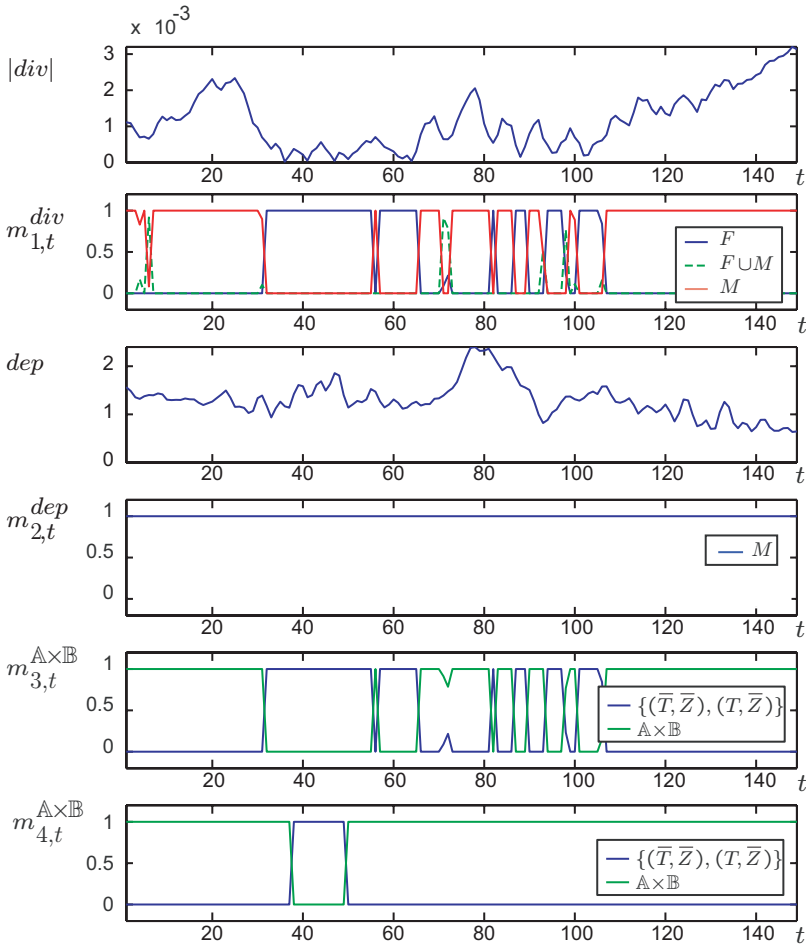


FIG. 4.9 – Etape 1 : Illustration des combinaisons au moyen de règles linguistiques sur une séquence possédant un mouvement de caméra de translation. Le mouvement attendu est $\{(T, \overline{Z})\}$. Comme le déplacement n'est que moyen et le divergent alterne entre moyen et faible, la croyance sur le mouvement (courbe $m^{\mathbb{A}\times\mathbb{B}}_{3,t}$) alterne entre doute total $\mathbb{A} \times \mathbb{B}$ et statique ou translation $\{(\overline{T}, \overline{Z}), (T, \overline{Z})\}$. Devant cette succession de masses sur des propositions différentes, le filtrage (courbe $m^{\mathbb{A}\times\mathbb{B}}_{4,t}$) reporte la croyance sur l'union des propositions, c'est-à-dire le doute total $\mathbb{A} \times \mathbb{B}$.

Si on reprend l'exemple précédent,

$$\begin{array}{lcl} m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(\overline{T},\overline{Z}),\ (T,\overline{Z})\right\}\right) & = & 0.14 \\ m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(T,\overline{Z})\right\}\right) & = & 0.80 \\ m_3^{\mathbb{A}\times\mathbb{B}}\left(\mathbb{A}\times\mathbb{B}\right) & = & 0.06 \end{array}$$

on obtient le résultat suivant :

$$\begin{array}{lcllcl} m_3^{\Omega}(S\cup D) & = & m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(\overline{T},\overline{Z}),\ (T,\overline{Z})\right\}\right) + m_3^{\mathbb{A}\times\mathbb{B}}\left(\mathbb{A}\times\mathbb{B}\right) & = & 0.20 \\ m_3^{\Omega}(D) & = & m_3^{\mathbb{A}\times\mathbb{B}}\left(\left\{(T,\overline{Z})\right\}\right) & = & 0.80 \end{array}$$

Pour chaque image t, une fenêtre temporelle de taille L_3 centrée à t est considérée. La proportion d'images ayant une hypothèse statique sur la fenêtre est déterminée en combinant la BBA $m_{3,t}^{\Omega}$ sur le produit cartésien $\Omega' = \Omega_{t-(L_3-1)/2} \times \ldots \times \Omega_{t+(L_3-1)/2}$ où chaque Ω_i est associé à l'image i de la fenêtre centrée sur l'image t étudiée. Si une masse du produit cartésien Ω' a au moins $\alpha\%$ des images sur l'hypothèse statique alors cette masse est reportée sur l'hypothèse statique S pour l'image t étudiée. De même, si une masse de Ω' a au moins $100-\alpha\%$ des images sur l'hypothèse dynamique alors celle-ci est affectée sur l'hypothèse dynamique D pour l'image t étudiée. Si ce n'est pas le cas, alors la masse est renvoyée sur l'hypothèse $S\cup D$. La proportion des images attribuant une masse sur l'hypothèse statique est déterminée par $n = ceil(L_3.\alpha)$ où $ceil()$ est la fonction arrondie par excès et L_3 est la durée de la fenêtre temporelle autour de l'image étudiée. Le nombre d'images devant avoir une masse sur l'hypothèse dynamique est obtenu par $L_3 - n + 1$.

Prenons l'exemple avec $L_3 = 3$ et $\alpha = 50\%$ avec $\Omega_1 = \{S_1, D_1\}$, $\Omega_2 = \{S_2, D_2\}$ et $\Omega_3 = \{S_3, D_3\}$. Le nombre d'images devant attribué une masse à l'hypothèse statique est de $n = 2$ alors que pour l'hypothèse dynamique le nombre d'images est de $L_3 - n + 1 = 2$. La distribution des masses de croyance de l'image 2 est modifiée suivant les croyances des images voisines (ici image 1 et image 3). On montre dans l'équation 4.4 comment les propositions du produit Cartésien Ω' sont redistribuées sur Ω :

$$\begin{array}{lcl} S_1 \times S_2 \times S_3 & \rightarrow & S \\ S_1 \times S_2 \times D_3 & \rightarrow & S \\ S_1 \times S_2 \times S_3 \cup D_3 & \rightarrow & S \\ S_1 \times D_2 \times S_3 & \rightarrow & S \\ S_1 \times D_2 \times D_3 & \rightarrow & D \\ S_1 \times D_2 \times S_3 \cup D_3 & \rightarrow & S \cup D \\ \ldots & & \\ S_1 \cup D_1 \times S_2 \cup D_2 \times S_3 & \rightarrow & S \cup D \\ S_1 \cup D_1 \times S_2 \cup D_2 \times D_3 & \rightarrow & S \cup D \\ S_1 \cup D_1 \times S_2 \cup D_2 \times S_3 \cup D_3 & \rightarrow & S \cup D \end{array} \tag{4.4}$$

Dans le cas où sur trois images successives, on a une forte masse de croyance portée sur l'hypothèse statique pour les images 1 et 3, et une forte masse de croyance attribuée à un mouvement dynamique pour l'image 2 :

$$\begin{array}{lcllcl} m_3^{\Omega_1}(S_1\cup D_1) & = & 0.3 & m_3^{\Omega_1}(S_1) & = & 0.7 \\ m_3^{\Omega_2}(S_2\cup D_2) & = & 0.4 & m_3^{\Omega_2}(D_2) & = & 0.6 \\ m_3^{\Omega_3}(S_3\cup D_3) & = & 0.1 & m_3^{\Omega_3}(S_3) & = & 0.9 \end{array} \tag{4.5}$$

On obtient alors la BBA m_5^{Ω} :

$$\begin{array}{lcl} m_5^{\Omega}(S) & = & m^{\Omega_1}(S_1) \cdot m^{\Omega_2}(S_2 \cup D_2) \cdot m^{\Omega_3}(S_3) + \\ & & m^{\Omega_1}(S_1) \cdot m^{\Omega_2}(D_2) \cdot m^{\Omega_3}(S_3) \\ & = & 0.63 \\ m_5^{\Omega}(S \cup D) & = & 0.37 \\ m_5^{\Omega}(D) & = & 0 \end{array} \quad (4.6)$$

Cet exemple montre que la distribution des masses dépend de la proportion des croyances sur les hypothèses statique et dynamique. Comme deux images sur trois ont une forte croyance sur l'hypothèse statique, la masse résultante soutient la croyance sur l'hypothèse statique et la masse attribuée à D est nulle.

Basée sur cette règle, une BBA $m_{5,t}^{\Omega}$ sur Ω est définie pour chaque image t. Elle est étendue sur $\mathbb{A} \times \mathbb{B}$ en utilisant les relations $\{(\overline{T}, \overline{Z})\} = S$, $\{(\overline{T}, Z), (T, \overline{Z}), (T, Z)\} = D$ et $\mathbb{A} \times \mathbb{B} = \Omega$, et elle est combinée avec $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$ en utilisant la combinaison conjonctive. La BBA résultante est $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ et si la masse attribuée à l'ensemble vide est non nulle alors celle-ci est transférée sur l'union des hypothèses. Le tableau 4.5.2 récapitule la combinaison des deux BBA.

Tab. 4.2 – Combinaison des BBA $m_{5,t}^{\mathbb{A}\times\mathbb{B}}$ et $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$ en utilisant la règle de combinaison conjonctive et en gérant l'ensemble vide.

		$m_{5,t}^{\mathbb{A}\times\mathbb{B}}$		
		$\{(\overline{T}, \overline{Z})\}$	$\{(T, \overline{Z}), (\overline{T}, Z), (T, Z)\}$	$\{\mathbb{A} \times \mathbb{B}\}$
$m_{4,t}^{\mathbb{A}\times\mathbb{B}}$	$\{(\overline{T}, \overline{Z})\}$	$\{(\overline{T}, \overline{Z})\}$	$\emptyset \rightarrow \mathbb{A} \times \mathbb{B}$	$\{(\overline{T}, \overline{Z})\}$
	$\{(T, \overline{Z})\}$	$\emptyset \rightarrow \{(T, \overline{Z}), (\overline{T}, \overline{Z})\}$	$\{(T, \overline{Z})\}$	$\{(T, \overline{Z})\}$
	$\{(\overline{T}, Z)\}$	$\emptyset \rightarrow \{(\overline{T}, Z), (\overline{T}, \overline{Z})\}$	$\{(\overline{T}, Z)\}$	$\{(\overline{T}, Z)\}$
	$\{(T, Z)\}$	$\emptyset \rightarrow \{(T, Z), (\overline{T}, \overline{Z})\}$	$\{(T, Z)\}$	$\{(T, Z)\}$

La figure 4.10 illustre la deuxième étape de notre approche avec les paramètres suivants $\alpha = 50\%$ et $L_3 = L_2 = L_1 = 13$. La séquence est filmée à caméra fixe. Par exemple, une masse de croyance est affectée à la proposition « zoom ou statique » (entre les images 3 et 21 de $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$) et la séparation permet de redistribuer la masse sur la proposition « statique ». En revanche, la séparation « dynamique/statique » n'est pas suffisante pour allouer la croyance à la proposition « statique » sur tout le segment. En effet, sur une certaine plage de 71 à 90 de $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$, aucune masse est associée à la proposition « statique » d'où la masse résultante (courbe $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$) est assignée à la proposition « statique ou dynamique ». C'est l'intégration de cette plage qui va permettre de retrouver le statique.

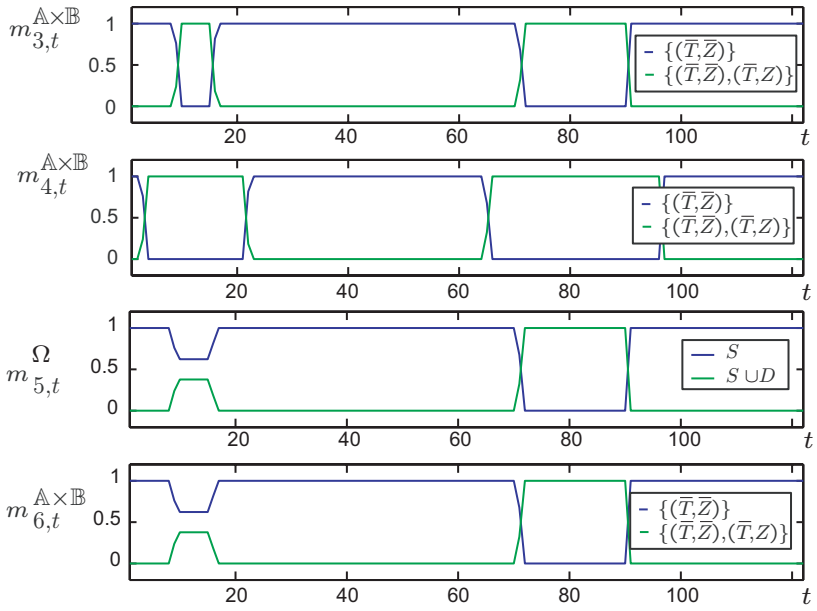


FIG. 4.10 – Etape 2 : Illustration du filtrage et séparation statique/dynamique sur une séquence filmée à caméra fixe. Le mouvement attendu est le statique $\{(\overline{T},\overline{Z})\}$. Or sur deux segments, une masse de croyance est affectée à la proposition « zoom ou statique » (entre les images 3 et 21 et les images 65 et 96 de $m_{4,t}^{\mathbb{A}\times\mathbb{B}}$). Pour le premier segment, la séparation statique/dynamique permet de redistribuer la masse sur la proposition « statique » (courbe $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$). En revanche, pour le second, la séparation ne permet pas de retrouver le statique puisque aucune masse est associée à la proposition « statique » entre les images 71 et 90 de $m_{3,t}^{\mathbb{A}\times\mathbb{B}}$.

4.5.3 Intégration temporelle du zoom et de la translation

Cette dernière étape (Fig. 4.5) réalise une étude plus globale par le passage d'une description du mouvement au niveau de l'image à une description au niveau du segment (ensemble d'images contenant un même type de mouvement). Cela consiste à segmenter la séquence en mouvements cohérents (translation ou zoom) puis à estimer l'amplitude du mouvement sur chaque segment. En décrivant le mouvement sur chaque segment, cette intégration a pour but de conserver uniquement les mouvements d'amplitude et de durée conséquentes.

A l'issue de l'étape 2 (Fig. 4.5), une distribution de masse $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ est disponible pour chaque image t. En transformant cette fonction de masse en loi de probabilité pignistique (Eq. 4.1), le mouvement de caméra potentiel peut être déterminé sur chacune des images. Ainsi, un mouvement potentiel est ensuite déterminé au niveau du segment par regroupement d'images successives ayant le même type de mouvement. Puis, une mesure sur l'amplitude du mouvement est calculée sur chaque segment. Le zoom nécessite le calcul d'un coefficient d'agrandissement alors que la translation demande la détermination du déplacement maximum normalisé. Chacun de ces 2 paramètres décrit aussi bien l'amplitude que la durée du mouvement et permet de représenter le segment. Par exemple, si la caméra est fixe mais soumise à des vibrations alors une translation sera détectée localement (analyse sur un voisinage de quelques images) mais le calcul du déplacement sur le segment permettra de retrouver le statique.

4.5.3.1 Cas du zoom

Dès que la probabilité pignistique $BetP^{\mathbb{A}\times\mathbb{B}}(\{(\overline{T},Z),(T,Z)\})$ sur une image devient supérieure à un seuil δ alors un début de zoom est détecté et cet instant t_0 est conservé. Quand cette grandeur $BetP^{\mathbb{A}\times\mathbb{B}}(\{(\overline{T},Z),(T,Z)\})$ devient inférieure à δ alors le mouvement de zoom s'interrompt et cet instant t_f est mémorisé. Le segment entre les instants t_0 et t_f contient un zoom potentiel qui est analysé pour s'assurer de sa présence. Comme le divergent n'est pas une grandeur très adaptée pour représenter le zoom, le coefficient d'agrandissement est introduit. La figure 4.11 montre un schéma illustrant le zoom avant. Le rectangle gris foncé de l'instant t_0 va s'agrandir jusqu'à atteindre à l'instant t_f la taille de l'image représentée par le rectangle gris clair.

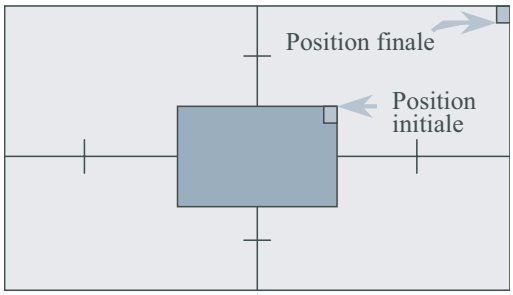


FIG. 4.11 – Schéma illustrant le zoom avant, la région de couleur gris foncé (petit rectangle) va s'agrandir jusqu'à atteindre la taille de l'image.

Nous traitons le cas à une seule dimension. Soit $a_1^{'}(t)$ le paramètre du modèle affine à l'instant t et v_x la vitesse à la position x_i fournie par $v_x = a_1^{'}(t) \cdot x_i$ en supposant les autres coefficients nuls (cas pour un zoom parfait). La position à l'instant $t+1$ est donnée par

$x_i^{'} = x_i + v_x = x_i \cdot (1 + a_1^{'}(t))$. D'où le rapport k_x entre la position à l'instant final t_f et la position à l'instant initial t_0 est donné par :

$$k_x = \frac{\text{Position finale}}{\text{Position initiale}} = \prod_{t=t_0}^{t_f-1} (1 + a_1^{'}(t)).$$

Si le zoom détecté est un zoom avant, le rapport k_x correspond à un coefficient d'agrandissement ($k_x > 1$), noté ag_x. En revanche, si le mouvement est un zoom arrière alors le rapport k_x représente un coefficient de réduction ($k_x < 1$) et par convention l'inverse de ce rapport $ag_x = 1/k_x$ est appelé coefficient d'agrandissement. Ainsi quel que soit le zoom (avant ou arrière), nous définissons uniquement le coefficient d'agrandissement. Dans le cas d'une image (2 dimensions), un coefficient d'agrandissement ag_x est défini suivant l'axe des abscisses et un ag_y suivant l'axe des ordonnées. Pour obtenir un seul coefficient d'agrandissement ag, les deux coefficients ag_x et ag_y sont simplement multipliés. Le coefficient d'agrandissement ag représente le rapport entre la surface de l'image finale et la surface de l'image initiale qui s'est agrandie jusqu'à l'image finale. La figure 4.12 montre un exemple de coefficients d'agrandissement. Pour une plage de zoom donnée, il est possible que le signe du divergent change, ce qui signifie un changement de sens du zoom. Afin d'en tenir compte, nous déterminons temporellement sur la plage de zoom les sous-segments ayant le divergent de même signe et un coefficient d'agrandissement est calculé sur chacun de ces sous-segments.

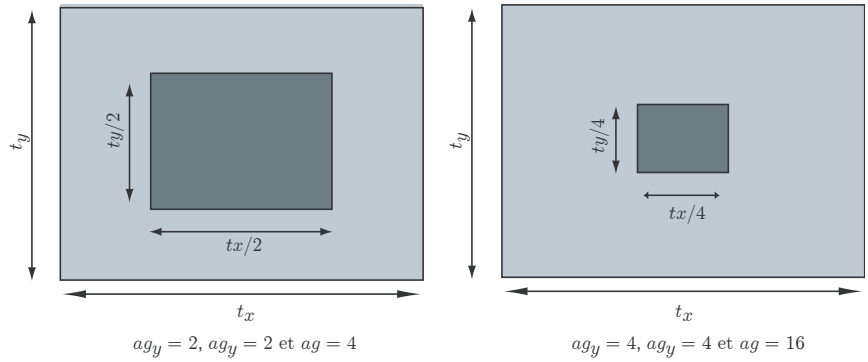


FIG. 4.12 – Exemple de coefficients d'agrandissement. Le rectangle gris foncé correspond à la partie qui va être agrandie jusqu'à atteindre la taille de l'image (rectangle gris clair).

Finalement, le coefficient d'agrandissement caractérise la puissance du zoom sur le segment ou sur le sous-segment. L'amplitude du coefficient d'agrandissement est donc utilisée pour conserver un zoom. Il est transformé en valeurs symboliques suivant les trois ensembles $Zoom$, $\overline{Zoom}$ et $\Omega_Z = Zoom \cup \overline{Zoom}$. La figure 4.13 montre les masses associées suivant le coefficient d'agrandissement.

Ainsi, une fonction de masse $m_7^{\Omega_Z}$ sur le cadre de discernement $\Omega_Z = \{\overline{Zoom}, Zoom\}$ est construite à partir du coefficient d'agrandissement défini sur un segment potentiel de zoom. Pour pouvoir ensuite la combiner, cette BBA $m_7^{\Omega_Z}$ est étendue sur $\mathbb{A} \times \mathbb{B}$ en utilisant les relations $\{(\overline{T}, Z), (T, Z)\} = Zoom$, $\{(\overline{T}, \overline{Z}, (T, \overline{Z})\} = \overline{Zoom}$ et $\mathbb{A} \times \mathbb{B} = Zoom \cup \overline{Zoom}$. La BBA $m_7^{\mathbb{A}\times\mathbb{B}}$ obtenue sur le segment est alors associée à chacune des images de ce segment. Le passage de la description du segment à l'image permet de combiner la fonction de masse $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ précédemment définie avec celle-ci et la fonction de masse résultante devient : $m_{8,t}^{\mathbb{A}\times\mathbb{B}} =$

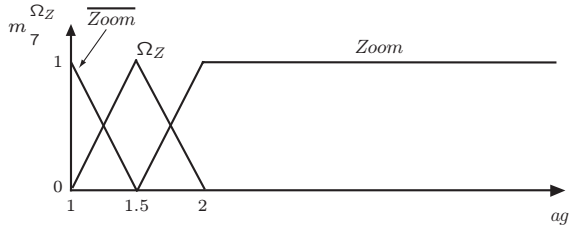


FIG. 4.13 – Définition de la fonction de masse pour le coefficient d'agrandissement.

$m_{6,t}^{\mathbb{A}\times\mathbb{B}} \textcircled{\cap} m_{7,t}^{\mathbb{A}\times\mathbb{B}}$. Le tableau 4.3 récapitule la combinaison des masses et la manière de gérer l'ensemble vide. Il est important de noter que, en cas de conflit, $m_{7,t}^{\mathbb{A}\times\mathbb{B}}$ étant plus fiable que $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ pour le « zoom », la masse associée à l'ensemble vide est reportée sur l'hypothèse du zoom provenant de $m_{7,t}^{\mathbb{A}\times\mathbb{B}}$ et sur l'hypothèse de la translation provenant de $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$.

TAB. 4.3 – Combinaison des fonctions de masse $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ et $m_7^{\mathbb{A}\times\mathbb{B}}$.

		$m_{6,t}^{\mathbb{A}\times\mathbb{B}}$			
		$\{(\overline{T},\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(\overline{T},Z)\}$	$\{(T,Z)\}$
$m_7^{\mathbb{A}\times\mathbb{B}}$	$\{(\overline{T},\overline{Z}),(T,\overline{Z})\}$	$\{(\overline{T},\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\emptyset \rightarrow \{(\overline{T},\overline{Z})\}$	$\emptyset \rightarrow \{(T,\overline{Z})\}$
	$\{(\overline{T},Z),(T,Z)\}$	$\emptyset \rightarrow \{(\overline{T},Z)\}$	$\emptyset \rightarrow \{(T,Z)\}$	$\{(\overline{T},Z)\}$	$\{(T,Z)\}$
	$\mathbb{A}\times\mathbb{B}$	$\{(\overline{T},\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(\overline{T},Z)\}$	$\{(T,Z)\}$

4.5.3.2 Cas de la translation

En procédant comme pour le zoom, dès que $BetP^{\mathbb{A}\times\mathbb{B}}(\{(T,\overline{Z}),(T,Z)\})$ sur une image devient supérieure δ alors un début de translation est détecté et l'indice de début t_0 est conservé. Quand $BetP^{\mathbb{A}\times\mathbb{B}}(\{(T,\overline{Z}),(T,Z)\})$ devient inférieure à δ alors ce mouvement de translation s'achève et cet instant t_f est mémorisé. Le segment potentiel de translation entre les instants t_o à t_f est analysé en calculant le déplacement maximum dep_{max} sur cette fenêtre.

$$t = \underset{t_k \in [t_o, t_f]}{\arg\max} \left(dep(t_o, t_k) \right)$$

$$dep_{max} = dep(t_o, t)$$

Le déplacement maximum dep_{max} est ensuite normalisé par la durée (de l'instant t_o à t) pour avoir une représentation relative sur le déplacement. Ainsi le déplacement maximum normalisé dep_{maxn} caractérise l'amplitude de la translation sur le segment potentiel. Cette grandeur est donc transformée en valeurs symboliques suivant les trois ensembles $Translation$, $\overline{Translation}$ et $\Omega_T = Translation \cup \overline{Translation}$. La figure 4.14 montre la fonction de masse $m_9^{\Omega_T}$ définie sur $\Omega_T = \{Translation, \overline{Translation}\}$ à partir du déplacement maximum normalisé dep_{maxn}.

La BBA $m_9^{\Omega_T}$ est ensuite étendue sur $\mathbb{A}\times\mathbb{B}$ en appliquant les relations $\{(T,\overline{Z}),(T,Z)\} = Translation$, $\{(\overline{T},\overline{Z},(\overline{T},Z)\} = \overline{Translation}$ et $\mathbb{A}\times\mathbb{B} = Translation \cup \overline{Translation}$. L'information obtenue sur le segment est transmise en associant la BBA $m_9^{\mathbb{A}\times\mathbb{B}}$ à chacune des

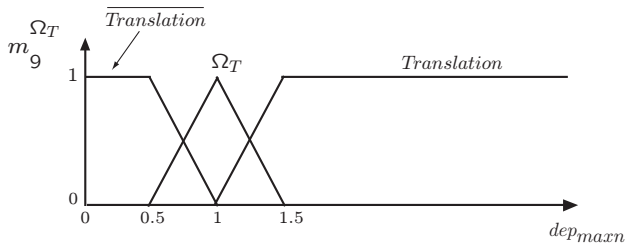


FIG. 4.14 – Définition de la fonction de masse suivant le déplacement maximum normalisé.

images du segment potentiel. Le passage de la description du segment à l'image permet de combiner la fonction de masse $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$ précédemment définie avec celle-ci et la fonction de masse résultante devient : $m_{10,t}^{\mathbb{A}\times\mathbb{B}} = m_{8,t}^{\mathbb{A}\times\mathbb{B}} \textcircled{\cap} m_{9,t}^{\mathbb{A}\times\mathbb{B}}$. Le tableau 4.4 montre la combinaison des BBA et la manière de gérer l'ensemble vide. En cas de conflit lors de la combinaison, $m_9^{\mathbb{A}\times\mathbb{B}}$ étant plus fiable, $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$ ne fait que la conforter au niveau de la translation et réciproquement $m_{9,t}^{\mathbb{A}\times\mathbb{B}}$ n'apportant pas d'information sur le zoom, $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$ ne la prend pas en compte au niveau du zoom.

Finalement la décision sur les mouvements de caméra est réalisée au niveau de chacune des images en choisissant le maximum de la probabilité pignistique calculé à partir de la BBA $m_{10,t}^{\mathbb{A}\times\mathbb{B}}$.

TAB. 4.4 – Combinaison des fonctions de masse $m_{8,t}^{\mathbb{A}\times\mathbb{B}}$ et $m_9^{\mathbb{A}\times\mathbb{B}}$.

		$m_{8,t}^{\mathbb{A}\times\mathbb{B}}$			
		$\{(\overline{T},\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(\overline{T},Z)\}$	$\{(T,Z)\}$
$m_9^{\mathbb{A}\times\mathbb{B}}$	$\{(\overline{T},\overline{Z}),(\overline{T},Z)\}$	$\{(\overline{T},\overline{Z})\}$	$\emptyset \rightarrow \{(\overline{T},\overline{Z})\}$	$\{(\overline{T},Z)\}$	$\emptyset \rightarrow \{(\overline{T},Z)\}$
	$(\{T,\overline{Z}),(T,Z)\}$	$\emptyset \rightarrow \{(T,\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\emptyset \rightarrow \{(T,Z)\}$	$\{(T,Z)\}$
	$\mathbb{A}\times\mathbb{B}$	$\{(\overline{T},\overline{Z})\}$	$\{(T,\overline{Z})\}$	$\{(\overline{T},Z)\}$	$\{(T,Z)\}$

Les figures 4.15, 4.16 et 4.17 illustrent la troisième étape de notre approche avec $\delta = 0.1$. L'exemple sur la séquence filmée à caméra fixe montre que le zoom potentiel sur la courbe $m_{6,t}^{\mathbb{A}\times\mathbb{B}}$ n'est pas conservé et donc la proposition zoom devient statique. De même, la proposition translation lors de l'intégration n'est pas vérifiée et donc seule la proposition zoom est conservée dans la figure 4.16.

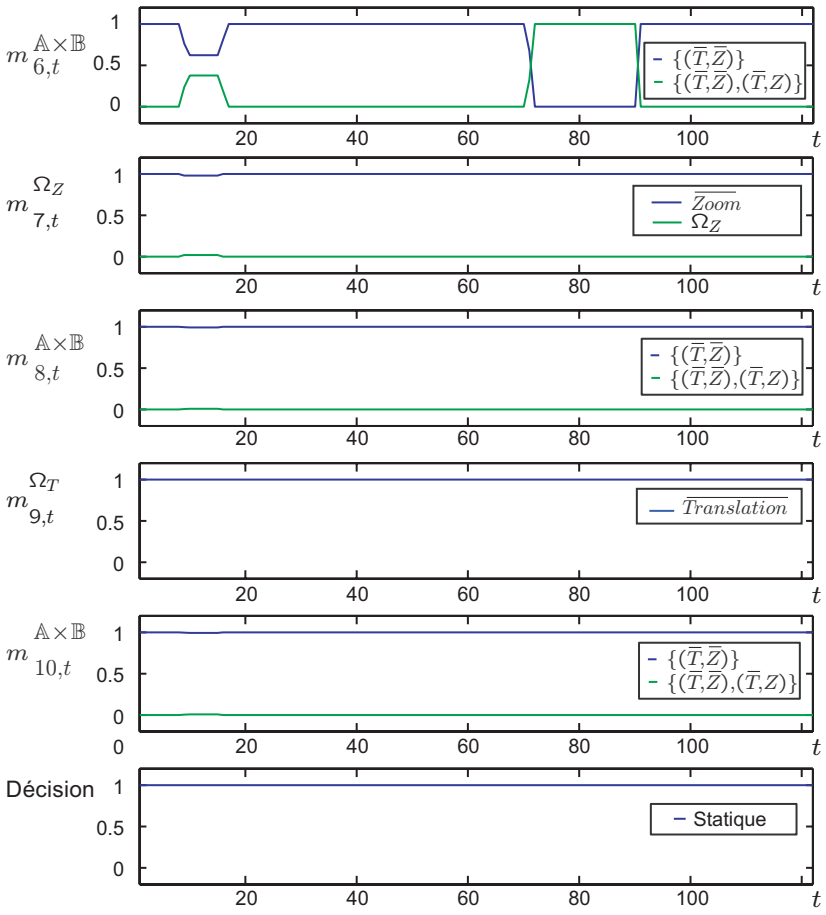


FIG. 4.15 – Etape 3 : Illustration de l'intégration temporelle sur une séquence filmée à caméra fixe. Le mouvement attendu est $\{(\overline{T},\overline{Z})\}$. Or parfois le zoom devient possible $\{(\overline{T},\overline{Z}),(\overline{T},Z)\}$ sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{6,t}$. L'intégration du zoom (courbe $m^{\Omega_Z}_{7,t}$) permet de supprimer ces zones et donc conduit à ne conserver que la proposition $\{(\overline{T},\overline{Z})\}$ sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{8,t}$.

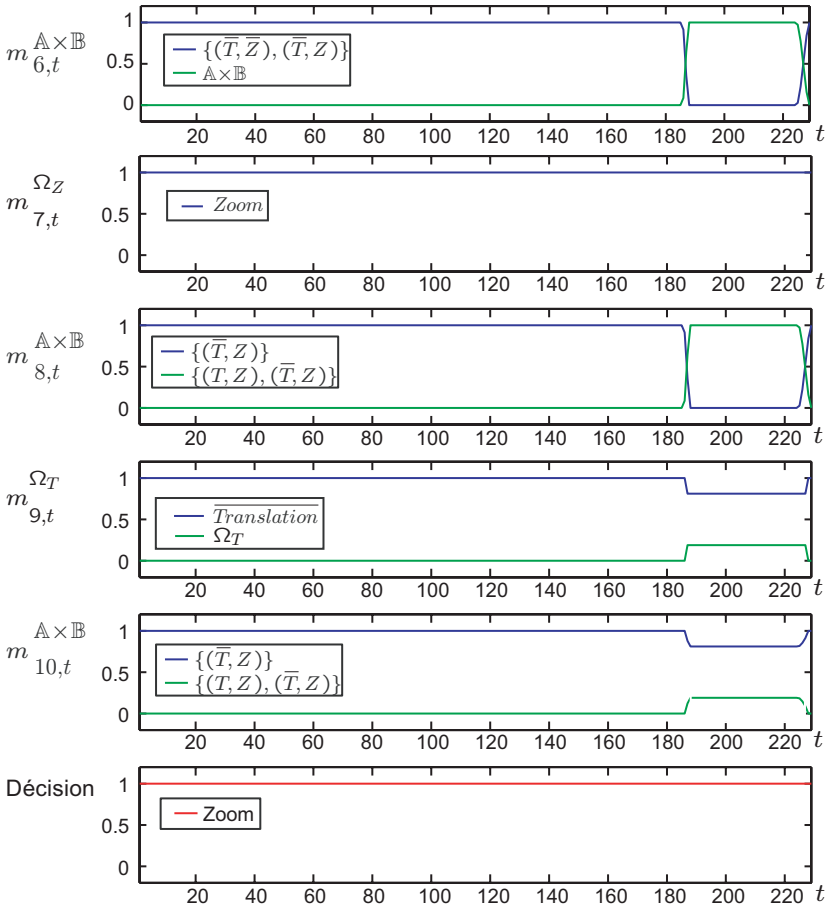


FIG. 4.16 – Etape 3 : Illustration de l'intégration temporelle sur une séquence ayant un mouvement de caméra de zoom arrière. Le mouvement attendu est $\{(\overline{T},Z)\}$. Or parfois la translation et le statique deviennent possibles $\{(\overline{T},\overline{Z}),(\overline{T},Z)\}$ et $\mathbb{A}\times\mathbb{B}$ sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{6,t}$. L'intégration du zoom (courbe $m^{\Omega_Z}_{7,t}$) permet de le conserver et donc entraîne la suppression du statique sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{8,t}$. Puis, l'intégration de la translation (courbe $m^{\Omega_T}_{9,t}$) permet de le supprimer et donc conduit à ne conserver que la proposition $\{(\overline{T},Z)\}$ sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{10,t}$.

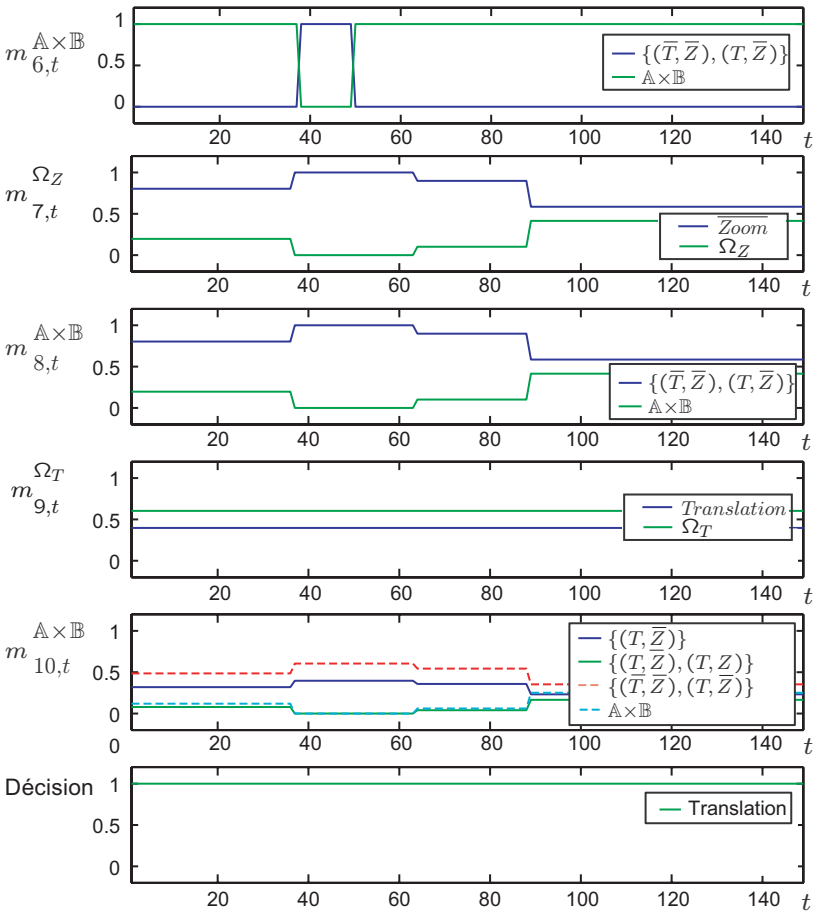


FIG. 4.17 – Etape 3 : Illustration de l'intégration temporelle sur une séquence possédant un mouvement de caméra de translation. Le mouvement attendu est $\{(T,\overline{Z})\}$. Or beaucoup de doute $\mathbb{A}\times\mathbb{B}$ est présent sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{6,t}$. L'intégration du zoom (courbe $m^{\Omega_Z}_{7,t}$) permet de l'annuler et donc conduit à ne conserver que la translation ou le statique sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{8,t}$. Puis, l'intégration de la translation (courbe $m^{\Omega_T}_{9,t}$) aboutit à une croyance de 0.4 pour la translation. Plusieurs propositions sont donc conservées sur la courbe $m^{\mathbb{A}\times\mathbb{B}}_{10,t}$. Les deux plus fortes croyances sont sur statique ou translation $\{(\overline{T},\overline{Z}),(T,\overline{Z})\}$ et translation $\{(T,\overline{Z})\}$, et le calcul de la probabilité pignistique permet de retrouver le mouvement de translation.

4.6 Quantification des mouvements de caméra

Cette phase (Fig. 4.1) consiste à décrire de manière quantitative chaque segment où un mouvement de caméra a été déterminé. En s'appuyant sur les résultats des paragraphes précédents, les choix ci-dessous ont été effectués. Un mouvement de zoom est représenté par le coefficient d'agrandissement ainsi que le sens du zoom. En revanche, le mouvement de translation détecté à un instant t_0 jusqu'à un instant t_f est caractérisé par le déplacement total et le chemin parcouru.

Pour le mouvement de translation, une autre caractéristique importante est sa direction. Cependant, comme ce mouvement ne se limite pas aux quatre orientations principales (du bas vers le haut, du haut vers le bas, de la gauche vers la droite et de la droite vers le gauche), une quantification floue est utilisée pour permettre des transitions progressives entre ces quatre orientations.

Ainsi, pour chaque image contenue dans un segment de translation, un poids est attribué suivant la direction et le sens du mouvement en fonction de la phase du vecteur $\overrightarrow{dep}(t)$. La figure 4.18 montre les fonctions de pondération des quatre orientations. La phase est segmentée en 4 zones correspondant aux quatre orientations. Chaque zone possède une largeur de $2\pi/3$, centrée sur l'axe horizontal ou vertical. Deux zones voisines se chevauchent sur un intervalle de largeur $\pi/6$. Par exemple, un mouvement diagonal du bas-gauche vers le haut-droit est caractérisé par les quatre indicateurs suivants (Zone 1, Zone 2, Zone 3, Zone 4) = $(0.5, 0.5, 0, 0)$.

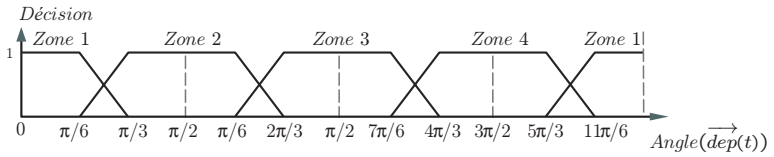


FIG. 4.18 – Fonctions d'appartenance suivant les 4 orientations. Par exemple, la zone 1 correspond à une translation horizontale de la droite vers la gauche.

En résumé, le classifieur de mouvement fournit les informations suivantes. Chaque mouvement (statique, translation et zoom) est décrit par une décision binaire attribuée au niveau de chaque image. Le mouvement de translation possède également quatre indicateurs sur l'orientation (gauche-droite, droite-gauche, bas-haut et haut-bas) définis sur chaque image où un mouvement de translation est détecté. Enfin, le mouvement de zoom a un indicateur sur le sens du zoom (zoom avant +1 et zoom arrière -1).

De plus, chaque segment où un mouvement de translation est identifié est décrit par le chemin parcouru et le déplacement total alors qu'un segment ayant un zoom est représenté par le coefficient d'agrandissement.

4.7 Evaluation de la classification des mouvements

La classification des mouvements de caméra doit être évaluée pour s'assurer de la performance de la méthode. Nous effectuons d'abord une étude sur des extraits de vidéos contenant un unique mouvement de caméra, puis nous procédons à une étude sur des vidéos contenant plusieurs types de mouvements. Par la suite, nous appliquerons la méthode avec les seuils suivants : $L_1 = L_2 = L_3 = 13$ taille de la fenêtre d'analyse qui correspond à environ une demi

seconde, $\alpha = 50\%$ proportion des images statiques et dynamiques (étape 2) et $\delta = 0.1$ le seuil de détection d'un mouvement de zoom ou de translation (étape 3).

4.7.1 Analyse de mouvements uniques

Il s'agit de vérifier l'efficacité de la méthode sur une base de vidéos. L'étude est réalisée sur des extraits ayant un mouvement de caméra unique. Nous présentons successivement le corpus, les mesures d'évaluation puis les résultats obtenus.

4.7.1.1 Corpus

Pour évaluer la performance de la classification des mouvements de caméra, des extraits de vidéos avec des contenus riches et variés (séquences de sport, série « The Avengers »,...) ont été annotés suivant les trois types de mouvement de caméra considérés (Fig. 4.19). Lors de la lecture des vidéos, des extraits contenant un unique mouvement de caméra ont été sélectionnés. Ces mouvements correspondent à des mouvements souhaités par le réalisateur et sont appelés mouvements principaux. Néanmoins, ces extraits peuvent aussi contenir des petits mouvements non souhaités (par exemple, légère translation dans un segment statique ou légère vibration haut bas dans un extrait de translation horizontale). Ces mouvements secondaires sont visibles lorsque les vidéos sont visionnées à une cadence faible. Notre approche doit être insensible aux mouvements secondaires et ne devra considérer que les mouvements principaux.

Le corpus se compose de 42 extraits de vidéos (4605 images) de durée variable de l'ordre de quelques secondes chacune :

- 8 extraits (899 images au total) n'ayant pas de mouvement (caméra fixe)
- 21 extraits (2053 images au total) contenant un mouvement de translation (6 mouvements de droite à gauche, 7 de haut en bas, 6 de gauche à droite et 2 bas en haut)
- 13 extraits (1663 images au total) contenant des zooms (7 zooms avant et 6 zooms arrière)

4.7.1.2 Mesures d'évaluation

Afin d'évaluer la méthode, nous avons utilisé les mesures classiques que sont le rappel et la précision ainsi que la retombée. Le rappel R évalue la capacité du classifieur à classer les vidéos dans la base pour un mouvement recherché et il se définit par la relation suivante : $R = N_{CT}/N_C$ où N_C désigne le nombre de vidéos dans la base contenant le mouvement souhaité (nombre de vidéos correctes) et N_{CT} le nombre de vidéos trouvées par le classifieur et ayant le mouvement recherché (nombre de vidéos correctes et trouvées). La précision P évalue la capacité du classifieur à retrouver uniquement des vidéos ayant le mouvement souhaité. Il est obtenu par $P = N_{CT}/N_T$ où N_T désigne le nombre de vidéos retrouvées par le classifieur (nombre de vidéos trouvées) et N_{CT} le nombre de vidéos trouvées par le classifieur et ayant le mouvement recherché (nombre de vidéos correctes et trouvées). La retombée D (ou fallout) évalue la capacité du classifieur à retourner des vidéos ne contenant pas le mouvement souhaité : $D = N_{IT}/N_I$ où N_I désigne le nombre total de vidéos dans la base ne contenant pas le mouvement souhaité (nombre de vidéos incorrectes) et N_{IT} le nombre de vidéos trouvées par le classifieur ne contenant pas le mouvement souhaité (nombre de vidéos incorrectes et trouvées).

En outre, la classification d'un extrait de vidéos dépend du mouvement de caméra attribué à chacune de ces images. Comme des mouvements non apparents peuvent être présents dans les

FIG. 4.19 – Exemple d'extraits de vidéos contenus dans la base. Deux extraits sont filmés à caméra fixe (en haut), deux comportent un mouvement de translation (au milieu) et enfin deux contiennent un zoom (en bas). Pour chaque extrait, l'image de gauche (resp de droite) correspond à la première (resp dernière) image de l'extrait.

vidéos et perturber localement la classification, nous considérons qu'une vidéo est correctement identifiée si elle possède au moins X% des images correctement classées. Pour avoir une bonne représentation de la performance de la classification, le rappel, la précision et la retombée sont calculés à un niveau complet (100% des images de la vidéo sont correctement classées) et à un niveau partiel (au moins 80% des images de la vidéo sont correctement classées). Pour la retombée, le niveau partiel est plus bas avec au moins 20% des images détectées mais ne contenant le mouvement souhaité.

4.7.1.3 Résultats

Les résultats de la classification des mouvements sont interprétés à deux niveaux. Tout d'abord, il s'agit de s'assurer que la classification binaire est exacte (présence de statique, translation ou zoom) puis il faut vérifier que la description du mouvement de la translation et du zoom est également correcte.

Le tableau 4.5 montre les résultats de la classification des mouvements avec le rappel et la précision calculés à 100% (niveau complet) et 80% (niveau partiel), et avec la retombée à 100% et 20%. Nous pouvons constater de bons résultats au niveau complet. Ceci témoigne de la robustesse de la méthode. Si on s'intéresse au mouvement de zoom, le rappel indique qu'une vidéo n'est pas retrouvée. Il s'agit d'un zoom qui est en fait détecté à 73%. Le début de cette vidéo a un zoom léger et il est apparenté à un statique. En ce qui concerne le mouvement de translation, il manque une vidéo au niveau du rappel complet, celle-ci est détectée à 95% et possède une plage statique au début de la vidéo. Pour ce qui est du statique, une vidéo a été mal classée (retombée à 20%). Il s'agit de la vidéo contenant un zoom détecté à 73%. Par conséquent, la classification des mouvements présente de bonnes performances avec une précision de 100%, un rappel $> 92\%$ et une retombée $< 3\%$ pour les trois mouvements de

caméra.

Le tableau 4.6 illustre la classification suivant l'orientation du mouvement de caméra. Nous constatons encore la performance de la méthode. Effectivement, si le mouvement est correctement classé (Tab. 4.5), la description apportée par l'orientation du mouvement est également correcte (Tab. 4.6). Nous pouvons conclure que le classifieur est fiable.

Tab. 4.5 – Résultats de la classification avec rappel et précision (100% et 80%) et retombée (100% et 20%).

		Translation	**Zoom**	**Statique**
Rappel	**Complet**	*95.24 (20/21)*	*92.31 (12/13)*	*100 (8/8)*
	Partiel à 80%	*100 (21/21)*	*92.31 (12/13)*	*100 (8/8)*
Précision	**Complet**	*100 (20/20)*	*100 (12/12)*	*100 (8/8)*
	Partiel à 80%	*100 (21/21)*	*100 (12/12)*	*100 (8/8)*
Retombée	**Complet**	*0 (0/21)*	*0 (0/29)*	*0 (0/34)*
	Partiel à 20 %	*0 (0/21)*	*0 (0/29)*	*2.94 (1/34)*

Tab. 4.6 – Résultats de la classification locale avec rappel et précision à 80%.

	Translation (droite à gauche)	**Translation (haut vers bas)**	**Translation (gauche à droite)**	**Translation (bas vers haut)**	**Zoom avant**	**Zoom arrière**
Rappel à 80%	*100 (6/6)*	*100 (7/7)*	*100 (6/6)*	*100 (2/2)*	*100 (7/7)*	*83.33 (5/6)*
Précision à 80%	*100 (6/6)*	*100 (7/7)*	*100 (6/6)*	*100 (2/2)*	*100 (7/7)*	*100 (5/5)*

4.7.2 Analyse de mouvements composés

La classification des mouvements est illustrée ici en considérant des vidéos contenant plusieurs mouvements de caméra. Il s'agit de vérifier le bon fonctionnement de la méthode lorsque les mouvements recherchés sont superposés (zoom et translation) ou successifs à l'intérieur d'un même extrait.

La figure 4.20 montre un exemple d'extrait de vidéos comprenant plusieurs mouvements de caméra, d'abord une plage de mouvement de zoom arrière et de translation de gauche à droite, puis une plage de zoom arrière et enfin une plage statique. Cet exemple permet aussi de mettre en évidence la description des mouvements de caméra. On retrouve les indicateurs des différents mouvements de caméra (les 4 orientations de la translation, le chemin parcouru, le déplacement total normalisé ainsi que le sens du zoom et le coefficient d'agrandissement).

Afin d'étudier la classification des mouvements de caméra, nous avons annoté trois vidéos suivant les trois types de mouvement de caméra. Les vidéos sont celles qui ont été employées dans le chapitre 3 pour le résumé de vidéo : un documentaire sportif sur le saut à ski (appelé « Documentaire ») avec 20 plans et 3271 images, une série « The Avengers » (appelée « Série ») avec 27 plans et 2412 images et un journal télévisé (appelé « Journal ») avec 42 plans et 6870 images. En supposant les plans connus, les différents mouvements sont estimés par notre méthode et comparés à la vérité terrain. L'évaluation est effectuée par le rappel et la précision au niveau de l'image (calcul du nombre d'images correctement identifiées pour chaque mouvement).

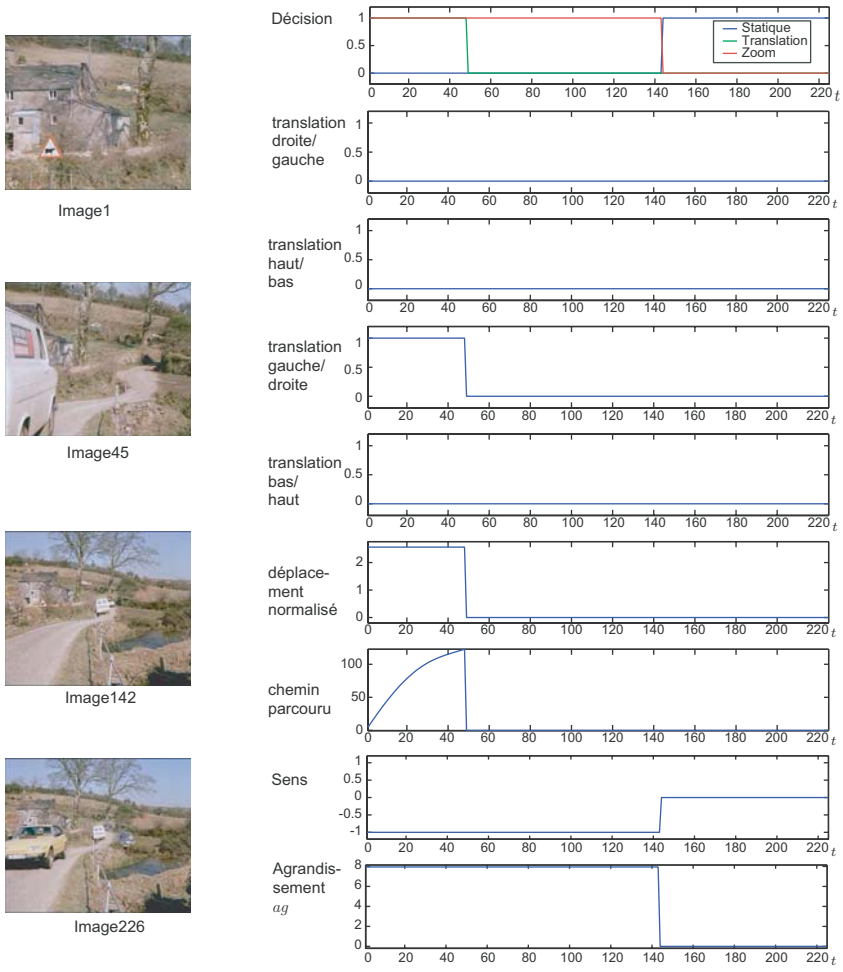


FIG. 4.20 – Résultats de la classification sur un extrait de vidéo comprenant plusieurs mouvements de caméra.

Néanmoins, le mouvement de caméra est parfois difficile à déterminer à certain endroit de la vidéo (ambiguïté entre mouvements) et la frontière entre deux mouvements successifs n'est pas forcément facile à trouver. Ainsi la vérité terrain n'étant pas parfaite, à ces types d'erreurs se rajoutent les erreurs de classification provenant de notre méthode. Ceci permet de nuancer les chiffres présentés dans le tableau 4.7. Nous pouvons remarquer que les résultats sont bons avec un rappel et une précision supérieurs à 78% pour les trois vidéos. Un exemple de la classification des mouvements de caméra est présenté dans la figure 4.21 et correspond au premier plan de la vidéo « Série » où un mouvement de translation est suivi d'une plage statique. Nous pouvons remarquer que les mouvements identifiés par la méthode sont similaires à ceux de la vérité terrain. Comme la frontière entre les mouvements n'est pas exactement au même endroit, le rappel et la précision ne sont respectivement que de 81% et 100% pour le statique et de 100% et 95% pour la translation alors que l'estimation des mouvements semble être correcte.

TAB. 4.7 – Résultats de la classification des mouvements sur trois vidéos.

	Documentaire	**Journal**	**Série**
Rappel	*85 (2884/3401)*	*86 (5990/6967)*	*90 (2409/2663)*
Précision	*79 (2884/3632)*	*85 (5990/7045)*	*89 (2409/2718)*

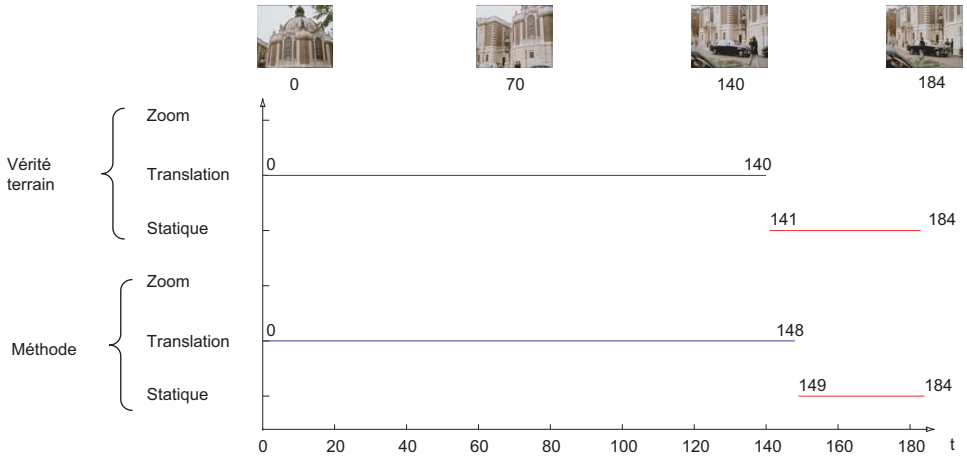


FIG. 4.21 – Exemple de classification des mouvements de caméra sur le premier plan de la vidéo « Série ».

4.8 Conclusion

Nous avons présenté une méthode de classification des mouvements de caméra basée sur le Modèle des Croyances Transférables. Elle consiste à localiser dans une vidéo les mouvements de translation, de zoom et les plages statiques. La classification des mouvements se distingue par son architecture de fusion qui se divise en trois étapes (Fig. 4.5) et qui repose sur un certain nombre d'hypothèses. La première étape consiste à convertir les paramètres d'un modèle

affine décrivant le mouvement en valeurs symboliques pour ensuite les combiner suivant des règles et obtenir des fonctions de masse sur les mouvements de caméra au niveau de chaque image. Il est intéressant de remarquer que les règles de combinaison sont conçues pour éviter dans la mesure du possible les mouvements secondaires (mouvements de faible amplitude). Puis, un filtrage selon le MCT est réalisé et a pour effet de remettre en cause la croyance en un mouvement suivant le voisinage de l'image étudiée. La deuxième étape concerne la séparation entre les images statiques et dynamiques. Elle repose sur l'hypothèse que si un certain nombre d'images autour de l'image étudiée ont une croyance sur l'hypothèse statique alors l'image étudiée est considérée comme statique. Enfin la dernière étape consiste à regrouper les images successives suivant la croyance portée sur un même type de mouvement puis à estimer l'amplitude du mouvement sur le segment. L'avantage de l'analyse au niveau du segment est de conserver seulement les mouvements d'amplitude et de durée conséquentes. La quantification des mouvements identifiés est ensuite effectuée (par exemple, coefficient d'agrandissement pour un zoom) pour les décrire plus facilement et les utiliser dans d'autres applications. Les mouvements de translation et de zoom sont également caractérisés de manière plus locale avec la direction (zoom avant, zoom arrière, translation de gauche à droite...).

Nous avons mené une étude afin de s'assurer de la performance de la méthode. Tout d'abord, nous avons présenté des résultats sur des vidéos ne contenant qu'un seul type de mouvement. Les résultats obtenus en terme de rappel, précision et retombée nous permettent de conclure que la méthode de classification est efficace pour déterminer les mouvements de caméra. Puis, nous avons évalué la méthode sur des vidéos contenant des mouvements de caméra pouvant se superposer ou se succéder. Bien que la vérité terrain sur des vidéos réelles soit parfois difficile à obtenir à quelques endroits des vidéos (ambiguïté entre mouvements ou placement des frontières entre mouvements), les résultats sont bons avec un rappel et une précision supérieurs à 78%.

Notre approche pourrait également considérer d'autres types de mouvement comme la rotation autour de l'axe optique. Le MCT ne posera pas de problème théorique pour les intégrer. Néanmoins, le fait de ne pas prendre en compte certains mouvements de caméra peut aboutir à des classifications erronées par notre méthode. Il est possible de travailler dans un cadre plus général car le MCT peut aussi définir une classe de rejet des mouvements non considérés pour ne pas perturber la classification. Cependant, les trois types de mouvement de caméra que nous cherchons à identifier sont les plus répandus dans les vidéos. C'est pourquoi notre méthode reste adéquate pour classer les mouvements de caméra.

Nous allons employer dans le chapitre 5 la classification des mouvements de caméra pour construire une nouvelle méthode de résumé de vidéo. Suivant l'enchaînement et l'amplitude des mouvements de caméra, des règles peuvent être établies pour sélectionner judicieusement les images clés. Comme l'identification des mouvements de caméra nécessite la connaissance des changements de plan, nous allons développer dans le chapitre 6 une méthode de détection de changements de plan basée sur le Modèle des Croyances Transférables.

Chapitre 5

Création de résumé de vidéo à partir du mouvement de caméra

Nous allons exploiter ici l'information apportée par le mouvement de caméra, caractéristique de plus haut niveau, pour construire le résumé de vidéo. Le changement temporel de contenu dans les vidéos peut être révélé par l'amplitude et l'enchaînement des mouvements de caméra et leur détection peut éviter la redondance temporelle des images clés pour le résumé. Nous proposons ainsi une méthode originale de résumé de vidéo à partir du mouvement caméra.

Sommaire

5.1 Introduction

Dans ce chapitre, nous nous intéressons à la création de résumé de vidéo en utilisant l'information fournie par le mouvement de caméra. Comme décrit dans le chapitre 4, l'identification et la quantification des mouvements de caméra apportent des renseignements susceptibles d'être pertinents pour la création de résumé.

Plusieurs méthodes de résumé de vidéo ont été proposées à partir du mouvement de caméra. Parmi ces méthodes, nous pouvons distinguer celles qui n'exploitent pas directement le mouvement de caméra pour construire le résumé. Dans [Kop04], le mouvement de caméra est employé pour détecter les objets en mouvement, caractéristique qui est ensuite utilisée pour le résumé. Dans d'autres travaux [Shi03b, Ma05b], le mouvement de caméra sert uniquement à partitionner les plans en autant de segments que de mouvements détectés. De la même manière, Zhu et al. [Zhu05] segmentent les plans suivant les mouvements de caméra détectés. Puis, les vecteurs de mouvement du flux MPEG sont utilisés pour définir sur chaque segment, une courbe d'intensité de mouvement. Si le segment est considéré comme statique, le minimum global sur cette courbe est choisi comme image clé. Sinon le segment est divisé en deux à l'endroit du maximum global et les images clés correspondant au minimum sont sélectionnées. Par conséquent, les plans sont segmentés suivant les mouvements de caméra détectés et la sélection des images clés n'est pas effectuée suivant le mouvement de caméra mais en fonction d'un descripteur sur l'intensité du mouvement (flux MPEG).

Les méthodes de résumé qui ont directement été conçues à partir du mouvement de caméra reposent essentiellement sur la présence ou l'absence de mouvement. Par exemple, Cherfaoui et al. [Che94] détectent les plans puis déterminent la présence ou l'absence de mouvement de caméra. Les plans avec un mouvement de caméra sont représentés par trois images clés alors que les plans tournés à caméra fixe n'en possèdent qu'une. Taniguchi et al. [Tan97] fabriquent une image composite pour chaque plan contenant un mouvement de caméra. Afin de capturer les aspects dynamiques de la vidéo, Peker et al. [Pek03] élaborent une méthode de résumé en sélectionnant les segments avec des mouvements importants (de caméra ou des objets). Dans [Kau02], les segments ayant un mouvement de caméra ont des images clés qui sont ajoutées afin de montrer l'arrière plan. Néanmoins, ces approches reposent sur des considérations simples qui exploitent peu l'information apportée par le mouvement de caméra.

Afin de réduire la redondance temporelle entre les images, des méthodes ont été développées pour mesurer, à partir du mouvement de caméra, la ressemblance entre les images. Dans [Por03], le mouvement dominant (modèle affine) est estimé pour décrire le mouvement de la caméra. Cette estimation est ensuite utilisée pour identifier les plans contenant un mouvement de caméra significatif et nécessitant plus d'une image clé. Une mesure de similarité entre deux images est alors définie en calculant le recouvrement entre elles. Plus le recouvrement est faible plus le contenu est différent. Finalement, à partir des paramètres du mouvement affine, le recouvrement est déterminé et permet de représenter chaque plan par un graphe orienté. Le chemin le plus court est calculé afin d'obtenir le résumé. De même, Fauvet et al. [Fau04] déterminent à partir de l'estimation du mouvement dominant (modèle affine), les régions entre deux images successives qui ont été perdues ou sont apparues. La surface apparue sur tout le plan, obtenue en additionnant la surface des régions apparues entre les paires d'images, est utilisée pour déterminer le nombre d'images clés à sélectionner. Puis la fonction cumulative des surfaces apparues entre la première image du plan et l'image courante est employée pour déterminer les images clés. Néanmoins, ces approches s'appuient sur une description de bas niveau qui consiste à mesurer le recouvrement entre les images. Elles reposent sur des pro-

priétés géométriques et locales (nombre de pixels conservés ou perdus entre deux images) et n'orientent pas la sélection des images en fonction du type de mouvement détecté.

Enfin, des méthodes plus élaborées consistent à étudier l'attention visuelle (i.e les endroits où le regard se porte lors du visionnage de la vidéo). Dans [Ma02], une courbe d'attention est obtenue en combinant plusieurs valeurs d'attention associées à chaque image de la séquence vidéo. Ces valeurs sont obtenues à partir de différents index et symbolisent le degré d'importance de l'image. Parmi les différentes valeurs d'attention, l'une d'elles provient du mouvement de caméra. La méthode de résumé consiste à sélectionner les images où l'attention est maximale. Néanmoins, la méthode combine un grand nombre de mesures d'attention et ne met pas en évidence l'apport du mouvement de caméra pour construire le résumé.

Nous proposons une nouvelle méthode de résumé de vidéo basée sur les mouvements de caméra. Elle consiste à étudier l'amplitude et l'enchaînement des mouvements. A partir de ces deux critères, différentes règles sont élaborées pour construire le résumé. Par exemple, la sélection d'images clés sera différente si l'amplitude du mouvement de translation est forte ou faible. De même, si un mouvement de translation est suivi d'un segment statique, alors cet enchaînement de mouvements pourra être décrit par l'image au début de la translation et l'image au milieu du segment statique. Ainsi la sélection des images va dépendre à la fois de l'amplitude et de l'enchaînement des mouvements de caméra. L'avantage de la méthode est qu'elle n'a pas besoin de comparer des images entre elles (mesure de similarité ou de recouvrement entre images au niveau des pixels) et repose uniquement sur la classification des mouvements de caméra.

La méthode de résumé de vidéo, présentée dans la partie 5.2, se décline en trois variantes. La première considère uniquement l'amplitude des mouvements de caméra. La deuxième étudie l'enchaînement des évènements ou plus précisément l'enchaînement des mouvements de caméra. Enfin la troisième combine l'amplitude et l'enchaînement des mouvements de caméra pour fournir un résumé de vidéo. Cette dernière, qui reprend les deux premières variantes, constitue la méthode de résumé à partir du mouvement de caméra. Afin de juger de la performance de la méthode, une méthode d'évaluation est décrite dans la partie 5.3 et permet la comparaison de différentes méthodes.

5.2 Méthode de construction de résumé de vidéo à partir du mouvement de caméra

Nous allons présenter les trois variantes de la méthode : résumé fonction de l'amplitude des mouvements, résumé fonction de l'enchaînement des mouvements et enfin résumé fonction de l'amplitude et de l'enchaînement des mouvements. Ces différentes variantes requièrent les paramètres extraits lors de la méthode de classification des mouvements de caméra et décrits dans le chapitre 4.

Nous effectuons un bref rappel des caractéristiques fournies par cette méthode. Elle consiste à localiser dans une vidéo les mouvements de translation (verticale ou/et horizontale), de zoom et l'absence de mouvement de caméra. La méthode fournit également une quantification des mouvements identifiés. Ainsi un segment où un zoom a été détecté est représenté par le coefficient d'agrandissement ag et la direction du zoom (arrière ou avant). Un segment de translation est quant à lui décrit par le chemin parcouru noté ch et le déplacement total noté dpt. Dans la suite de ce chapitre, les plans des vidéos sont supposés connus et l'identification et la quantification des mouvements de caméra sont fournies par la méthode de classification.

A partir de ces caractéristiques, différentes stratégies sont adoptées pour construire le résumé de vidéo.

5.2.1 Résumé fonction de l'amplitude des mouvements de caméra

Cette approche a pour objectif de sélectionner des images clés en fonction de l'amplitude du mouvement de caméra. La sélection des images repose sur des observations élémentaires pour représenter l'intégralité de la vidéo avec un minimum de redondance. Par exemple, un mouvement de translation avec une forte amplitude nécessitera plus d'images clés qu'un segment statique puisque le contenu visuel sera d'autant plus dissimilaire d'une image à l'autre que le mouvement sera important. Un segment de zoom aura aussi une sélection différente suivant l'amplitude de son coefficient d'agrandissement. Compte tenu du type de mouvement détecté et de l'amplitude du mouvement, des règles sont établies pour déterminer les images clés.

Pour un segment de translation, le coefficient c_r est calculé afin de déterminer si la trajectoire est rectiligne.

$$c_r = \frac{|ch - dpt|}{ch}$$

Le coefficient c_r est compris entre 0 et 1 et informe sur la trajectoire du mouvement. Plus c_r est petit, plus le mouvement est rectiligne. Au contraire, plus c_r est grand, plus le mouvement de translation dans le segment possède des changements de direction. Par conséquent, si le coefficient c_r est inférieur à un seuil δ_r, le mouvement est considéré rectiligne. Dans ce cas, si le déplacement total dpt est important c'est-à-dire supérieur à un seuil δ_{dt}, la première image et la dernière image du segment sont sélectionnées. Seule la dernière image est sélectionnée si le déplacement total dpt est faible, inférieur au seuil δ_{dt}. En revanche, si le coefficient c_r est supérieur à δ_r, le mouvement possède des changements de direction. Les images du début, du milieu et de fin du segment sont sélectionnées si le déplacement total dpt est supérieur à un seuil δ_{dt} sinon seule la dernière image est choisie. En ce qui concerne un segment de zoom, les images clés sont sélectionnées en fonction du coefficient d'agrandissement ag. Si l'agrandissement est important c'est-à-dire supérieur à un seuil δ_{ag}, la première image et la dernière image du segment sont sélectionnées. Dans le cas contraire, seule la dernière est choisie. Enfin, pour un segment statique, l'image au milieu du segment est choisie. Après une étude expérimentale, nous avons choisi les seuils suivants : $\delta_r = 0.5$, $\delta_{dt} = 300$ et $\delta_{ag} = 5$. La sélection des images clés en fonction de l'amplitude des mouvements de caméra est résumée dans l'algorithme 5.1.

Algorithme 5.1 Sélection des images clés en fonction du type et de l'amplitude du mouvement de caméra

INITIALISATION : $\delta_r = 0.5$, $\delta_{dt} = 300$ et $\delta_{ag} = 5$
Pour Chaque mouvement de caméra détecté **Faire**
  Si Le mouvement est une translation **Alors**
    Si $c_r < \delta_r$ **Alors**
      //Mouvement rectiligne
      Si $dpt > \delta_{dt}$ **Alors**
        //Mouvement d'amplitude conséquente
        Première et dernière images du segment sélectionnées
      Sinon
        Dernière image du segment sélectionnée
      Fin Si
    Sinon
      //Mouvement considéré comme non rectiligne
      Si $dpt > \delta_{dt}$ **Alors**
        //Mouvement d'amplitude conséquente
        Images au début, au milieu et à la fin du segment sélectionnées
      Sinon
        Dernière image du segment sélectionnée
      Fin Si
    Fin Si
  Sinon Si Le mouvement est un zoom **Alors**
    Si $ag > \delta_g$ **Alors**
      Première et dernière images du segment sélectionnées
    Sinon
      Dernière image du segment sélectionnée
    Fin Si
  Sinon
    //Segment statique
    Image au milieu du segment choisie
  Fin Si
Fin Pour

La figure 5.1 illustre la sélection des images en fonction de l'amplitude des mouvements. Il s'agit d'une séquence vidéo nommée « Baseball » qui est un extrait d'un match de baseball et qui possède 9 plans. En partant du bas vers le haut sur l'axe des ordonnées, nous avons respectivement la position des plans, l'identification du statique (absence de mouvement), de la translation et du zoom, et enfin la sélection des images clés. Par exemple, le plan n°1 qui va de l'image 0 à l'image 59 est identifié comme statique et l'image clé correspond à l'image 29. De même, le plan n°7 qui va de l'image 378 à l'image 503 contient deux segments, un segment statique de l'image 378 à l'image 448 suivi d'un segment de zoom de l'image 449 à l'image 503. Les images clés sélectionnées pour ce plan sont les images 413 et 503.

Nous pouvons constater que cette approche est adaptée pour les plans qui ne présentent qu'un seul type de mouvement (par exemple, plans n°1 et n°2). Le nombre d'images est suffisant pour décrire le plan. En revanche, pour les plans avec beaucoup de mouvements de caméra (plans n°3 et n°7), le nombre d'images sélectionnées par cette approche n'est pas toujours adéquat pour les résumer. Par exemple, le plan n°3 possède quatre images clés, ce qui est relativement important pour le décrire. La figure 5.2 montre la vidéo « Baseball » (échantillonnage de 1 image sur 25) et les images clés sélectionnées en fonction de l'amplitude des mouvements de caméra.

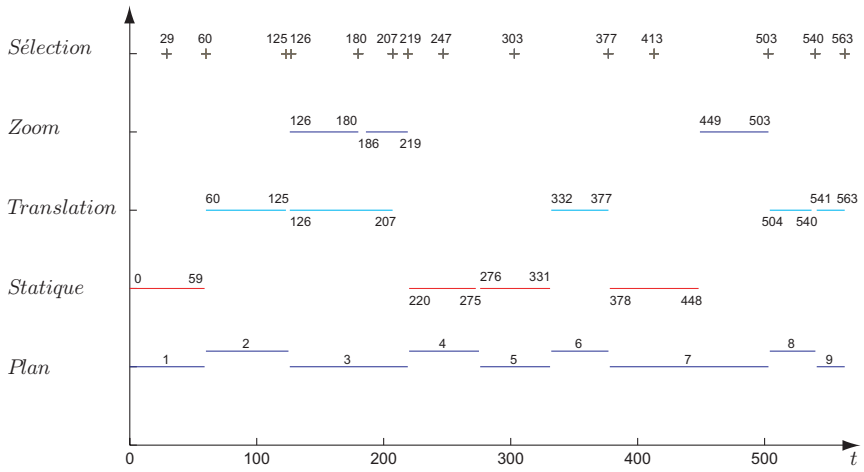
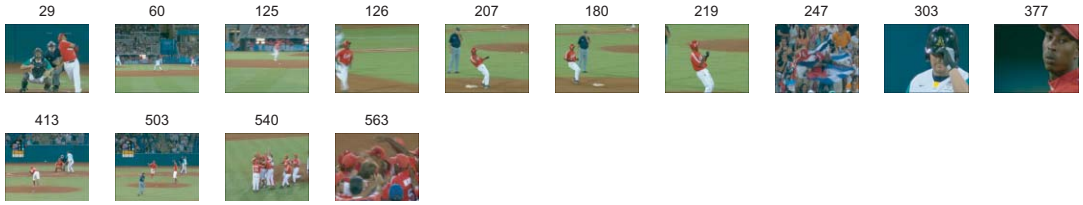


FIG. 5.1 – Sélection des images clés pour la vidéo « Baseball » en fonction de l'amplitude des mouvements de caméra. La vidéo possède 9 plans et contient différents mouvements (zoom, translation et statique). Par exemple, le premier plan est statique et l'image située au milieu du plan est sélectionnée pour le représenter. D'autres plans possèdent plusieurs mouvements successifs ou superposés (par exemple plan n°3).

Les règles définies permettent de sélectionner des images clés pour représenter la vidéo. Cependant, celles-ci ne tiennent pas compte des mouvements de caméra voisins. Par exemple, si on considère un plan possédant deux mouvements : statique suivi d'une translation de forte amplitude. Avec ces règles, une image sera sélectionnée dans la zone statique et deux images au début et à la fin du segment de translation. On note qu'une redondance apparaît entre les deux premières images sélectionnées. La prise en compte de l'enchaînement des mouvements de caméra pourra supprimer cette redondance.

(a) Echantillonnage de la vidéo (1 image sur 25).



(b) Images sélectionnées en fonction de l'amplitude des mouvements.

FIG. 5.2 – Résumé de la vidéo « Baseball » en fonction de l'amplitude des mouvements.

5.2.2 Résumé fonction de l'enchaînement des mouvements de caméra

Cette approche consiste à déterminer les images clés en fonction de l'enchaînement des mouvements détectés. Il faut tout d'abord trier par ordre chronologique les différents segments de mouvement dans chaque plan, puis proposer une technique de sélection d'images suivant l'enchaînement des mouvements.

5.2.2.1 Tri par ordre chronologique des mouvements de caméra

L'algorithme du tri rapide ou « quicksort » inventé par C.A.R Hoare [Hoa61] a été employé et adapté pour trier les segments de mouvement dans l'ordre chronologique. Le principe du tri est de considérer un tableau d'éléments où chaque élément comprend deux termes : l'indice de début et l'indice de fin du segment où un mouvement a été détecté. L'élément situé au milieu du tableau est ensuite considéré comme l'élément séparateur. L'algorithme divise en deux le tableau : l'un avec les éléments inférieurs ou égaux à l'élément séparateur et l'autre avec les éléments supérieurs. Puis le tri se poursuit de manière itérative sur les deux sous-tableaux jusqu'à obtenir tous les éléments triés.

Cependant, cet algorithme nécessite la définition d'une relation d'ordre entre les éléments du tableau. Or les segments qui sont fournis par la méthode de classification des mouvements de caméra et qui forment les éléments à trier peuvent présenter des chevauchements ou des inclusions. Le chevauchement peut effectivement exister entre deux mouvements (zoom et translation) ou être présent en raison du filtrage temporel introduit par la méthode de classification. De même, l'inclusion peut aussi avoir lieu entre deux mouvements de caméra (zoom et translation).

L'adaptation de l'algorithme de tri concerne la comparaison entre deux éléments du tableau. Elle consiste à effectuer un test sur le chevauchement des segments. Si tel est le cas,

le segment qui a le plus petit indice de début est considéré comme inférieur à l'autre. Nous vérifions ensuite si les deux segments ne sont pas inclus l'un dans l'autre. Si une inclusion est détectée, nous choisissons par convention le segment inclus comme inférieur à l'autre. Enfin, dans le cas où les segments ne présentent pas de chevauchement ou d'inclusion, le segment qui a l'indice de début le plus petit est défini comme inférieur à l'autre.

La figure 5.3 montre un exemple de tri de segments comprenant une inclusion et un chevauchement. Le premier terme de chaque élément correspond au numéro de l'image où un mouvement de caméra commence et le deuxième correspond à la fin du mouvement.

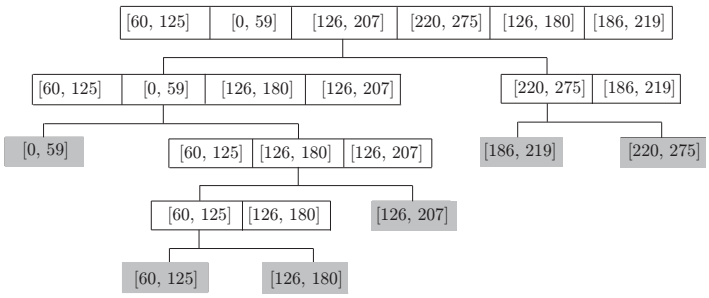


FIG. 5.3 – Exemple de tri de segments par ordre chronologique.

Une fois les segments de mouvements triés, le chevauchement entre les segments est supprimé. Si un chevauchement est présent entre deux segments, l'image au milieu de la zone de chevauchement est choisie comme fin du premier segment et début du segment suivant. La figure 5.4.a montre un exemple où plusieurs mouvements sont présents à l'intérieur d'un même plan. Les segments sont triés par ordre chronologique et le tri est affiché au-dessus de chaque segment de mouvement de caméra. Le segment de l'image 126 à l'image 207 et le segment de l'image 186 à l'image 219 se chevauchent et sont remplacés dans la figure 5.4.b par le segment de l'image 126 à de l'image 196 et le segment de l'image 197 à l'image 207. L'inclusion de l'image 126 à l'image 180 n'est en revanche pas modifiée.

5.2.2.2 Sélection des images clés en fonction de l'enchaînement des mouvements

Nous allons créer un résumé de vidéo qui repose sur l'enchaînement des mouvements de caméra. Pour réduire la redondance des images et représenter l'intégralité du contenu de la vidéo, nous avons choisi de sélectionner les images uniquement au début, au milieu ou à la fin d'un segment. Nous avons alors établi des règles heuristiques qui dépendent de l'enchaînement de deux mouvements de caméra pour obtenir les images clés. Par exemple, si un segment statique est suivi d'un mouvement de translation, la fin du statique et la fin de la translation sont choisies comme images clés. Ainsi les informations apportées par le mouvement de caméra sont directement utilisées pour déterminer les images clés. Le tableau 5.1 récapitule comment les images clés sont sélectionnées en fonction de deux mouvements consécutifs. Le procédé est répété de manière itérative sur tous les segments de mouvement du plan.

Cette technique ne traite que deux mouvements consécutifs. Or la sélection d'images suivant les deux premiers mouvements peut paraître redondante par rapport aux deux mouvements suivants. Supposons par exemple que trois mouvements consécutifs soient détectés dans un plan : statique, translation puis statique. En appliquant les règles définies dans le tableau 5.1, nous obtenons les résultats montrés dans le tableau 5.2. Les itérations corres-

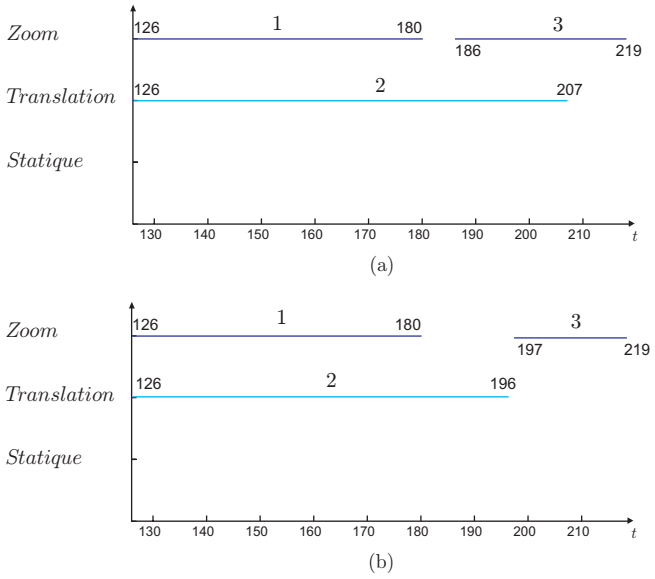


FIG. 5.4 – Exemple de tri des segments sur un plan qui comprend une inclusion et un chevauchement. (a) Tri par ordre chronologique des segments de mouvements de caméra (Les numéros au-dessus des segments correspondent à l'ordre chronologique). (b) Suppression des chevauchements.

TAB. 5.1 – Règles de sélection des images clés ([* * *]=au début, au milieu ou à la fin du segment). La colonne de gauche représente le premier mouvement détecté et la première ligne correspond au second mouvement détecté. Par exemple, si un segment statique est suivi d'un segment de translation, la dernière image du segment statique sera sélectionnée ainsi que la dernière image de la translation.

2^e mouvement / 1^{er} mouvement	Statique	Translation	Zoom
Statique		[0 0 1 , 0 0 1]	[0 0 1 , 0 0 1]
Translation	[1 0 0 , 1 0 0]		[1 0 0 , 1 0 0]
Zoom	[1 0 0 , 1 0 0]	[1 0 0 , 1 0 0]	

pondent au traitement de deux mouvements. Nous effectuons ensuite un « ou logique » pour déterminer les images clés sur tout le plan. Il est immédiat de constater la redondance des images de début et de fin du segment 2. Pour s'assurer que l'image de fin d'un segment ne soit pas sélectionnée en même temps que l'image du début du segment suivant, un post filtrage est réalisé au sein de chaque plan. Il consiste à détecter la présence de deux 1 consécutifs et à supprimer le deuxième. Il a pour effet de supprimer les images voisines. Finalement le résultat reste cohérent puisque deux images sont sélectionnées : une à la fin du segment statique 1 (ou début du segment de translation) et une à la fin du segment de translation (ou au début du segment statique 3).

Tab. 5.2 – Illustration de la sélection des images clés. La première itération correspond au traitement des mouvements 1 et 2. De même, la deuxième itération correspond à l'enchaînement des deux mouvements suivants (segments 2 et 3). Après avoir réalisé un « ou logique » entre les différentes itérations, les images clés sont obtenues.

	Segment 1	Segment 2	Segment 3
Mouvement détecté	Statique	Translation	Statique
1^{re} itération	[0 0 1]	[0 0 1]	
2^{e} itération		[1 0 0]	[1 0 0]
Ou logique	[0 0 1]	[1 0 1]	[1 0 0]
Final (filtrage)	[0 0 1]	[0 0 1]	[0 0 0]

Enfin, si un plan ne contient qu'un seul mouvement de caméra, alors les règles suivantes sont appliquées pour sélectionner les images. Si le plan est statique, l'image au milieu du plan est choisie comme image clé. En revanche, si le mouvement de caméra est une translation ou un zoom dans le plan alors l'image de fin du plan est choisie. La figure 5.5 présente la sélection des images en fonction de l'enchaînement des mouvements de caméra sur la vidéo « Baseball » et la figure 5.6 montre les images clés correspondantes pour le résumé. Deux plans possèdent une succession de mouvements (plan n°3 et plan n°7). Par exemple, le plan n°7 comprend une partie statique suivie d'un zoom et la sélection des images se situe à la fin du segment statique et à la fin du zoom. En revanche, la figure 5.7 montre un exemple d'un plan de la vidéo « Journal » qui possède beaucoup de mouvements successifs. Nous pouvons observer que le segment statique de l'image 994 à l'image 1006 possède deux images clés. Ceci montre la limite de cette approche. En effet, prendre deux images pour représenter un segment statique peut paraître redondant. Une autre limite est la non prise en compte des amplitudes des mouvements. Par exemple, un mouvement de translation de faible amplitude ou de courte durée nécessite moins d'images qu'une translation de forte amplitude. De plus, l'inclusion des mouvements n'est également pas considérée par cette approche.

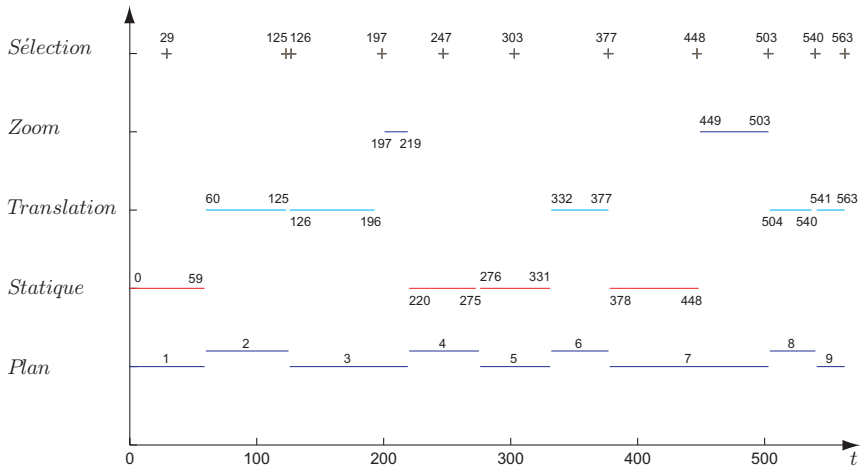


FIG. 5.5 – Images sélectionnées pour la vidéo « Baseball » en fonction de l'enchaînement des mouvements. La vidéo possède 9 plans et contient différents mouvements. Les plans n°3 et n°7 possèdent plusieurs mouvements et les images sont sélectionnées en fonction de l'enchaînement des mouvements.

FIG. 5.6 – Résumé de la vidéo « Baseball » en fonction de l'enchaînement des mouvements.

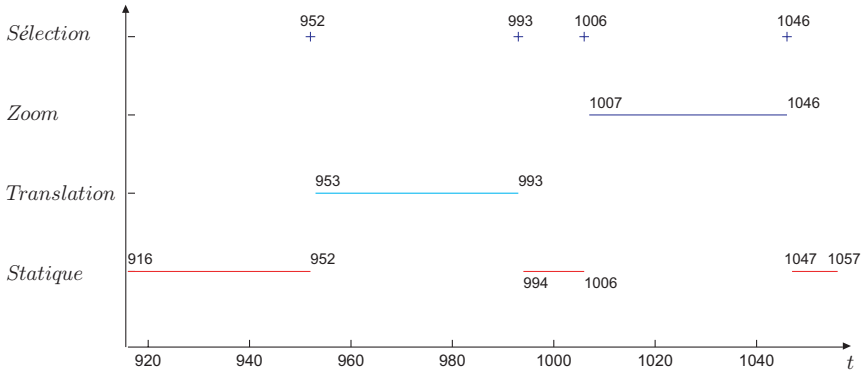


FIG. 5.7 – Images sélectionnées pour un plan de la vidéo « Journal » en fonction de l'enchaînement des mouvements. Par exemple, le segment statique de l'image 994 à l'image 1006 possède deux images clés pour le représenter, ce qui peut être un inconvénient de l'approche.

5.2.3 Résumé fonction de l'amplitude et de l'enchaînement des mouvements de caméra

Cette approche est une combinaison des deux précédentes. Il s'agit de prendre en compte l'amplitude et l'enchaînement des mouvements de caméra. Après un tri des segments de mouvement et une suppression des chevauchements, les mouvements détectés qui ont une amplitude ou une durée faible sont traités comme des segments statiques. S'il s'agit d'un mouvement de translation de durée T avec un déplacement total dpt, le déplacement total normalisé $dpt_n = dpt/T$ est alors calculé. Celui-ci sera considéré comme un segment statique si la durée T est inférieure à un seuil δ_T et si le déplacement total normalisé dpt_n est inférieur à un seuil δ_n. De même, un zoom de durée T avec un agrandissement ag est considéré comme un segment statique si la durée T est inférieure à un seuil δ_T et si l'agrandissement ag est inférieur δ_g. Dans notre expérimentation, les seuils ont été fixés de manière empirique à $\delta_n = 1.5$, $\delta_g = 1.8$ et $\delta_T = 50$ images.

Les images sont ensuite sélectionnées suivant les règles établies lors de l'enchaînement des mouvements. La seule différence, dans le cas de mouvements de forte amplitude, est que des images clés peuvent être ajoutées pour le résumé. Un mouvement sera déclaré de forte amplitude, si l'approche de sélection des images clés suivant l'amplitude des mouvements nécessite au moins deux images pour le représenter. Dans ce cas, les images choisies par l'approche de résumé suivant l'amplitude des mouvements de caméra sont ajoutées au résumé.

Si on reprend l'exemple précédent avec trois mouvements consécutifs détectés dans un plan : statique, translation de forte amplitude puis statique. Les images clés sont obtenues en appliquant les règles sur l'amplitude et l'enchaînement des mouvements. Le tableau 5.3 illustre la manière de sélectionner les images. Si le mouvement de translation est représenté par les images de début, du milieu et de fin du segment selon les règles établies suivant l'amplitude des mouvements, alors ces images sont conservées. Un « ou logique » est simplement effectué entre les sélections des images suivant l'amplitude et l'enchaînement des mouvements pour déterminer les images clés sur le plan. Cela permet de conserver l'image au centre du segment de translation. Le post filtrage est ensuite réalisé à chaque fin de segment pour supprimer les redondances éventuelles entre la fin d'un segment et le début du suivant. Finalement trois images sont sélectionnées pour décrire le plan : une image à la fin du segment statique 1 (ou au début du segment de translation), une au milieu du segment de translation et une à la fin du segment de translation (ou au début du segment statique 3).

Tab. 5.3 – Illustration de la sélection des images clés suivant l'amplitude et l'enchaînement des mouvements. Les images clés sont obtenues en réalisant un « ou logique ». Le filtrage temporel est réalisé (sur chaque plan) à chaque fin de segment pour supprimer les images voisines entre la fin d'un segment et le début du suivant.

	Segment 1	Segment 2	Segment 3
Mouvement détecté	Statique	Translation	Statique
Amplitude	[0 0 0]	[1 1 1]	[0 0 0]
1^{re} itération	[0 0 1]	[0 0 1]	
2^{e} itération		[1 0 0]	[1 0 0]
Ou logique	[0 0 1]	[1 1 1]	[1 0 0]
Final (filtrage)	[0 0 1]	[0 1 1]	[0 0 0]

En revanche, si le segment est de courte durée (inférieure au seuil δ_T), seule l'image au milieu du segment est choisie pour le représenter. De la même manière, les images sélectionnées pour décrire un segment statique sont remplacées par l'image située au milieu du segment. Cela évite de sélectionner une image en début et en fin d'un segment statique. En ce qui concerne les inclusions, si le mouvement inclus est de forte amplitude, alors le segment le contenant est également décrit par l'image au centre de celui-ci.

La figure 5.8 présente la sélection des images suivant l'amplitude et l'enchaînement des mouvements de caméra pour la vidéo « Baseball » et la figure 5.9 montre les images clés correspondantes. Par exemple, le plan n°7 comprend une partie statique suivie d'un zoom, la différence par rapport à la méthode précédente est la sélection des images qui se situe au milieu du segment statique et à la fin du zoom, et non à la fin du statique et à la fin du zoom. De plus, les plans où un seul mouvement est détecté ont des images sélectionnées en fonction de l'amplitude du mouvement.

La figure 5.10 montre l'exemple du plan de la vidéo « Journal » contenant beaucoup de mouvements successifs. Les images sont sélectionnées sur les segments statiques et aucune n'est choisie sur les segments de translation et de zoom. En effet, ces mouvements sont intercalés entre deux segments statiques qui sont supposés représenter le début et la fin de ces mouvements. Finalement, nous présentons dans la figure 5.11 le résumé de la vidéo « Documentaire » suivant les trois variantes de la méthode de résumé. Nous pouvons remarquer que le résumé fonction de l'amplitude et de l'enchaînement des mouvements est le résumé qui fournit le moins d'images clés avec 34 images contre 40 pour le résumé fonction de l'enchaînement des mouvements et 56 pour le résumé fonction de l'amplitude des mouvements. Pour chaque plan de la vidéo « Documentaire », c'est le résumé créé à partir de l'amplitude et de l'enchaînement des mouvements de caméra qui présente visuellement le moins de redondance.

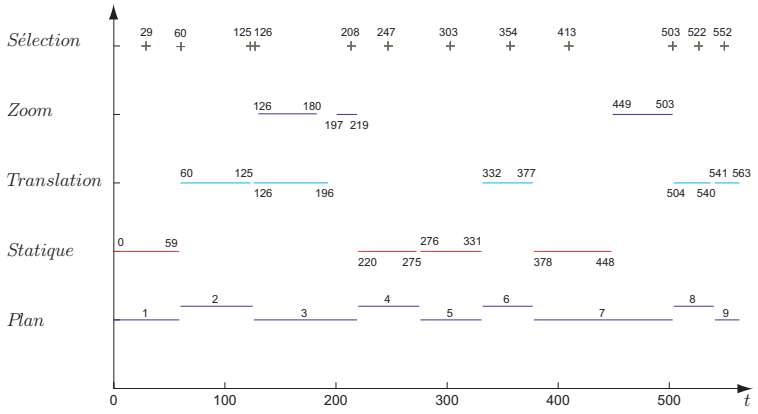


FIG. 5.8 – Images sélectionnées pour la vidéo « Baseball » en fonction de l'amplitude et de l'enchaînement des mouvements de caméra.

Nous avons développé une méthode de résumé qui exploite l'information apportée par le mouvement de caméra. Cette méthode se décline en 3 variantes : résumé selon l'amplitude des mouvements de caméra, résumé selon l'enchaînement des mouvements de caméra et résumé selon l'amplitude et l'enchaînement des mouvements de caméra (résultat de la combinaison des deux premières variantes). Le résumé obtenu par cette dernière semble donner visuellement de bons résultats. Toutefois, afin de valider cette méthode de sélection des images clés, nous proposons la conception d'une méthode d'évaluation pour juger de la performance des résumés.

FIG. 5.9 – Résumé de la vidéo « Baseball » en fonction de l'amplitude et de l'enchaînement des mouvements de caméra.

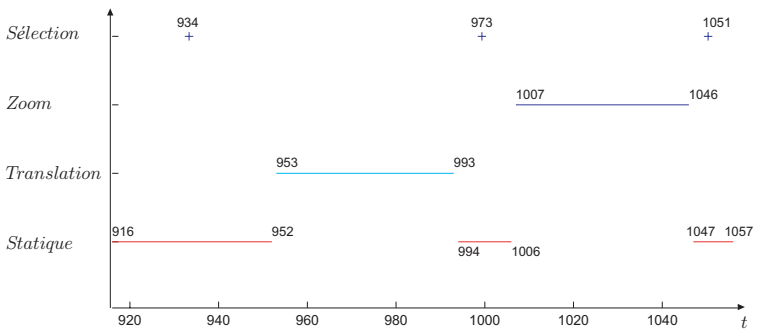
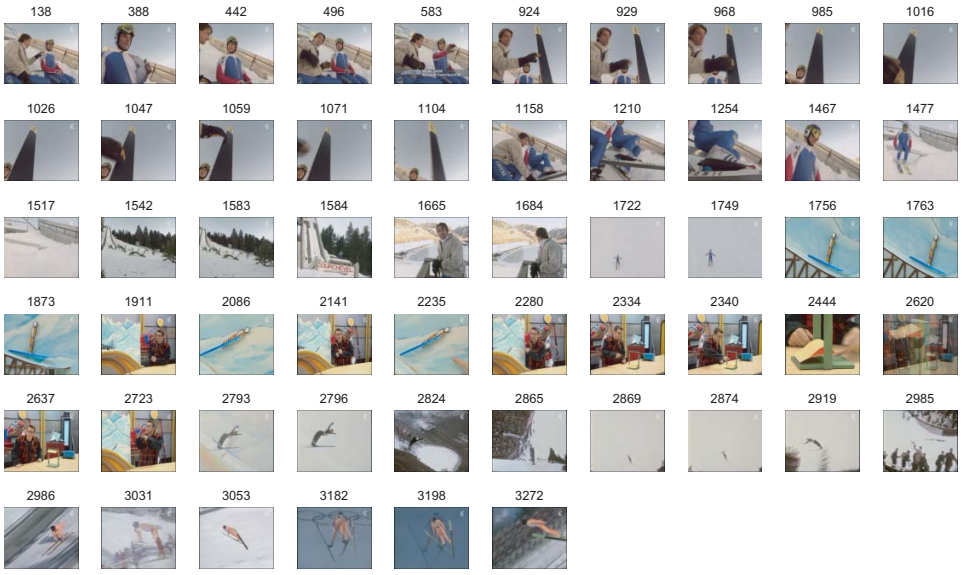
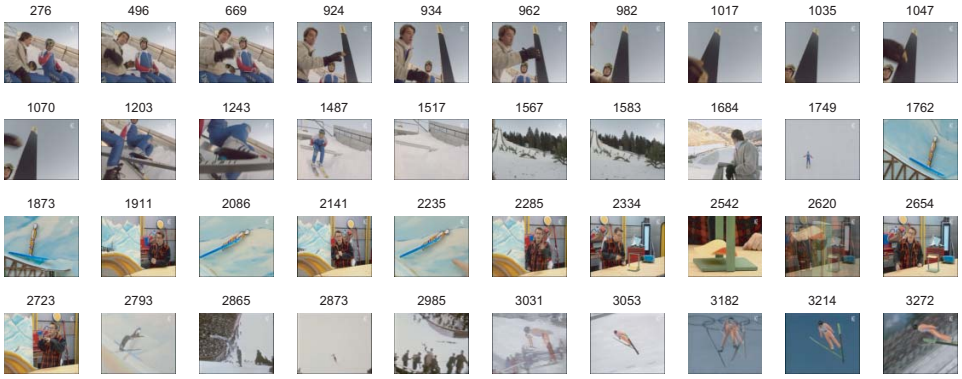


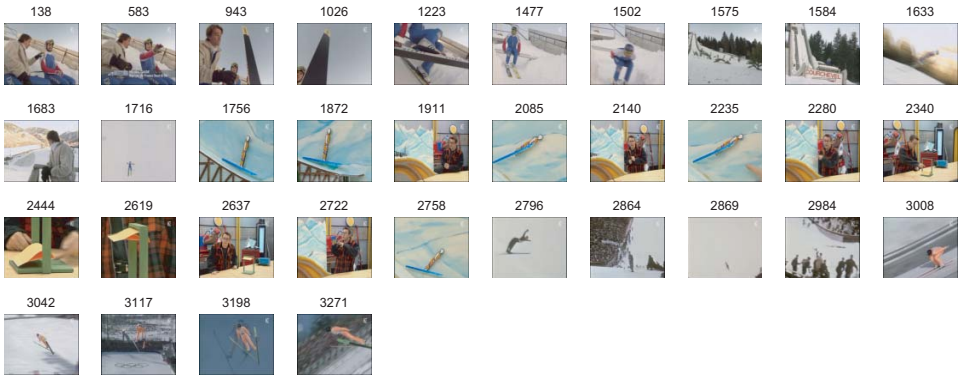
FIG. 5.10 – Images sélectionnées pour un plan de la vidéo « Journal » en fonction de l'amplitude et de l'enchaînement des mouvements. L'image au milieu de chaque segment statique est sélectionnée comme image clé. Comme les mouvements de translation ou de zoom ne sont pas d'amplitude suffisante, seuls le début et la fin de ces segments sont décrits par les images des segments statiques.



(a) Images sélectionnées en fonction de l'amplitude des mouvements.



(b) Images sélectionnées en fonction de l'enchaînement des mouvements.



(c) Images sélectionnées en fonction de l'amplitude et de l'enchaînement des mouvements.

FIG. 5.11 – Résumé de la vidéo « Documentaire » suivant les trois variantes de résumé.

5.3 Évaluation des méthodes de création de résumé de vidéo

Les méthodes de résumé de vidéo doivent être évaluées pour s'assurer de la pertinence des images proposées. Cependant, la qualité d'un résumé de vidéo repose sur des considérations subjectives. C'est le jugement humain qui détermine la valeur d'un résumé. Tout d'abord, nous allons donner un bref descriptif des méthodes existantes concernant l'évaluation des résumés. Puis, nous exposerons notre méthode d'évaluation.

5.3.1 Méthodes d'évaluation

Comme déjà mentionnées précédemment, les méthodes de création de résumé de vidéo ont été très étudiées dans la littérature. Néanmoins, il n'existe pas de méthode standard pour évaluer les différents résumés de vidéo. La qualité d'un résumé de vidéo est difficile à quantifier et repose souvent sur des mesures d'évaluation subjectives. Il est en général demandé à des sujets de donner leur opinion sur des résumés fournis par leur méthode.

Quatre familles d'évaluation peuvent être dégagées pour juger des méthodes de création de résumé de vidéo.

La première famille de méthodes permet la comparaison de résumés. L'expérience consiste à visionner la vidéo puis à présenter deux résumés : l'un issu de sa propre méthode de création de résumé et l'autre provenant d'une méthode concurrente. Dans [Cor04], il est demandé aux sujets de choisir le résumé qui représente le mieux le contenu. La méthode concurrente peut provenir d'un échantillonnage régulier de la vidéo ou d'une version simplifiée de sa propre méthode. L'objectif est donc de montrer par l'intermédiaire des réponses des sujets que le résumé proposé par sa méthode est meilleur que le résumé concurrent. Ce genre d'approche est difficile à généraliser lorsqu'il y a plus d'un résumé à comparer. De plus, le résumé concurrent est souvent créé de manière simple et n'est pas forcément très performant pour représenter la vidéo. John Boreczky et al. [Bor00] proposent une évaluation plus sophistiquée qui se divise en deux phases : une évaluation sur la capacité à se déplacer dans le résumé, puis une comparaison de différents résumés. Pour chaque vidéo, la première phase consiste à retrouver des segments pertinents dans le résumé qui répondent aux questions posées. La mesure d'évaluation est le temps de réponse moyen mis par les sujets suivant les résumés proposés. La seconde phase permet de comparer trois méthodes de résumé. Deux résumés sont présentés, l'un à gauche de l'écran et l'autre à droite. Pour chaque paire présentée, plusieurs questions sont posées comme par exemple quel est le résumé le plus efficace ou le résumé le plus plaisant visuellement. Les réponses moyennes sur tous les sujets sont étudiées et permettent de conclure sur la méthode de résumé la plus performante.

La deuxième famille consiste à créer manuellement un résumé, sorte de vérité terrain de la vidéo. Celui-ci peut être obtenu soit en sélectionnant des images représentatives pour décrire une vidéo soit en décomposant la vidéo en scènes. Dans ce dernier cas, la notion de scène est alors définie et représente le regroupement de plans juxtaposés ou non qui possèdent la même unité de lieu. Puis, la méthode de résumé automatique est comparée à celle de la vérité terrain et des mesures comme le rappel et la précision sont déterminées pour évaluer le résumé. Néanmoins, la comparaison (regroupement d'images clés similaires ou comparaison de scènes) est effectuée soit manuellement soit par des distances auxquelles il est délicat de donner un sens. Par exemple, Ferman et al. [Fer03] évaluent leur résumé en demandant à un observateur neutre de signaler les images clés oubliées et les images clés redondantes. Les critères d'évaluation sont donc le nombre d'images clés oubliées et le nombre d'images

redondantes. Dans [Che03b, Odo03, Zhu03a], leur méthode de résumé est évaluée en comparant les scènes détectées à celles obtenues par la vérité terrain.

La troisième famille demande à des sujets de mesurer le niveau de satisfaction du résumé proposé. Un sujet visionne une vidéo puis il lui est demandé de juger le résumé selon une échelle donnée. Des questions peuvent aussi être posées aux sujets pour mesurer le degré de performance du résumé proposé. Dans [Sha04], la qualité du résumé est évaluée en demandant aux sujets d'attribuer une note entre un et cinq pour quatre critères : clarté, concision, cohérence, qualité globale. De même, le résumé de vidéo dans [Yu03] est évalué suivant la clarté, la concision et la cohérence. Dans [Ma02], le sujet doit d'abord donner une appréciation pour chaque plan sur l'unique image clé sélectionnée (bonne, mauvaise ou neutre) puis il doit porter une appréciation sur le nombre d'images clés par plan (correct, beaucoup trop, trop peu). Dans [Lu04], trois questions sont posées sur le résumé : qui, quoi et cohérence. En fonction du nombre d'acteurs retournés (qui ?), du nombre d'événements retournés (quoi ?) et du score indiqué par les sujets sur la cohérence du résumé, la qualité de leur résumé est mesurée. Ngo et al. [Ngo03] proposent deux critères d'évaluation pour juger le résumé : le degré d'information et le degré de plaisance (agréable). Le premier critère renseigne sur la capacité du résumé à représenter toute l'information de la vidéo en évitant la redondance et le second évalue la performance de l'algorithme à fournir des segments agréables pour le résumé. Dans leur expérimentation, un résumé est présenté à 10% de la longueur de la vidéo originale, puis un autre à 25%, et enfin la vidéo originale. Il est demandé à chaque sujet, après chaque visionnage, d'assigner un score entre zéro et cent pour les deux critères considérés. Dans [Lu05], les résumés sont évalués en introduisant le degré de signification. Le résumé est présenté à 15% de la longueur de la vidéo originale, puis à 30%. Il est demandé à chaque sujet, après chaque visionnage, de répondre à des questions sur les événements majeurs dans la vidéo (comme qui fait quoi ?). En fonction des réponses données par le sujet, un score est déterminé entre zéro et cent sur le degré de signification du résumé. Un autre critère est aussi utilisé et consiste simplement à déterminer le meilleur résumé (à un même niveau de compression) entre deux méthodes de résumé. Néanmoins il est assez difficile de juger un résumé proposé suivant une échelle donnée. Si la granularité de l'échelle est trop petite, le jugement sur le niveau de satisfaction sera d'autant plus délicat à donner puisque la différence entre deux niveaux sera mal perçue. La granularité de l'échelle entre les sujets n'est pas nécessairement cohérente puisqu'il est difficile de faire la différence entre un bon et un très bon résumé. Nous pouvons également constater que chaque auteur a employé ses propres critères d'évaluation. De plus, l'évaluation s'effectue entre la vidéo originale et le résumé proposé. La comparaison entre plusieurs résumés n'est pas immédiate.

Devant la subjectivité de ces méthodes d'évaluation, la quatrième famille propose de définir des mesures objectives et applicables sur n'importe quel type de vidéos. L'avantage de ces approches est qu'elles ne nécessitent pas la participation d'experts pour évaluer le résumé. Par exemple, dans [Yah01a], une mesure objective est définie. Elle consiste à tirer des extraits de longueur donnée dans la vidéo et à vérifier qu'il existe au moins une image du résumé qui soit similaire à chacun des extraits. Ciocca et al. [Cio05] utilisent trois mesures objectives pour évaluer le résumé. Une mesure de fidélité est définie en construisant une distance entre les images clés et les images de la vidéo. Plus cette distance est faible, plus le résumé est fidèle à la vidéo. Une mesure sur le degré de reconstruction par plan est également considérée. Elle consiste à interpoler les images de chacun des plans à partir des images clés puis à calculer une distance entre le plan original et celui interpolé. Enfin, la dernière mesure est le taux de compression entre la longueur de la vidéo et la longueur du résumé. Dans [Gon01], une

métrique sur la redondance de la vidéo est élaborée.

L'inconvénient des méthodes d'évaluation objective est qu'elles reposent sur des comparaisons entre des descripteurs de bas niveau. Ces mesures ne peuvent pas refléter la qualité sémantique du résumé. Comme pour la compression des images, l'image avec le rapport signal sur bruit le plus élevé (mesure objective) n'est pas forcément celle qui a la meilleure qualité visuelle. La qualité d'un résumé doit être fournie par le jugement humain. En ce qui concerne les méthodes subjectives (intervention du jugement humain), elles présentent l'inconvénient de ne pas pouvoir être réutilisées automatiquement. De plus, ces approches sont souvent manuelles et nécessitent un temps important pour évaluer la méthode de création de résumé.

Nous allons créer une méthode d'évaluation classée dans la deuxième famille. Elle consiste à créer un résumé de référence (vérité terrain) à partir de résumés obtenus par différents sujets, puis à effectuer une comparaison automatique entre le résumé de référence et les résumés soumis par différentes méthodes. Cette technique d'évaluation présente l'intérêt de pouvoir tester rapidement n'importe quel type de méthode de résumé. En effet, la construction du résumé de référence est une étape fastidieuse qui nécessite l'intervention de sujets mais une fois celui-ci obtenu, la comparaison entre le résumé de référence et le résumé fourni par une méthode est immédiate. Cette méthode permet également de prendre en compte les approches de résumé qui sélectionnent des images clés au niveau de chaque plan et les approches de résumé hiérarchique qui fournissent des résumés de taille variable.

La méthode d'évaluation que nous proposons est similaire à celle de Huang et al. [Hua04]. Néanmoins, bien que leur évaluation se passe au niveau de la vidéo, leur méthode de construction du résumé de référence s'effectue au niveau des plans. De plus, ils n'ont pas de connaissance sur le degré d'importance d'un plan par rapport à un autre, et donc ils ne définissent pas de degré d'importance entre les images. Leur évaluation ne permet pas de prendre en compte les méthodes de résumé qui fournissent une mesure d'importance pour chaque image clé sélectionnée. De plus, le résumé de référence n'est pas hiérarchique. La longueur est fixée par les sujets et ne permet pas de juger des approches « grossier à fin » (« coarse to fine » en anglais) qui demandent en entrée un nombre d'images clés souhaité par l'utilisateur. La méthode d'évaluation automatique que nous proposons a été développée dans un cadre plus général et présente l'avantage de résoudre les problèmes cités ci-dessus : importance des images clés les unes par rapport aux autres, taille du résumé variable, prise en compte des résumés hiérarchiques. Notre méthode repose sur une expérience qui permet à un sujet de créer manuellement un résumé de vidéo selon différents niveaux de résolution.

5.3.2 Création d'un résumé par un sujet

Le but de l'expérience est de concevoir un résumé pour différentes vidéos données. Il s'agit de demander à des sujets de regarder une vidéo puis de créer manuellement un résumé. A partir des différents résumés, une méthode est proposée pour générer le résumé « optimal » appelé résumé de référence afin qu'il soit comparé aux résumés proposés par divers algorithmes.

5.3.2.1 Les vidéos choisies

La sélection des vidéos est une étape importante qui peut conditionner les résultats. Deux critères sont à prendre en compte : le contenu et la durée. Nous avons choisi trois vidéos avec des contenus variés et des durées différentes parmi celles présentées dans le chapitre 3 : le documentaire sportif (nommé « Documentaire ») sur le saut à ski avec 20 plans et 3271

images, la série « The Avengers » (nommée « Série ») avec 27 plans et 2412 images et le journal télévisé (nommé « Journal ») avec 42 plans et 6870 images. Chaque vidéo se compose d'images couleur (288x352 pixels) affichées à une fréquence de 25 images par seconde.

Il est à noter que ces vidéos sont de courte durée. La plus longue dure environ 5 minutes. A titre de comparaison, la vidéo la plus longue utilisée dans [Hua04] a 3114 images et a un nombre maximum de 20 plans. Le fait de ne pas choisir des vidéos de grande taille s'explique par la durée assez longue de l'annotation par un sujet. Il s'agit donc de trouver un bon compromis entre une durée suffisante des vidéos et une durée raisonnable pour l'expérience. Dans notre expérimentation, la création d'un résumé d'une vidéo nécessite entre 20 et 35 minutes par sujet.

5.3.2.2 Les sujets

12 personnes ont passé l'expérience pour chacune des trois vidéos. Tous les sujets avaient une vue normale ou corrigée et ils connaissaient les enjeux de l'expérience : création d'un résumé de vidéo.

5.3.2.3 Protocole expérimental

Les sujets passent l'expérience individuellement. Chaque sujet reçoit les consignes suivantes : D'une part, le résumé doit être aussi court que possible et préserver la totalité du contenu. D'autre part, le résumé doit être aussi neutre que possible. C'est donc le sujet qui discerne lui-même le degré d'acceptation du résumé.

La procédure de création d'un résumé de vidéo se déroule en trois étapes.

1^{re} *étape : Visionnage de la vidéo*

Dans la première étape, le sujet visionne d'abord une vidéo (images et son) puis un résumé oral lui est demandé afin de s'assurer que le contenu de la vidéo a été compris. Il visionne à nouveau la vidéo et peut ensuite ajouter, s'il le souhaite, des remarques sur la vidéo.

2^{e} *étape : Annotation des extraits de la vidéo*

Dans la deuxième étape, la vidéo est visionnée sous forme d'extraits présentés dans l'ordre chronologique en haut à gauche de l'écran (Fig. 5.12). Il est demandé au sujet d'indiquer le degré d'importance de l'extrait qu'il vient de visionner. Les extraits correspondent aux plans successifs de la vidéo. Ils sont présentés au sujet comme des extraits et aucune information n'est donnée sur les plans. Une fois l'extrait visionné, le sujet doit préciser le degré d'importance en indiquant, si selon lui, cet extrait est « très important », « important » ou « peu important » pour résumer la vidéo. Il clique simplement sur la notation correspondante en haut à droite de l'écran.

Puis il est demandé au sujet de choisir les images qui résument l'extrait. En bas à droite, les images sont présentées selon un échantillonnage régulier (une image sur dix). Le sujet doit sélectionner les images qui lui semblent être les plus représentatives pour le plan (de une au minimum à trois au maximum), sachant que la sélection doit être la plus concise possible et représenter l'intégralité de l'information.

Une fois que le sujet a terminé son annotation pour un extrait donné, il le valide et les résultats sont alors affichés dans la partie gauche en bas de l'écran pour garder en mémoire les annotations déjà effectuées.

Deux remarques peuvent être faites sur les choix de cette étape. La première concerne le nombre limité de niveaux d'importance. Seulement trois niveaux d'importance : « très important », « important » ou « peu important » ont été choisis. Une échelle avec plus de niveau aurait rendu la tâche plus délicate et peut-être déconcertante pour le sujet en raison de la difficulté à faire la différence entre les niveaux. La deuxième porte sur l'échantillonnage des images de l'extrait. Nous avons choisi l'échantillonnage d'une image sur dix pour éviter d'avoir le plan complet à afficher à l'écran, ce qui rendrait la tâche de sélection des images fastidieuse. Du fait de la redondance des images temporellement, il est apparu judicieux de faire cet échantillonnage et donc 5 images affichées à l'écran correspondent à 2 secondes de la vidéo.

La deuxième étape est illustrée dans la figure 5.12 et les images présentées proviennent de la vidéo « Documentaire ». Le sujet a ici indiqué que l'extrait était important pour résumer la vidéo. Il a également sélectionné une seule image (image n°2) pour résumer cet extrait. La sélection s'effectue par un clic de souris sur l'image qu'il souhaite et l'image correspondante est alors marquée par un cadre rouge. De plus, si l'image a été sélectionnée par erreur, le sujet peut la supprimer en cliquant à nouveau sur celle-ci. Comme les images sont affichées en bas à droite avec un sous-échantillonnage spatial par quatre, le sujet peut revoir l'image en résolution normale en passant la souris sur l'imagette. Celle-ci apparaîtra alors en haut à gauche. L'intérêt est d'être le plus insensible possible à l'effet de réduction des images et de présenter un grand nombre d'images à l'écran. Un autre intérêt de cet affichage est que le passage rapide d'une imagette à l'autre peut permettre de percevoir le mouvement entre les images malgré la sélection statique des images. L'annotation des extraits précédents est affichée en bas à gauche où ici 5 images ont été sélectionnées. Nous pouvons aussi apercevoir que l'imagette n°7 en bas à droite, ayant le curseur de la souris au dessus de celle-ci est affichée en résolution normale en haut à gauche. Finalement, cette étape va permettre de fournir un résumé au niveau de chaque plan de la vidéo.

3^{e} étape : Confirmation des annotations et construction d'un résumé court

Dans la troisième étape, une fois tous les extraits annotés, le résumé complet est affiché à l'écran. L'objectif est de fournir une vision globale du résumé pour le modifier et le valider. Chaque extrait est représenté par les images que le sujet a choisi et le degré d'importance est indiqué en dessous de chacune des images. Le passage d'un extrait au suivant est marqué par un espace au niveau des images. Il est alors demandé au sujet de modifier s'il le souhaite le degré d'importance des extraits en cliquant sur les images, puis de supprimer les images qui lui paraissent redondantes et enfin de n'en sélectionner qu'un nombre limité. Cette étape a pour but de fournir un résumé hiérarchique avec un niveau fin à l'échelle des plans et un niveau plus grossier à l'échelle de la vidéo. La figure 5.13 montre la troisième étape où il est demandé au sujet de modifier le degré d'importance des extraits.

Afin de bien comprendre l'expérience et le fonctionnement du logiciel, une phase d'apprentissage est effectuée avec une vidéo d'essai comportant 5 plans et 477 images. L'expérience est alors réalisée avec cette vidéo pour permettre au sujet de bien appréhender le protocole expérimental et de répondre à ses dernières interrogations. Néanmoins, l'annotation de cette vidéo n'est pas conservée. L'intérêt est d'éviter au maximum des effets sur l'annotation des vidéos dus à la non maîtrise du logiciel. De plus, l'ordre de présentations des vidéos est aléatoire d'un sujet à l'autre.

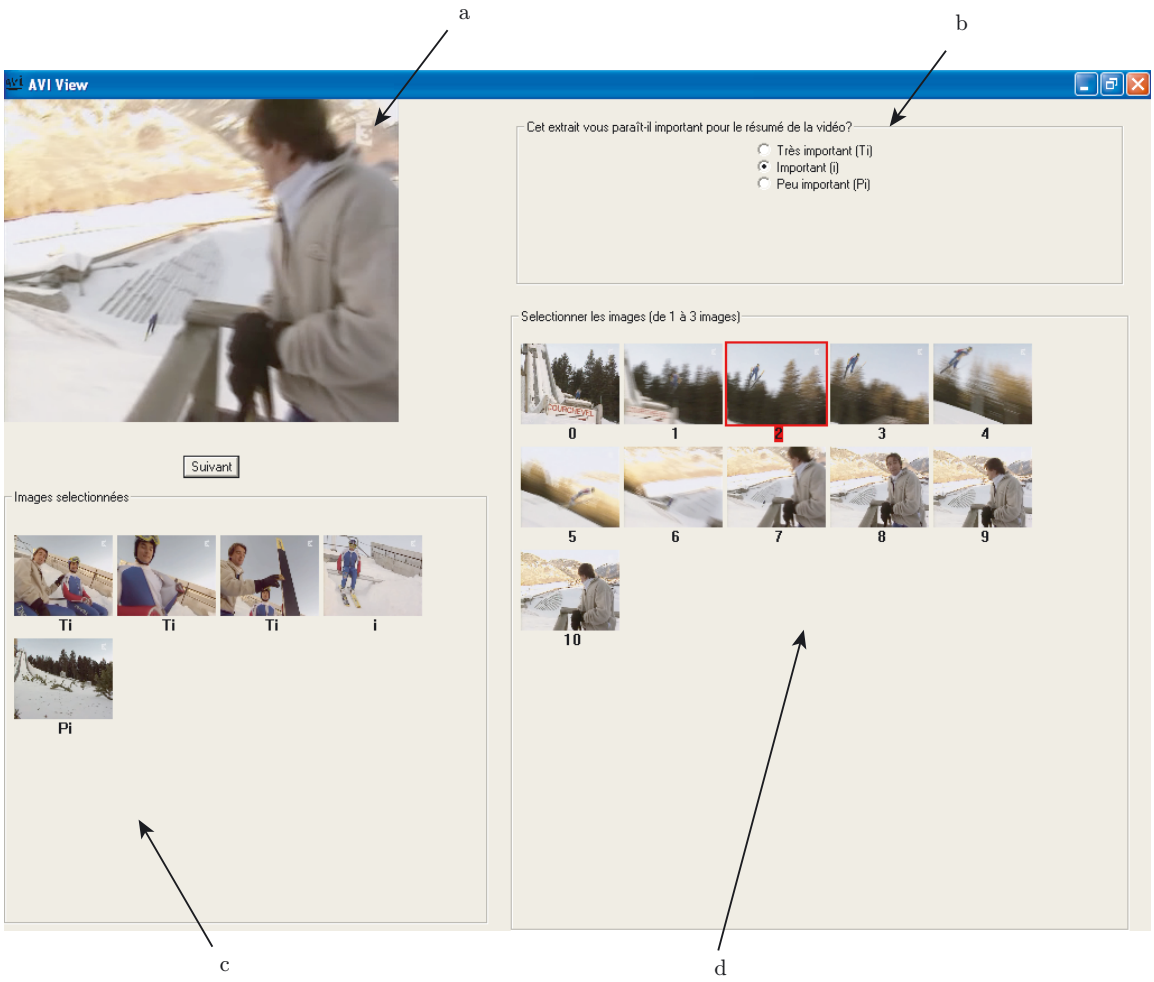


FIG. 5.12 – Deuxième étape de la création du résumé de référence pour la vidéo « Documentaire ». Le sujet doit indiquer le degré d'importance de l'extrait dans la zone b. Puis il doit sélectionner dans la zone d les images qui lui semblent pertinentes pour résumer l'extrait. Comme les images sont affichées avec un sous-échantillonnage spatial par quatre, le sujet peut les voir en résolution normale en passant la souris sur une imagette de la zone d pour que celle-ci apparaisse dans la zone a. Enfin, dans la zone c, les images déjà sélectionnées lors des extraits précédents sont affichées pour garder en mémoire la sélection.

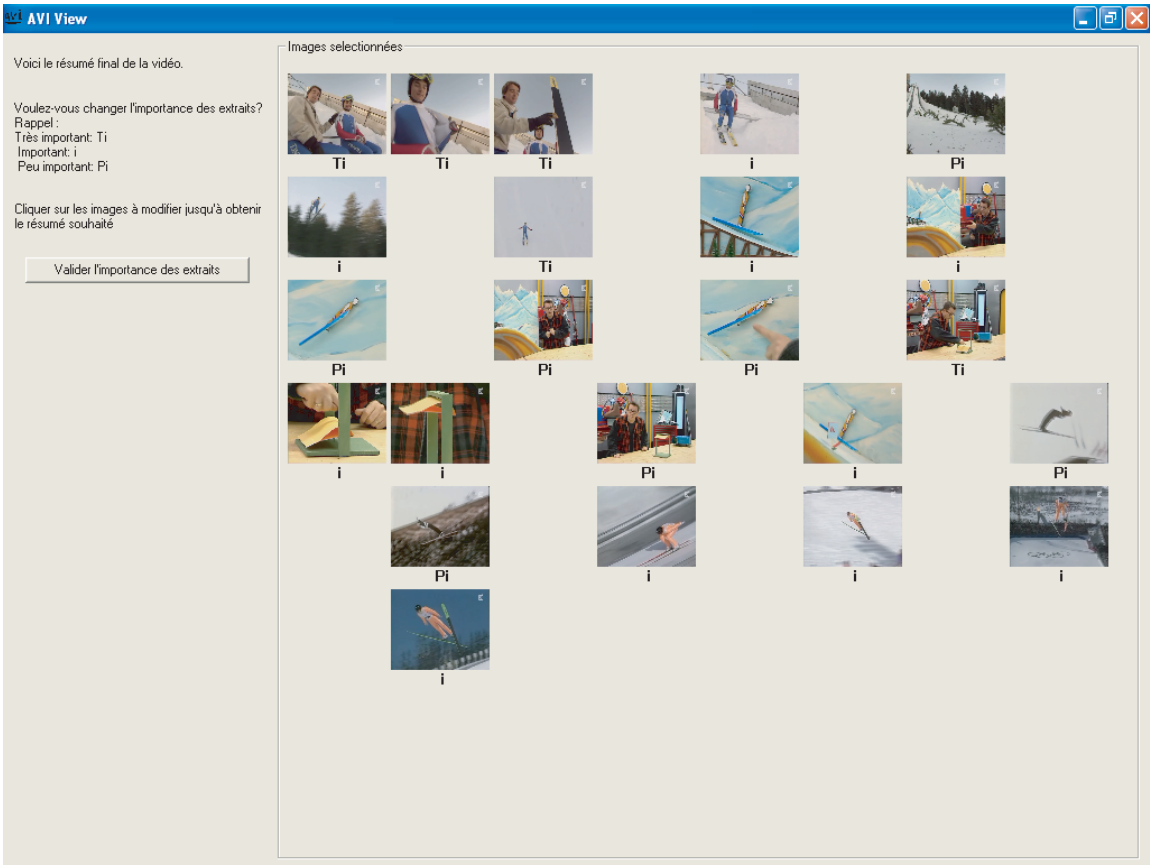


FIG. 5.13 – Troisième étape de la création du résumé de référence pour la vidéo « Documentaire ». Le résumé complet est affiché et correspond à la sélection d'un sujet lors de l'annotation des extraits (étape 2). Il est alors demandé au sujet de modifier si besoin le degré d'importance des extraits.

5.3.3 Construction d'un résumé de référence

La difficulté consiste à créer un résumé de référence à partir des résumés créés par les différents sujets. En partant de l'hypothèse que les résumés des sujets ont une véritable signification sémantique, il s'agit de trouver un résumé « optimal » qui prend en compte ces différents résumés. Néanmoins, les écarts entre des résumés ne se mesurent pas en appliquant une distance entre les descripteurs des images puisque le fossé entre descripteurs de bas niveau et contenu sémantique n'est pas encore comblé. Le procédé va s'appuyer sur des considérations élémentaires pour créer le résumé optimal. Nous développons deux méthodes pour créer un résumé de référence, l'une conçue au niveau de chaque plan appelée « résumé fin » et l'autre créée à partir de comparaison entre plans appelée « résumé court ».

5.3.3.1 Au niveau des plans

La construction au niveau des plans s'effectue seulement à partir des annotations de l'étape 2, c'est-à-dire des images sélectionnées pour représenter les différents extraits. Comme déjà mentionné précédemment, chaque extrait visionné correspond à un plan, et seules les images choisies par les sujets seront examinées et non les degrés d'importance des plans les uns par rapport aux autres. Comme le nombre possible d'images sélectionnées varie d'un sujet à l'autre, le nombre optimal d'images clés doit être d'abord déterminé pour représenter un extrait. La moyenne arithmétique pourrait être utilisée pour déterminer le nombre optimal. Néanmoins, comme la moyenne est influencée par les données atypiques, le médian est privilégié en raison de sa robustesse.

Une fois le nombre d'images clés trouvé, il faut ensuite déterminer comment les images choisies par les différents sujets sont réparties sur un plan donné. Néanmoins la répartition temporelle des images ne suffit pas puisqu'elle ne permet pas de prendre en compte le voisinage temporel des images. Comme les images sont échantillonnées une toutes les dix, deux images voisines peuvent être sélectionnées par différents sujets et avoir le même contenu. De plus, il faut aussi différencier les sujets qui ont sélectionné peu d'images de ceux qui en ont sélectionné beaucoup. En fonction du nombre d'images choisies par un sujet pour un plan donné, un poids est attribué à chacune de ces images. Si une seule image est sélectionnée pour un plan donné, le poids associé à l'image vaut trois alors que si trois images sont choisies, le poids n'est plus que de un par image. Cette stratégie assure un poids moyen par plan égal pour chaque sujet. Ceci reste cohérent avec le fait que si un sujet choisit beaucoup d'images, elles auront un poids faible et inversement.

Afin de prendre en compte le voisinage de l'image sélectionnée, une gaussienne centrée à l'endroit de l'image et d'écart type σ est positionnée selon un axe temporel. L'amplitude de la gaussienne est fonction du poids attribué précédemment. Si le sujet a choisi par exemple une seule image pour représenter le plan, alors une seule gaussienne sera placée sur l'axe temporel avec une amplitude de trois. L'écart type est un paramètre important pour la création du résumé de référence. Plus ce paramètre sera grand, plus les images sélectionnées par les différents sujets seront fusionnées. La figure 5.14 montre comment le poids des images voisines varie en fonction du paramètre σ. Comme les images à choisir sont affichées selon un échantillonnage régulier (une toutes les dix), le poids de l'image voisine dépend directement de ce paramètre et se situe à l'indice 10. Par exemple, si $\sigma = 20$ alors le poids de l'image voisine vaut 0.88.

Après accumulation des réponses des différents sujets, nous obtenons la répartition temporelle des images sélectionnées. La figure 5.15 présente les résultats sur la séquence « Docu-

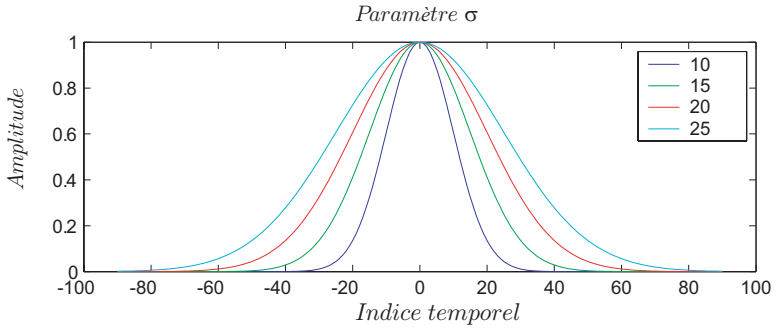


FIG. 5.14 – Paramètre σ en fonction de l'image choisie par le sujet. La gaussienne est positionnée à l'endroit où une image a été choisie. Par exemple, si le paramètre $\sigma = 10$ alors l'image voisine (à gauche ou à droite) aura un poids de 0.6 et l'image suivante aura un poids de 0.13 puisque les images sont affichées selon un échantillonnage régulier (une toutes les dix). En revanche, si le paramètre $\sigma = 20$, le poids de l'image voisine aura un poids de 0.88 et l'image suivante aura un poids de 0.6.

mentaire ». Nous pouvons constater par exemple que le premier plan est très long et possède beaucoup de maximums locaux alors que le deuxième plan n'a qu'un seul maximum. Les maximums symbolisent les endroits où les images doivent être sélectionnées pour résumer la vidéo puisque ce sont les endroits les plus choisis par les sujets. Après obtention des maximums en calculant la dérivée première et en recherchant les changements de signe, ils sont triés par ordre décroissant. Les maximums locaux voisins sont fusionnés pour éviter la présence de maximums locaux sur une fenêtre inférieure à 2 secondes (ou 50 images). De plus, tous les maximums locaux dont l'amplitude est inférieure à 20% du maximum global sont supprimés.

Finalement, pour chaque plan, nous retenons seulement les n premiers maximums locaux triés par ordre décroissant en fonction du nombre optimal d'images défini. Ils correspondent aux endroits où les images clés sont choisies pour résumer le plan et donc la vidéo. La figure 5.16 montre le résumé de référence de la vidéo « Documentaire ». Le paramètre σ est fixé à 20. La redondance entre certaines images du résumé peut paraître surprenante. Néanmoins, elle est logique du fait que les images appartiennent à des plans différents. Par exemple, les images n°1944 et n°2176 sont similaires mais ne proviennent pas des mêmes plans. Le choix du paramètre σ sera étudié lors de la présentation des résultats sur la méthode de résumé de vidéo à partir du mouvement de caméra.

5.3.3.2 Au niveau de la vidéo

La construction d'un résumé de référence au niveau de la vidéo est difficile à obtenir. En effet, la performance du résumé proposé doit dépendre du nombre de plans présents dans le résumé. Il s'agit de tenir compte de la taille du résumé proposé puisqu'il n'existe pas de résumé idéal avec une taille parfaite.

L'idée est de déterminer un classement des plans de la vidéo suivant leur importance. Puis, en fonction des plans représentés dans le résumé proposé, une mesure de performance pourra alors être établie. A partir des annotations des vidéos par les différents sujets, l'importance des plans est calculée. Un poids est attribué (respectivement 1, 0.5, 0) pour chaque plan suivant le degré d'importance donné à celui-ci par un sujet (respectivement très important,

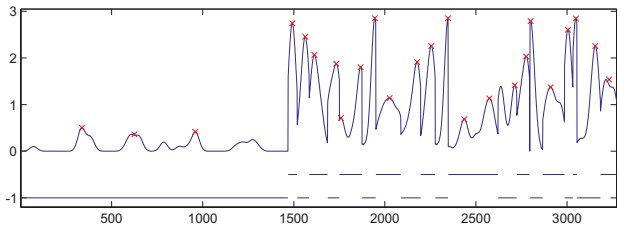


FIG. 5.15 – Répartition de la sélection des images sur la vidéo « Documentaire » normalisé par le nombre de sujet (Axe horizontal correspond aux numéros d'images). Les maximums sur cette courbe permettent la sélection des images clés. Les croix sur la courbe sont les images choisies pour résumer la vidéo. La fonction en escalier entre -1 et 0 permet de repérer les changements de plan. Dans cet exemple, le paramètre σ est fixé à 20.

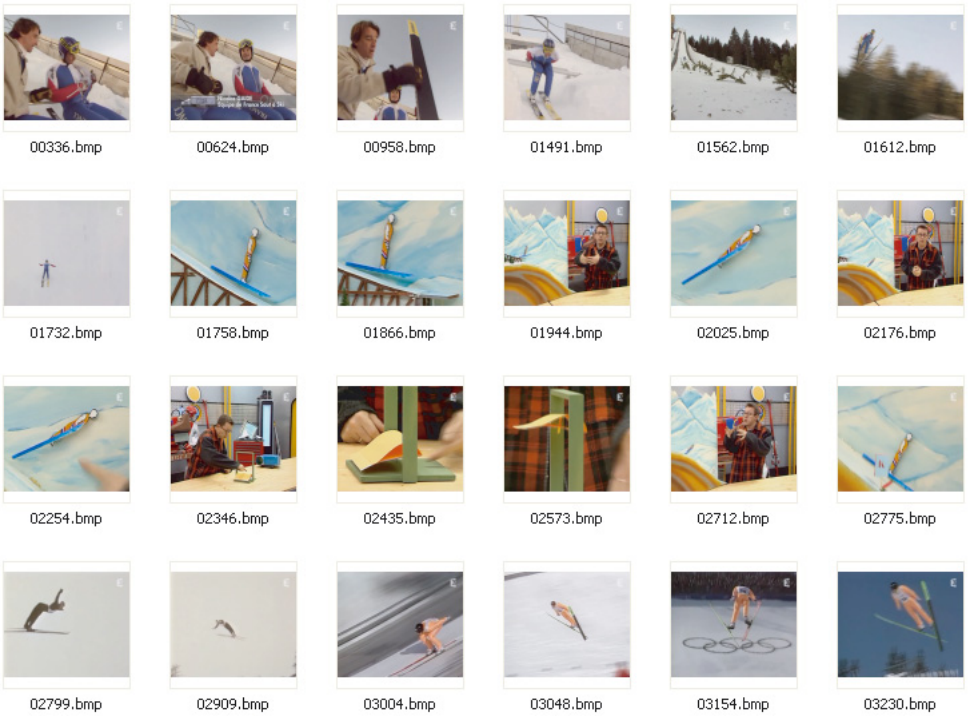
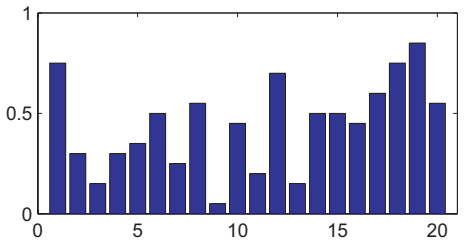


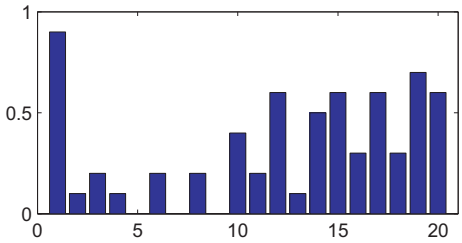
FIG. 5.16 – Exemple de résumé de référence pour les plans de la vidéo « Documentaire » avec le paramètre σ fixé à 20. Trois plans possèdent plus d'une image, le plan n°1 avec les trois premières images présentées, le plan n°6 avec les images 1758 et 1866 et enfin le plan n°12 avec les images 2435 et 2573.

important, peu important). Par exemple, si le plan est très important pour un sujet, le plan obtient le poids de un. Finalement, la moyenne de ces poids est calculée sur chaque plan. La figure 5.17.a montre la moyenne des degrés d'importance pour la vidéo « Documentaire ». Nous pouvons constater que les moyennes sont très différentes entre les plans. De plus, bien que les stratégies d'annotation peuvent être différentes selon les sujets, l'effet sur la moyenne des degrés d'importance est faible. Prenons l'exemple de 2 sujets où l'un annote tous les plans comme étant importants et l'autre avec un degré qui varie entre très important et peu important. Si les plans les plus importants sont recherchés, les poids fournis par le premier sujet n'apportent pas d'informations sur la sélection des plans. Ce n'est pas la moyenne absolue des poids qui importe mais les écarts entre les degrés d'importance par un même sujet qui déterminent les plans importants.

De plus, à la fin de chaque annotation par un sujet, un résumé leur est demandé avec un nombre d'images clés imposé. Le nombre demandé permet de fournir un résumé court. Dans notre expérimentation, le nombre est fixé au nombre de plans contenus dans la vidéo divisé par deux. Par exemple, la vidéo « Documentaire » a 20 plans et plus de 3000 images alors le résumé court devra posséder au maximum 10 images. Ainsi cette sélection permet aussi de connaître les plans les plus importants dans la vidéo. Afin d'en tenir compte, chaque plan qui se trouve dans le résumé court d'un sujet est extrait, puis un histogramme est obtenu sur les différents sujets. La figure 5.17.b présente l'histogramme. On peut s'apercevoir que certains plans ne sont jamais sélectionnés par les sujets.



(a) Degré d'importance des plans.



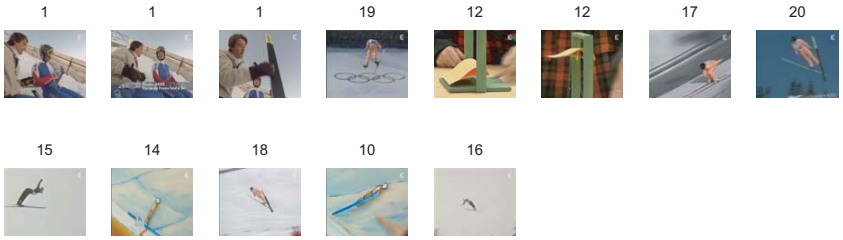
(b) Proportion des plans sélectionnés pour le résumé court.

Fig. 5.17 – Exemple des annotations par les sujets pour la vidéo « Documentaire ». (a) Moyenne des degrés d'importance par plan et (b) sélection moyenne des plans pour le résumé court.

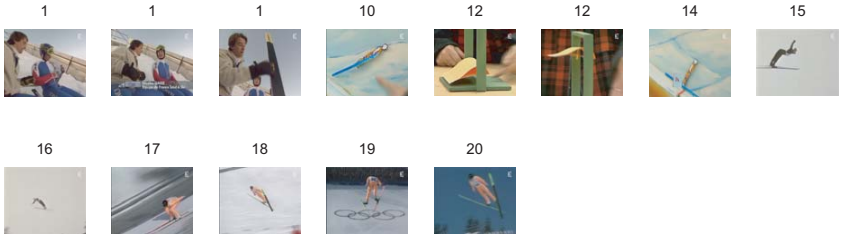
Comme les plans sélectionnés dans le résumé court font partie des plans les plus importants, ils doivent apporter une information sur le degré d'importance des plans. La moyenne des degrés d'importance pour un plan donné est donc pondérée par la sélection moyenne du plan correspondant pour connaître l'importance du plan. Un tri par ordre décroissant des importances des plans permet la création d'une hiérarchie sur le résumé de vidéo. Si seulement deux plans sont demandés pour résumer la vidéo, les deux premiers devront être retournés. La figure 5.18(a) montre les 10 premiers plans triés par ordre décroissant selon leur importance alors que la figure 5.18(b) montre le résumé présenté dans l'ordre chronologique.

5.3.4 Comparaison du résumé automatique avec le résumé de référence

La comparaison entre le résumé de référence et le résumé proposé par un algorithme appelé résumé candidat est une tâche délicate puisqu'elle nécessite la comparaison d'images, le



(a) Ordre décroissant des 10 premiers plans selon leur importance.



(b) Résumé de la vidéo avec les 10 premiers plans, triés par ordre chronologique.

FIG. 5.18 – Exemple de résumé de la vidéo « Documentaire ». Les numéros au dessus de chaque image correspondent aux numéros des plans et les images présentées sont celles qui résument chaque plan.

regroupement d'images et la comparaison entre plans. Deux approches sont proposées suivant le niveau de résolution demandé pour le résumé : sélection des images au niveau des plans et sélection des images au niveau de la vidéo.

5.3.4.1 Au niveau des plans

Le procédé de comparaison entre le résumé de référence et le résumé candidat s'effectue en 4 étapes. La figure 5.19 illustre la comparaison des résumés au niveau de chaque plan. Nous pouvons constater dans cet exemple que le résumé de référence a 3 images clés alors que le résumé candidat en possède 4.

La première étape consiste à déterminer les images du résumé de référence auxquelles chaque image du résumé candidat pourra être associée. Ainsi chaque image candidate est associée à deux images au maximum du résumé de référence, qui sont les images les plus proches temporellement sur la gauche et sur la droite dans le même plan. Par exemple, l'image B du résumé candidat est associée aux images 1 et 2 du résumé de référence (Fig. 5.19.a). En revanche, l'image D est associée uniquement à l'image 3.

La deuxième étape consiste à déterminer parmi les deux images potentielles du résumé de référence celle qui est la plus similaire à l'image du résumé candidat. Par exemple, l'image B qui peut être associée soit à l'image 1 ou 2 est finalement associée à l'image 1 (Fig. 5.19.b) car elle est supposée plus proche en terme de contenu. Cela demande la représentation des images par un descripteur et la définition d'une distance entre deux images. Néanmoins, il est difficile de comparer le contenu de deux images. Mais, comme les images appartiennent au même plan, cela signifie une continuité temporelle entre les images et la comparaison entre les images peut être effectuée en comparant leurs histogrammes couleur. En effet, deux histogrammes supposés similaires auront le même contenu puisque les images sont continues

temporellement. Il est donc assez improbable d'avoir deux histogrammes similaires avec des contenus différents à l'intérieur d'un même plan. Le descripteur utilisé ici est un histogramme couleur global et la distance entre histogrammes est obtenue par la norme L1. Nous avons choisi de ne pas présenter l'histogramme couleur, celui-ci n'étant pas indispensable pour comprendre la méthode. Cependant, nous l'avons également employé par la suite et une description détaillée pourra être trouvée à la section 6.2.1.1.

La troisième étape traite du cas où plusieurs images du résumé candidat sont associées à la même image du résumé de référence. Par l'exemple, les images A et B sont associées à la même image 1 (Fig. 5.19.b) et finalement seule l'image B est associée à l'image 1 (Fig. 5.19.c) puisque la distance entre les images 1 et B est supposée plus faible.

Enfin, la quatrième étape consiste à ne conserver que les regroupements où les distances sont inférieures à un seuil δ_s. Les images qui ont été regroupées jusqu'à présent peuvent avoir des distances d'histogrammes élevées. Le seuillage permet de conserver uniquement les images regroupées avec un contenu similaire. Le paramètre δ_s est fondamental et sera largement étudié lors de la présentation des résultats.

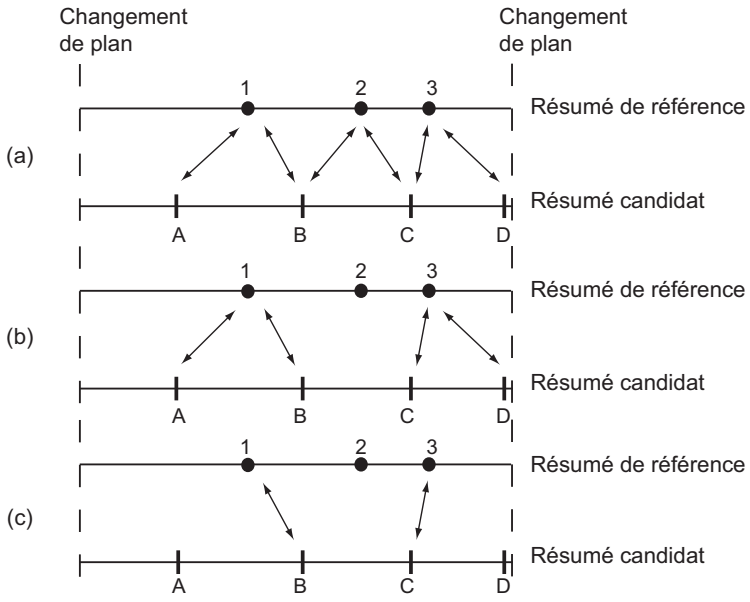


FIG. 5.19 – Illustration de la comparaison au niveau d'un plan entre le résumé de référence et le résumé candidat. Le résumé de référence possède 3 images (de 1 à 3) alors que le résumé candidat présente 4 images (de A à D).

La comparaison entre le résumé de référence et le résumé candidat permet finalement d'obtenir le nombre d'images regroupées. Les mesures standards que sont la précision, le rappel et le F_1 peuvent être utilisées pour évaluer le résumé candidat.

5.3.4.2 Au niveau de la vidéo

La comparaison entre différents résumés de vidéo va essentiellement dépendre des plans proposés et non des images soumises. L'évaluation doit prendre en compte le nombre de plans auxquels les images clés soumises appartiennent afin de pouvoir traiter les méthodes de résumé de vidéo avec plusieurs niveaux de résolution. Une méthode simple pour mesurer la qualité

du résumé est de comparer les plans proposés aux plans triés de la vérité terrain (§ 5.3.3.2). L'idée consiste à comparer les n plans soumis par la méthode avec les n premiers plans triés. En ce qui concerne les images clés soumises, elles peuvent être évaluées avec la méthode de comparaison au niveau des plans.

Pour un nombre donné de plans, le calcul du rappel va donner un critère de performance sur la méthode. Pour pouvoir comparer des méthodes de résumé de vidéo, il suffit de fournir le nombre de plans à retourner et de comparer le rappel entre les méthodes.

La figure 5.20 illustre l'évaluation de résumés candidats lorsque quatre plans sont demandés pour résumer la vidéo. Par exemple, la méthode de résumé n°1 fournit 5 images qui appartiennent à 4 plans (1, 2, 4 et 7). La méthode de comparaison au niveau des plans permet de juger du nombre d'images fournies pour représenter chaque plan. Ainsi ce sont seulement les plans retournés qui sont évalués et non le nombre d'images retournées pour chaque plan. C'est pourquoi la méthode de résumé n°1, qui possède deux images pour représenter le plan n°4, a un rappel de 3/4. La deuxième méthode a en revanche un rappel de 2/4.

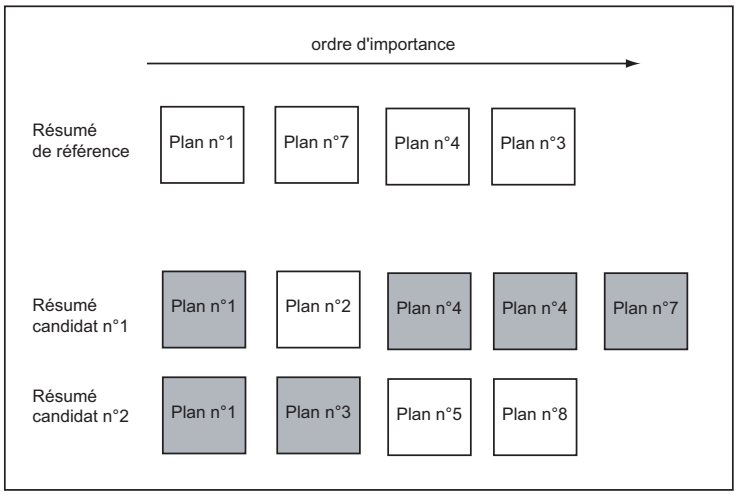


FIG. 5.20 – Illustration de la comparaison au niveau d'une vidéo entre le résumé de référence et deux résumés candidats. Les deux résumés candidats présentent 4 plans pour résumer les vidéos. Les images grisées du résumé candidat correspondent aux plans qui ont été retrouvés dans le résumé de référence. Le résumé n°1 a un rappel de 3/4 alors que le résumé n°2 a un rappel de 2/4.

5.3.5 Évaluation de résumés automatiques

Comme la méthode de création de résumé à partir du mouvement de caméra fournit un résumé au niveau des plans, nous allons étudier uniquement la méthode d'évaluation au niveau des plans. Six résumés candidats sont étudiés : trois sont des résumés élémentaires qui permettent de fixer les performances minimales à atteindre et trois sont des résumés provenant du mouvement de caméra. Le premier résumé est un résumé complètement aléatoire. Pour chacun des plans de la vidéo, un nombre d'images à sélectionner est tiré aléatoirement (nombre entre 1 et 3), puis les images sont tirées aléatoirement dans le plan. Le deuxième résumé est une version semi-aléatoire. Il consiste à tirer aléatoirement les images dans le plan, mais le nombre d'images à tirer est donné par le résumé de référence. Le troisième

résumé est créé en choisissant l'image du milieu de chaque plan. Enfin, les trois derniers résumés sont ceux qui ont été développés à partir du mouvement de caméra : résumé suivant l'amplitude des mouvements, résumé suivant l'enchaînement des mouvements, puis résumé suivant l'enchaînement et l'amplitude des mouvements (Section 5.2).

Le tableau 5.4 montre les résultats des 6 méthodes de résumé de vidéo. Nous pouvons constater que la méthode suivant l'amplitude et l'enchaînement des mouvements obtient les meilleurs résultats pour les trois vidéos. Pour la vidéo « Série », les méthodes n°2 et n°3 présentent des résultats qui sont du même ordre de grandeur que celle suivant l'amplitude et l'enchaînement des mouvements. Ceci confirme que les méthodes qui sélectionnent une seule image par plan (soit une image au milieu du plan ou tirée de façon aléatoire) sont relativement efficaces quand les plans sont de courte durée. La vidéo « Série » comprend 16 plans sur 28 de moins de 3 secondes alors que les vidéos « Documentaire » et « Journal » ont respectivement 8 plans sur 20 et 9 plans sur 42. Il est effectivement naturel de sélectionner une seule image pour ces plans. De plus, devant la similarité des images, l'image clé peut être sélectionnée de manière aléatoire. Néanmoins, il est préférable de la sélectionner à proximité du centre pour éviter d'avoir des images de mauvaise qualité (les images aux extrémités du plan peuvent être bruitées lors d'un changement de plan). Cependant, les résultats présentés confirment l'intérêt d'utiliser le mouvement de caméra pour sélectionner des images. Plus les plans seront de taille conséquente, plus le contenu sera susceptible de changer et donc la méthode sera d'autant plus efficace.

TAB. 5.4 – Résultats des six méthodes de résumé pour les trois vidéos. Le seuil δ_s de regroupement entre deux images est fixé à 0.3. (R : Rappel, P :Précision, F_1)

	Documentaire			Journal			Série		
Résumé	R	P	F_1	R	P	F_1	R	P	F_1
n°1	62.5 (15/24)	40.5 (15/37)	49.1	83.6 (46/55)	50.5 (46/91)	63.0	80.0 (24/30)	40.6 (24/59)	53.9
n°2	54.1 (13/24)	54.1 (13/24)	54.1	72.7 (40/55)	72.7 (40/55)	72.7	76.6 (23/30)	76.6 (23/30)	76.6
n°3	50.0 (12/24)	60.0 (12/20)	54.5	63.6 (35/55)	83.3 (35/42)	72.1	73.3 (22/30)	78.5 (22/28)	75.8
n°4	83.3 (20/24)	35.7 (20/56)	50.0	81.8 (45/55)	56.9 (45/79)	67.1	80.0 (24/30)	50.0 (24/48)	61.5
n°5	75.0 (18/24)	45.0 (18/40)	56.2	76.3 (42/55)	63.6 (42/66)	69.4	76.6 (23/30)	60.5 (23/38)	67.6
n°6	79.1 (19/24)	55.8 (19/34)	**65.5**	80.0 (44/55)	77.1 (44/57)	**78.5**	86.6 (26/30)	72.2 (26/36)	**78.7**

Cependant, la méthode de comparaison des résumés nécessite de fixer différents paramètres qui peuvent influencer les résultats. Dans la méthode de construction de résumé de référence, le paramètre à étudier est l'écart type de la gaussienne σ autour de l'image choisie par un sujet. Effectivement, si le paramètre σ est choisi petit, alors les images voisines sélectionnées par les sujets ne pourront être fusionnées. De même, si le paramètre σ est choisi grand alors les images seront facilement regroupées. Donc le nombre de maximums locaux à l'intérieur d'un plan dépend de ce paramètre σ. Les figures 5.21, 5.22 et 5.23 illustrent les résultats de la méthode de résumé avec la sélection de l'image au centre du plan et la méthode suivant l'amplitude et l'enchaînement des mouvements en fonction du paramètre σ. Nous pouvons aussi remarquer que le nombre d'images du résumé de référence pour les trois vidéos ne diminue pas de manière importante avec l'augmentation du paramètre σ. De plus, les résultats des deux méthodes présentées restent relativement stables en fonction du paramètre σ. Ainsi, nous pouvons conclure que ce paramètre σ ne remet pas en cause les performances des méthodes. Par la suite, ce paramètre σ sera fixé à 20.

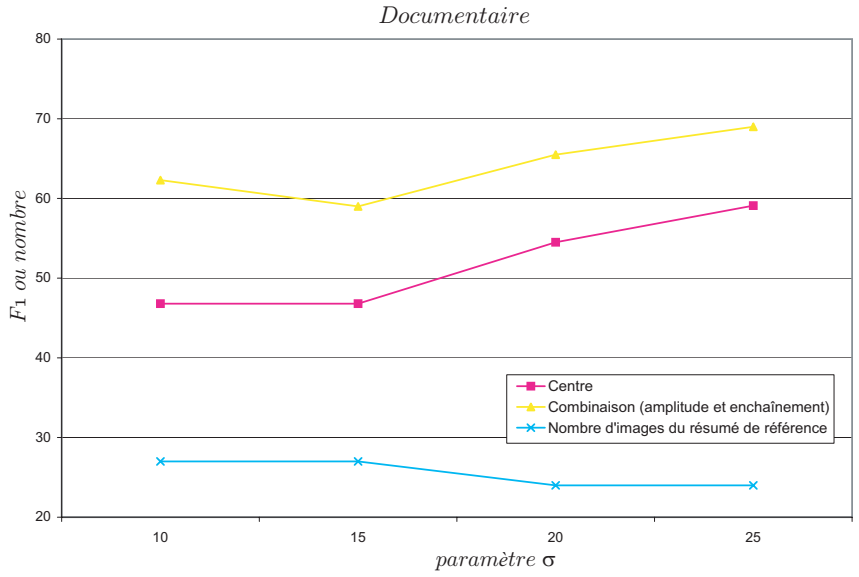


FIG. 5.21 – Etude de la méthode de résumé avec sélection de l'image au centre du plan et de la méthode suivant l'amplitude et l'enchaînement des mouvements en fonction du paramètre σ pour la vidéo « Documentaire ». Le seuil δ_s est fixé à 0.3.

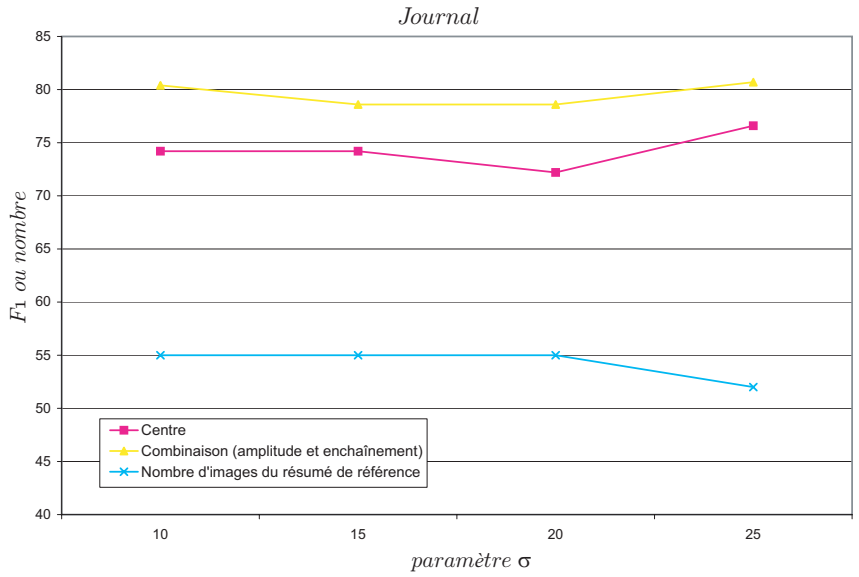


FIG. 5.22 – Etude de la méthode de résumé avec sélection de l'image au centre du plan et de la méthode suivant l'amplitude et l'enchaînement des mouvements en fonction du paramètre σ pour la vidéo « Journal ». Le seuil δ_s est fixé à 0.3.

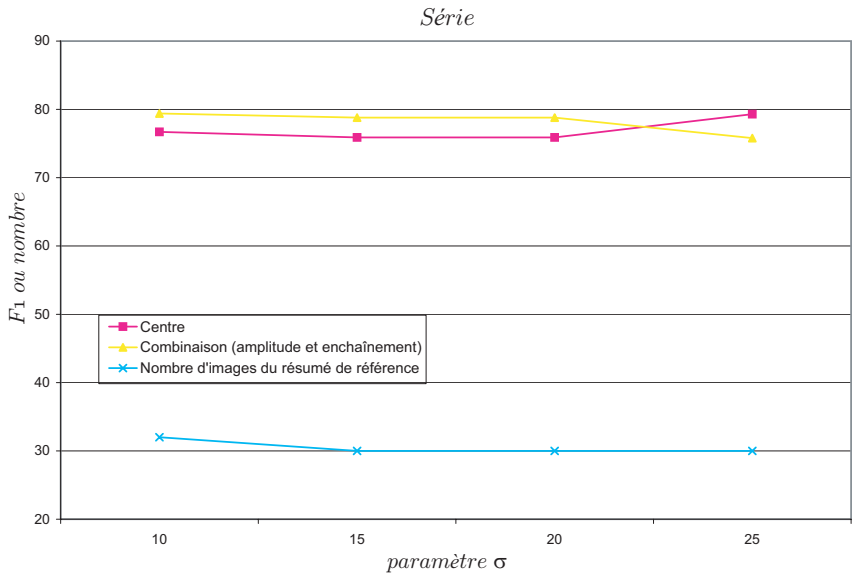


FIG. 5.23 – Etude de la méthode de résumé avec sélection de l'image au centre du plan et la méthode suivant l'amplitude et l'enchaînement des mouvements en fonction du paramètre σ pour la Vidéo « Série ». Le seuil δ_s est fixé à 0.3.

Enfin, en ce qui concerne la comparaison entre résumé de référence et résumé candidat, bien que la description des images s'effectue par l'histogramme couleur, les regroupements entre images sont conservés uniquement si les distances sont inférieures au seuil δ_s. Or, ce seuil joue un rôle important dans les résultats. En effet, si le seuil est choisi assez petit alors les images seront difficilement regroupées alors que si le seuil est trop grand, des images non similaires pourront être réunies. Les figures 5.24, 5.25 et 5.26 illustrent les résultats des différentes méthodes en fonction du seuil δ_s. Comme attendu, plus le seuil augmente, plus les performances augmentent (jusqu'à une certaine valeur). Néanmoins, quel que soit le seuil sélectionné, la méthode suivant l'amplitude et l'enchaînement des mouvements présente les meilleurs résultats pour les vidéos « Documentaire » et « Journal ». En ce qui concerne la vidéo « Série », la méthode la plus performante est aussi celle qui est basée sur l'amplitude et l'enchaînement des mouvements pour les seuils 0.1, 0.2, 0.3 et 0.4. En revanche, pour les seuils 0.5 et 0.6, c'est la méthode de résumé avec l'image au centre du plan qui est la plus performante. De manière générale, les performances obtenues pour les seuils 0.5 et 0.6 sont sensiblement similaires pour une même vidéo (présence de plat sur les courbes). Cela signifie que le paramètre δ_s est trop élevé et que des images dissemblables peuvent être regroupées. Le paramètre δ_s doit plutôt être choisi en dessous de 0.5 car la pente est non nulle.

5.4 Conclusion

Nous avons présenté une méthode originale de création de résumé de vidéo à partir du mouvement de caméra. Elle consiste à sélectionner des images suivant l'amplitude et l'enchaînement des mouvements de caméra. Cette méthode repose sur un système à base de règles. L'élaboration de ces règles permet de sélectionner des images clés en fonction des mouvements de caméra. Les règles que nous avons utilisées sont naturelles et ont pour objectif d'éviter la

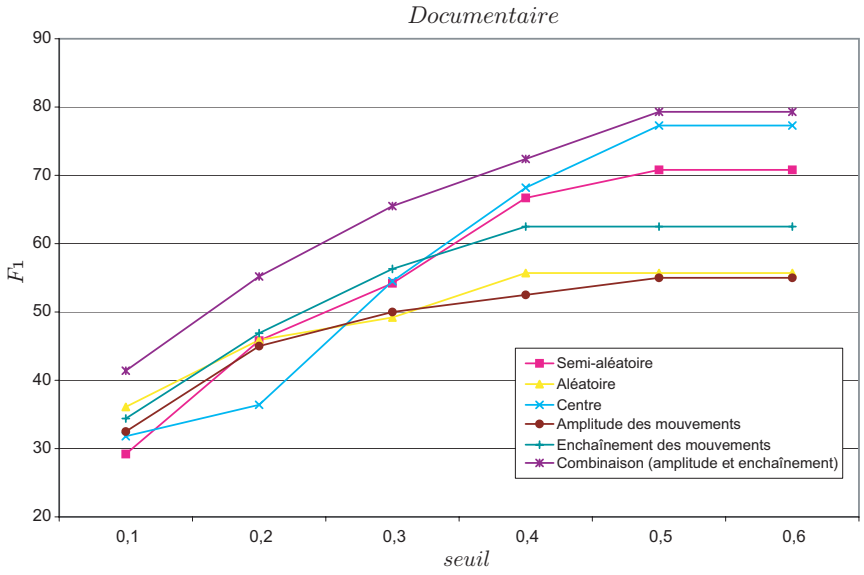


FIG. 5.24 – Etude du regroupement des images en fonction du seuil δ_s pour la vidéo « Documentaire ». La méthode de résumé suivant l'amplitude et l'enchaînement des mouvements fournit les meilleurs résultats quel que soit le seuil choisi.

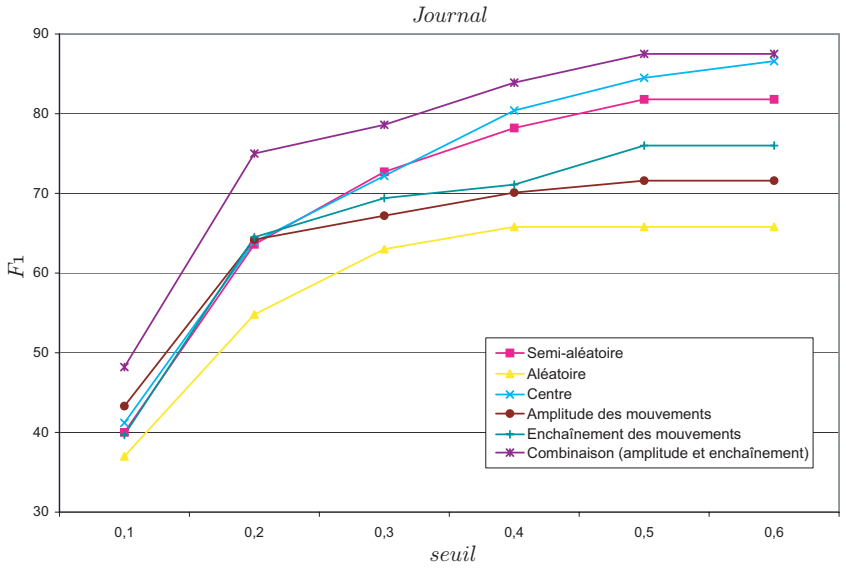


FIG. 5.25 – Etude du regroupement des images en fonction du seuil δ_s pour la vidéo « Journal ». La méthode de résumé suivant l'amplitude et l'enchaînement des mouvements fournit les meilleurs résultats quel que soit le seuil choisi.

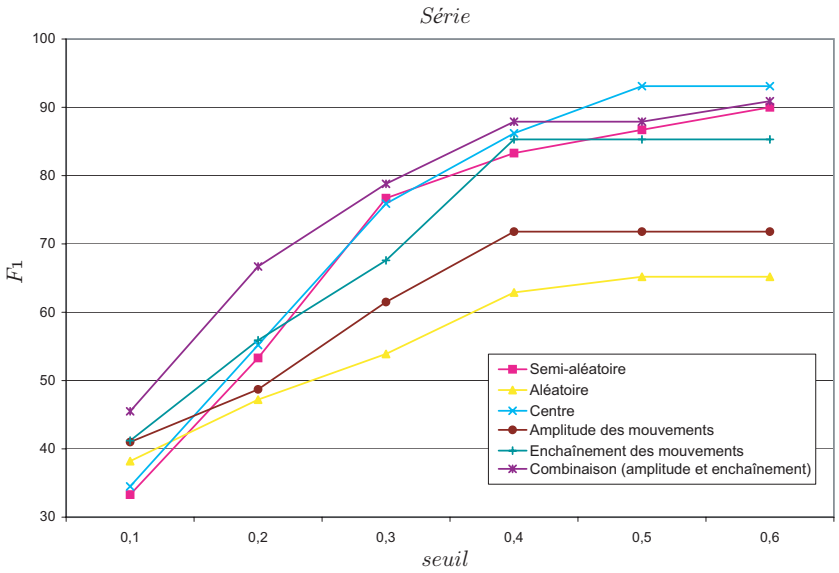


FIG. 5.26 – Etude du regroupement des images en fonction du seuil δ_s pour la vidéo « Série ». La méthode de résumé suivant l'amplitude et l'enchaînement des mouvements fournit les meilleurs résultats pour un seuil inférieur ou égal à 0.4. En revanche, pour un seuil de 0.5 ou 0.6, la troisième méthode (résumé avec l'image au centre du plan) est la plus performante.

redondance temporelle entre les images. Par exemple, si un segment statique est suivi d'un segment avec un mouvement de translation, les règles conçues permettent de désigner comme images clés l'image au milieu du segment statique et l'image de fin de la translation. Cette méthode permet d'introduire des informations de plus haut niveau pour construire le résumé. Elle présente aussi l'avantage de ne pas calculer des différences entre images à partir d'informations de bas niveau.

Une méthode originale d'évaluation a également été proposée pour comparer les différents résumés créés. Une expérience a été mise en place pour permettre à un sujet de créer manuellement un résumé pour une vidéo donnée. Une fois les résumés obtenus selon plusieurs sujets, nous avons conçu un résumé de référence au niveau des plans appelé « résumé fin » et un résumé de référence hiérarchique au niveau de la vidéo appelé « résumé court ». Le résumé de référence au niveau des plans a ensuite été comparé à différentes méthodes de résumé. Une étude a également été menée afin d'étudier l'influence des paramètres de la méthode de comparaison et les principaux effets sur les performances des résumés à tester. Parmi les méthodes de résumé proposées (résumé aléatoire, résumé semi-aléatoire, résumé en sélectionnant l'image au centre de chaque plan, résumé fonction de l'amplitude des mouvements de caméra, résumé fonction de l'enchaînement des mouvements de caméra et résumé fonction de l'amplitude et de l'enchaînement des mouvements de caméra), la méthode de création de résumé suivant l'amplitude et l'enchaînement des mouvements est celle qui a fourni les meilleurs résultats.

Le résumé hiérarchique de référence au niveau de la vidéo (« résumé court ») a aussi été proposé mais n'a pas été comparé. En effet, les méthodes de résumé présentées dans ce chapitre ont été conçues au niveau des plans. Il serait alors intéressant d'appliquer cette méthode d'évaluation à des résumés hiérarchiques et plus particulièrement aux résumés hiérarchiques fournis par la première méthode de résumé du chapitre 3.

Chapitre 6

Détection des changements de plan

La méthode de création de résumé décrite dans le chapitre 5 a été conçue à partir du mouvement de caméra. Néanmoins, leur extraction nécessite la connaissance des différents plans de la vidéo. Pour fabriquer un résumé de vidéo de manière automatique, nous proposons dans ce chapitre d'étudier et de détecter les changements de plan dans les vidéos. Ainsi la chaîne complète de la méthode de résumé de vidéo à partir du mouvement de caméra aura été traitée.

Sommaire

6.1 Introduction

Dans le chapitre 4, une méthode a été mise au point pour identifier les mouvements de caméra dans les vidéos. Elle est à l'origine d'une méthode de résumé, développée dans le chapitre 5, qui repose sur cette identification. Or, la méthode de classification des mouvements de caméra suppose la connaissance des changements de plan. Pour obtenir une chaîne complète, nous allons aborder dans ce chapitre une méthode de détection des changements de plan.

De façon générale, une vidéo comporte différents types de transitions qui permettent de passer d'un plan à l'autre. Par définition, une transition correspond au point de jonction entre deux plans. Il existe plusieurs types de transitions dans les vidéos. Celles-ci ont été regroupées suivant deux grandes familles de transitions :

Transition instantanée :
Cette famille de transitions (Fig. 6.1(a)) consiste à juxtaposer la fin d'un plan avec le début du plan suivant. Ces changements de plan sont les plus répandus dans les vidéos.

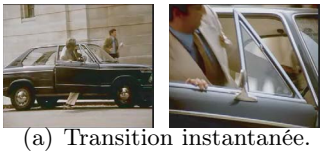
Transition progressive :
Les transitions progressives regroupent plusieurs types de transitions et elles correspondent à un changement progressif lors du passage d'un plan à un autre. Nous pouvons tout d'abord dégager les fondus (Fig. 6.1(b)). Ils consistent à faire apparaître (ou disparaître) progressivement un plan à partir d'une image d'intensité constante. En général, l'image d'intensité constante est une image noire, mais celle-ci peut être une image différente comme une image blanche. Un fondu est appelé soit fondu en ouverture si le plan apparaît soit fondu en fermeture si le plan disparaît. Un autre type de transition progressive souvent rencontré est le fondu enchaîné (Fig. 6.1(c)). Il est conçu par superposition des extrémités de deux plans consécutifs, en faisant disparaître le premier plan et apparaître le second graduellement. Enfin, le dernier type de transition progressive est le volet (Fig. 6.1(d)). Il correspond à une transition qui remplace le plan précédent par le suivant selon une ligne verticale traversant l'image. D'autres versions du volet peuvent être construites suivant la direction de la ligne qui traverse l'image. Parmi les différents types de transitions, les fondus et les fondus enchaînés sont les plus souvent rencontrés dans les vidéos. Néanmoins d'autres types de transitions peuvent être créés suivant l'imagination de son concepteur. Par exemple, une transition en damier pourrait également être réalisée entre deux plans.

Par la suite, nous avons choisi de nous intéresser à ces deux familles de transitions. En ce qui concerne les transitions progressives, nous avons étudié plus particulièrement les fondus et les fondus enchaînés, qui sont les plus fréquents.

Beaucoup de techniques ont été développées pour détecter les différents types de transitions. La plupart des méthodes utilisent des différences d'images basées sur des comparaisons de luminances ou d'histogrammes. A partir de caractéristiques visuelles (couleur, contour ou mouvement) extraites sur chaque image, une distance est définie soit entre deux images successives soit entre images contenues dans une fenêtre temporelle pour déterminer des différences entre elles. De fortes variations sur les différences renseignent sur la présence d'une discontinuité temporelle et donc d'une possible transition. Celle-ci peut être ainsi détectée par seuillage ou apprentissage sur les différences.

Suivant la nature de la transition, instantanée ou progressive, différentes approches ont été proposées. Pour les transitions instantanées, la méthode la plus simple pour les détecter consiste à comparer les pixels entre images successives [You03]. Cependant cette comparaison n'est pas insensible aux mouvements des objets et de la caméra, ce qui peut aboutir à de fausses détections. De ce fait, certaines approches développent des descripteurs moins sensibles aux mouvements comme les histogrammes [Qi03b, Zho05]. D'autres utilisent la compensation de mouvement [Por01] ou le suivi de caractéristiques [Whi03] pour créer une métrique entre les images. Enfin, des comparaisons entre histogrammes couleur locaux (par division de l'image en blocs) peuvent aussi être effectuées pour être moins sensibles aux mouvements des objets [Qi03b].

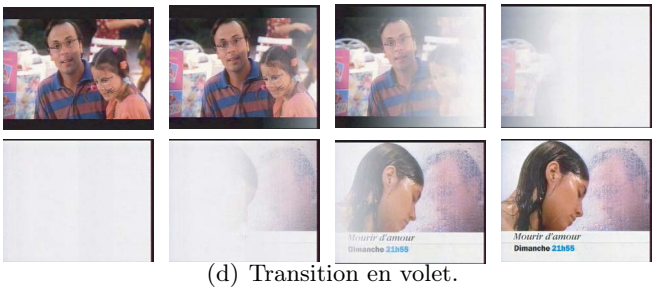
En ce qui concerne les transitions progressives, leur détection est plus difficile en raison de la faible différence entre deux images successives et le nombre inconnu d'images dans la transition. Pour être indépendantes de la longueur des transitions, les distances sont le plus



(a) Transition instantanée.

(b) Fondu en ouverture.

(c) Fondu enchaîné



(d) Transition en volet.

FIG. 6.1 – Exemple de transitions.

souvent calculées à l'intérieur d'une fenêtre temporelle. De plus, comme le mouvement de caméra entraîne souvent de fausses détections, certaines approches [Sae04, Ron04, Qi03b] l'estiment pour distinguer les changements progressifs (liés au mouvement de caméra) des transitions progressives. D'autres [Cai04] caractérisent les transitions à partir de la prédiction de blocs (utilisation du bloc matching).

Par ailleurs, une étape importante dans la détection est la détermination des seuils. La plupart des méthodes utilisent des seuils préfixés [Sae04, Por01, Cai04, Bes05]. D'autres approches définissent des seuils dynamiques dépendants du contenu des images [Ron04, Whi03, You03]. Pour éviter d'avoir recours à des seuils, certains [AlH03, Qi03b] choisissent un algorithme d'apprentissage pour discerner les transitions. Mais ce type d'approche nécessite une base d'apprentissage adéquate.

Enfin des méthodes proposent de détecter les plans en effectuant une analyse statistique. Dans [Zho05], une analyse en composantes indépendantes suivie d'un algorithme de regroupement par similarité des caractéristiques est réalisée pour déterminer les plans. Dans [Bes05], chaque image est caractérisée par un vecteur composé de plusieurs distances calculées à partir d'images d'une fenêtre temporelle suivant différents index (couleur et contour). Les seuils sont obtenus par une étude statistique qui consiste à projeter les vecteurs sur différents hyper-plans et à effectuer des sous-classifications successives.

D'autres travaux [AlH03, Sae04] se sont concentrés sur la détection de plans dans le domaine compressé MPEG. A partir des coefficients DC des images MPEG, Saez et al. [Sae04] proposent d'extraire deux descripteurs (luminance et contour) et utilisent un procédé de seuillage pour identifier les transitions. Le mouvement dominant où une transition progressive est détectée, est estimé pour éliminer les fausses alarmes. Dans [AlH03], un algorithme d'apprentissage à partir du pourcentage de macroblocks prédits dans le flux MPEG est utilisé pour détecter les transitions instantanées et éviter ainsi l'utilisation de seuils.

Néanmoins, peu de méthodes proposent une structure unifiée pour détecter à la fois les transitions instantanées et progressives. De plus, les méthodes demandent souvent l'utilisation de seuils et peu d'entre elles combinent différents descripteurs. Par conséquent, nous avons proposé une méthode qui se présente sous forme d'un arbre de décision défini à partir de règles. Le Modèle des Croyances Transférables a été employé pour faciliter la combinaison des descripteurs et détecter les transitions.

La suite du chapitre est organisée de la façon suivante. La section 6.2 présente la structure de la méthode de segmentation en plans de vidéos. Celle-ci nécessite l'extraction de plusieurs descripteurs, de la définition d'une similarité entre descripteurs et l'élaboration de règles pour détecter les différents types de transitions. Une évaluation est ensuite exposée dans la partie 6.3 pour juger de la performance de la méthode de détection des changements de plan.

6.2 Méthode de segmentation en plans de la vidéo

Nous proposons un système robuste qui se présente sous la forme d'un arbre de décision pour détecter à la fois les transitions instantanées et progressives. Il repose sur la combinaison de descripteurs locaux et globaux ainsi que sur l'estimation du mouvement. La description des images est réalisée par l'histogramme couleur (local et global) et l'estimation du mouvement dominant. L'originalité de notre détecteur réside dans la combinaison de ces descripteurs en utilisant le Modèle des Croyances Transférables afin d'être insensible aux changements liés aux mouvements de caméra et des objets. En effet, aucune transition ne doit être détectée en pré-

sence d'un mouvement dominant, supposé provenir du mouvement de la caméra, d'amplitude élevée. De même, l'extraction d'histogrammes couleur locaux suivie de comparaisons locales va permettre d'être insensible aux mouvements d'objets de grande taille. A partir de règles heuristiques, un *arbre de décision* est donc conçu pour détecter les transitions. La figure 6.2 présente la structure de la méthode. Pour chacune des étapes, une règle est appliquée. Les étapes 1 et 2 permettent une sélection exhaustive de transitions éventuelles, puis les étapes suivantes permettent un raffinement de la sélection des transitions éventuelles en appliquant différentes règles pour supprimer les fausses alarmes.

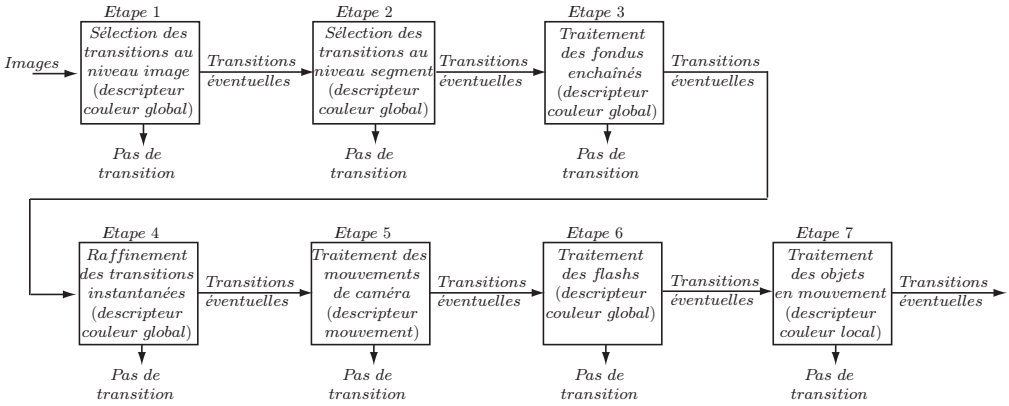


FIG. 6.2 – Principe de la détection des transitions.

6.2.1 Extraction des descripteurs et similarité

Pour représenter le contenu des images, des caractéristiques de bas niveau sont extraites sur chacune des images. Les descripteurs couleur et mouvement sont introduits, puis une mesure de similarité entre les différents descripteurs est ensuite présentée pour juger de la ressemblance entre les images.

6.2.1.1 Couleur

Parmi les descripteurs couleur, nous avons retenu l'histogramme couleur qui offre une grande simplicité et capture la distribution des couleurs dans l'image. L'espace couleur choisi est l'espace YCbCr qui permet de séparer la luminance de la chrominance. La composante Y correspond à la luminance et les chrominances Cb et Cr représentent respectivement les oppositions Rouge-Vert et Bleu-jaune. Comme le système visuel humain est plus sensible à la lumière qu'aux couleurs, cet espace YCbCr est utilisé dans le domaine compressé MPEG pour donner moins d'importance à la couleur. Cependant, nous ne quantifions pas uniformément l'espace YCbCr pour ne pas donner le même poids aux pixels proches du centre d'un bin de ceux qui sont localisés aux extrémités. Nous employons les ensembles flous pour donner à chaque pixel un degré d'appartenance à chacun des bins. Chaque composante de l'espace YCbCr est quantifiée en 8 bins comme montré dans la figure 6.3.

Pour chaque pixel p, les bins b_i, b_j et b_k sont définis respectivement pour chaque composante de l'espace YCbCr, de là découle le degré d'appartenance μ de chacun de ces bins :

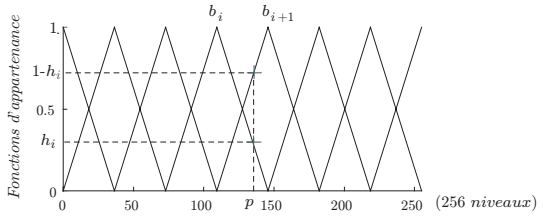


FIG. 6.3 – Degré d'appartenance d'un pixel à chaque bin.

$$\left\{ \begin{array}{l} \mu(b_i) = h_i \\ \mu(b_{i+1}) = 1 - h_i \end{array} \right. \quad \left\{ \begin{array}{l} \mu(b_j) = h_j \\ \mu(b_{j+1}) = 1 - h_j \end{array} \right. \quad \left\{ \begin{array}{l} \mu(b_k) = h_k \\ \mu(b_{k+1}) = 1 - h_k \end{array} \right. \tag{6.1}$$

L'histogramme couleur 3D flou h_{3d} est mis à jour de la façon suivante :

$$\text{Pour } (i,j,k = 1 \text{ à } 8), \quad h_{3d}(b_i, b_j, b_k) = h_{3d}(b_i, b_j, b_k) + \mu(b_i)\mu(b_j)\mu(b_k) \tag{6.2}$$

Après une normalisation par la taille de l'image, l'histogramme couleur flou 3D avec 8x8x8 bins est obtenu. Celui-ci correspond à l'histogramme couleur global. Néanmoins, pour être indépendant aux mouvements des objets, l'extraction d'un histogramme couleur local est aussi réalisée. L'image est simplement divisée en imagettes et sur chacune de ces imagettes, l'histogramme couleur est calculé. La normalisation des histogrammes s'effectue alors par la taille de l'imagette. Dans notre expérimentation, l'image est divisée en 9 blocs (3x3). Finalement, un histogramme couleur peut être extrait soit de manière locale ou globale sur chaque image de la vidéo.

6.2.1.2 Mouvement

Le mouvement est un critère important à prendre en compte pour détecter les transitions entre les plans d'une vidéo. La comparaison entre les images ne peut pas se limiter à des histogrammes couleur. En effet, si un mouvement de caméra de forte amplitude a lieu, la cohérence temporelle entre les descripteurs couleur des images sera difficile, et par conséquent les comparaisons entre les images y seront sensibles. Il est donc nécessaire d'estimer le mouvement de caméra. Nous utilisons l'algorithme développé dans la section 4.3 qui fournit une estimation paramétrique du mouvement dominant entre deux images successives.

$$\left\{ \begin{array}{l} v_x(p_i) = c_1 + a_1 \cdot x_i + a_2 \cdot y_i \\ v_y(p_i) = c_2 + a_3 \cdot x_i + a_4 \cdot y_i \end{array} \right. \tag{6.3}$$

Les coefficients $[c_1, \cdots, a_4]$ du modèle affine sont estimés et caractérisent le mouvement de caméra $v = (v_x(p_i), v_y(p_i))$ en fonction de la position du pixel $p_i = (x_i, y_i)$ entre deux images successives.

6.2.1.3 Similarité entre deux images

La similarité entre deux images est souvent déterminée à partir d'une distance entre les descripteurs. Parmi les différentes distances, nous utilisons la norme L1 entre les histogrammes couleur (distance dite de Manhattan qui correspond à la somme des valeurs absolues des écarts entre deux vecteurs caractéristiques). Cette distance notée d_c est comprise entre 0 et 2 avec

une valeur de 0 si les histogrammes sont identiques et une valeur de 2 si les histogrammes sont complètement différents.

En ce qui concerne les histogrammes couleur locaux, nous calculons les distances (Norme L1) sur chacune des 9 imagettes. Le minimum de ces 9 distances est choisi pour représenter la distance entre deux images. L'intérêt de cette distance est qu'elle est peu sensible aux mouvements des objets. En effet, il suffit que la distance entre deux imagettes soit faible pour considérer les images proches. Du fait de la normalisation des histogrammes, les distances entre histogrammes locaux ou globaux possèdent la même dynamique, comprise entre 0 et 2. Ainsi, quelle que soit la distance utilisée entre deux images (locale ou globale), elle sera notée d_c et il sera précisé s'il s'agit de la distance couleur locale ou globale.

Pour pouvoir appliquer facilement des règles sur la variable numérique d_c et détecter les transitions, nous avons employé le Modèle des Croyances Transférables (MCT). Nous ne rappellerons pas ici cette théorie (MCT), une description a déjà été donnée à la section 4.4. La variable numérique d_c peut être exprimée selon des termes linguistiques, elle est donc transformée en valeurs symboliques : très faible (TF), faible (F), moyen (M), grand (G) et très grand (TG). Soit $\Omega_{d_c} = \{TF, F, M, G, TG\}$ le cadre de discernement de la variable numérique d_c. Une fonction de masse (BBA) m_c est définie et permet l'attribution d'une masse à chaque terme linguistique ou groupe (très faible, faible, moyen, grand et très grand). Après une étude expérimentale, les fonctions d'appartenance qui permettent d'attribuer les masses ont été fixées comme indiqué sur la figure 6.4. Par exemple, pour une valeur numérique $d_c = 0.0425$, les masses non nulles sont les suivantes : $m_c(TF) = 0.5$, masse de croyance sur distance très faible, et $m_c(TF \cup F) = 0.5$, masse de croyance sur le doute entre distance très faible et distance faible.

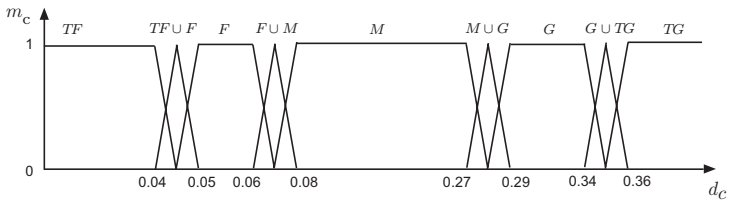


FIG. 6.4 – Définition de BBA pour la distance d_c entre histogrammes couleur (locaux ou globaux).

Pour ce qui est du mouvement de caméra, nous nous intéressons uniquement aux mouvements de translation c'est-à-dire aux coefficients $[c_1, c_2]$ (pour plus de détails sur les coefficients, se reporter à la section 4.3). En effet, c'est essentiellement le mouvement de translation de forte amplitude qui génère des distances élevées entre les histogrammes couleur. L'information obtenue par $dp(t) = \sqrt{(c_1^2 + c_2^2)}$ renseigne sur le déplacement de translation entre deux images successives aux instants t et $t+1$. Plus le déplacement dp est élevé plus la translation est importante. De même que pour la distance d_c, la variable numérique dp est transformée en valeurs symboliques : faible (F) et grand (G). Nous considérons le cadre de discernement $\Omega_{dp} = \{F, G\}$ pour distinguer les mouvements de translation de faible ou grande amplitude. Une fonction de masse (BBA) m_m est obtenue suivant les fonctions d'appartenance définies dans la figure 6.5.

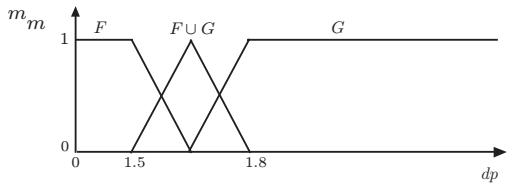


FIG. 6.5 – Définition de BBA pour le déplacement dp en caractérisant les grands et faibles mouvements de translation.

6.2.2 Règles pour la détection des transitions

Comme déjà signalé, nous proposons une méthode de détection des transitions fondée sur le Modèle des Croyances Transférables. Cette théorie présente l'intérêt de pouvoir implémenter facilement des règles qui répondent à notre expertise.

Le modèle de détection que nous proposons est divisé en 7 étapes (Fig. 6.2). L'idée est simplement de supprimer les images de la vidéo n'appartenant pas à une transition. Dès la première étape, nous sélectionnons des transitions éventuelles de manière assez exhaustive pour n'en oublier aucune. En terme de rappel et de précision, il s'agit d'avoir un rappel égal à 1 avec une précision éventuellement faible. Puis, par étapes successives, on opère à un raffinement de la sélection des transitions. Soit $Tr = \{T, \overline{T}\}$ le cadre de discernement de la transition où T est une hypothèse sur la présence d'une transition et $\overline{T}$ est une hypothèse sur l'absence de transition. L'attribution des masses de croyance sur Tr est effectuée à partir de règles que nous allons définir.

La première étape (Fig. 6.2) consiste à localiser les transitions potentielles. Comme la couleur est relativement stable entre deux images successives, chaque image est représentée par son histogramme couleur global. Le calcul des distances d_c entre deux images successives permet de repérer d'éventuelles transitions. Comme nous ne souhaitons conserver que les images où les distances sont importantes, nous formulons la règle d'implication suivante : Si la distance $d_c(t)$ entre les images t et $t+1$ est très faible (TF) alors l'image t n'appartient pas à une transition ($\overline{T}$), sinon l'image t appartient à une transition (T). Cette règle d'implication permet de passer du cadre de discernement d_c au cadre $Tr = \{T, \overline{T}\}$. Une fois, les masses obtenues sur le cadre de discernement Tr pour chaque image de la vidéo, nous calculons la probabilité pignistique sur l'ensemble Tr puis l'hypothèse qui a le maximum est choisie comme état de l'image (soit transition ou non transition). Finalement, la décision est réalisée au niveau de chaque image. Les images adjacentes appartenant à une transition sont alors regroupées pour obtenir les segments éventuels de transition dans la vidéo. Ainsi chaque segment contient un certain nombre d'images (deux ou plusieurs images).

Comme les transitions ont été détectées aux endroits où les distances ne sont pas très faibles (TF), une contrainte est rajoutée. Les transitions sont conservées uniquement s'il existe au moins une distance d_c dans la transition qui soit moyenne (M), grande (G) ou très grande (TG). La deuxième étape (Fig. 6.2) consiste à appliquer la règle d'implication suivante : Si le maximum des distances $d_c(t)$ sur un segment donné est moyen (M), grand (G) ou très grand (TG) alors le segment correspond à une transition (T) sinon ce segment n'est pas une transition ($\overline{T}$). Des masses de croyance sont alors obtenues sur Tr et l'hypothèse qui a le maximum de croyance est choisie comme état de la transition. Le segment est alors soit

conservé soit supprimé.

La troisième étape (Fig. 6.2) concerne uniquement les segments possédant plus d'une image. Supposons un fondu enchaîné entre les instants t_1 et t_2. L'intensité de l'image $I(p_i, t)$ à l'intérieur de la transition est alors modélisée par l'équation 6.4 et la distance entre deux images consécutives $d\left(I(p_i, t), I(p_i, t+1)\right)$ est donnée par la norme de leur différence (Eq. 6.5) :

$$I(p_i, t) = I(p_i, t_1)\alpha(t) + I(p_i, t_2)(1 - \alpha(t)) \tag{6.4}$$

$$\begin{aligned} d\left(I(p_i, t), I(p_i, t+1)\right) &= \|I(p_i, t+1) - I(p_i, t)\| \\ &= \|\left(I(p_i, t_2) - I(p_i, t_1)\right)\left(\alpha(t+1) - \alpha(t)\right)\| \\ &= \|\alpha(t+1) - \alpha(t)\|\, d\left(I(p_i, t_1), I(p_i, t_2)\right) \end{aligned} \tag{6.5}$$

où $\alpha(t)$ est une fonction décroissante avec $\alpha(t_1) = 1$ et $\alpha(t_2) = 0$. Si on suppose la fonction $\alpha(t)$ linéaire alors $\|\alpha(t+1) - \alpha(t)\| = 1/(t_2 - t_1)$. La somme des distances entre l'image à l'instant t_1 et l'image à l'instant t_2 est alors équivalente à une distance entre les première et dernière images de la transition (Eq. 6.6). Ainsi la somme des distances doit être du même ordre de grandeur que la distance entre deux images d'une transition instantanée.

$$\sum_{t=t_1}^{t_2-1} d\left(I(p_i, t), I(p_i, t+1)\right) = d\left(I(p_i, t_1), I(p_i, t_2)\right) \tag{6.6}$$

Le même raisonnement peut être repris non pas sur les images en niveaux de gris $I(p_i, t)$ mais sur les histogrammes couleur $h_{3d}(b_i, b_j, b_k, t)$:

$$h_{3d}(b_i, b_j, b_k, t) = h_{3d}(b_i, b_j, b_k, t_1)\alpha(t) + h_{3d}(b_i, b_j, b_k, t_2)(1 - \alpha(t))$$

Une nouvelle règle d'implication est alors appliquée : Pour toute transition progressive, si la somme des distances $d_c(t)$ est grande (G) ou très grande (TG) alors le segment correspond à une transition (T) sinon ce segment n'est pas une transition ($\overline{T}$). Le segment est soit conservé soit supprimé suivant les masses de croyance portées sur les ensembles de Tr. De plus, si une transition progressive détectée n'est pas un fondu enchaîné mais un fondu ou si la transition progressive détectée correspond à un segment contenant une transition instantanée, cette étape la conserve puisque certaines distances à l'intérieur de la transition seront élevées.

L'étape 4 (Fig. 6.2) a pour objectif d'améliorer la détection des transitions instantanées. En effet, une transition progressive a pu être détectée à la place d'une transition instantanée. Par exemple, si la fin d'un plan ou le début du suivant a un mouvement de caméra conséquent, les images correspondantes peuvent avoir des distances $d_c(t)$ moyennes. Bien que la transition soit instantanée, elle peut être détectée comme une transition progressive. L'idée est alors de vérifier que la distance maximale à l'intérieur d'une transition progressive soit très grande (TG) pour s'assurer de la présence d'une transition instantanée. Celle-ci est alors réduite aux images où les distances sont grandes (G) ou très grandes (TG). La règle suivante est alors appliquée : Pour une transition progressive entre les instants t_1 et t_2, si la distance maximale $d_c(t)$ est très grande (TG) et si la distance $d_c(i)$ à l'instant $i = t_1$ est très faible (TF), faible (F) ou moyenne (M) alors l'image est supprimée de la transition et on recommence avec $i = i+1$ sinon l'image est conservée et le processus est arrêté. De même, ce procédé est répété en commençant par les images de droite de la transition. Si la distance maximale $d_c(t)$ est

très grande (TG) et si la distance $d_c(i)$ à l'instant $i = t_2$ est très faible (TF), faible (F) ou moyenne (M) alors l'image est supprimée de la transition et on recommence avec $i = i - 1$ sinon l'image est conservée et le processus est arrêté. En revanche, si la distance maximale $d_c(t)$ n'est pas très grande (TG) alors l'image est conservée. Le tableau 6.1 résume la règle d'implication et montre comment la combinaison des masses est réalisée.

TAB. 6.1 – Attribution des masses en fonction de la distance $d_c(i)$ et de la distance maximale dans la transition.

		$d_c(i)$				
		TF	F	M	G	TG
$max(d_c(t))$	TG	$\overline{T}$	$\overline{T}$	$\overline{T}$	T	T
	G, M, F, TF	T	T	T	T	T

Néanmoins, cette étape est réalisée uniquement si l'intensité des images de la transition n'est pas uniforme pour ne pas traiter le cas des fondus. A partir de l'histogramme couleur global, nous déterminons la composante b_i^* de l'histogramme ayant une proportion majoritaire dans l'image (Eq.6.7). Ainsi, si la composante b_i^* est égale à 1 (ou respectivement 8) alors l'image est plutôt de teinte noire (ou respectivement blanche). Nous considérons que l'image a une teinte uniforme si au moins $\alpha\%$ des pixels sont contenus dans la composante majoritaire. Dans notre expérimentation, α est choisi égal à 70%. Donc s'il existe au moins une image de teinte noire ou blanche dans la transition potentielle alors cette règle n'est pas appliquée.

$$h(b_i, t) = \sum_{b_j=1, b_k=1}^{8} h_{3d}(b_i, b_j, b_k, t) \text{ et } b_i^* = \arg\max_{b_i=1..8} (h(b_i, t)) \tag{6.7}$$

Pour éviter les fausses alarmes dues aux mouvements de caméra, le mouvement dominant est estimé sur chaque transition progressive. Le déplacement $dp(t)$ entre deux images successives va permettre de distinguer les véritables transitions de celles qui ont été détectées en raison des variations des histogrammes dues aux mouvements de caméra. Pour avoir une représentation du mouvement sur une transition progressive candidate, nous ne calculons pas la moyenne des déplacements mais le médian des déplacements sur la transition. La moyenne des déplacements est sensible aux données aberrantes. De plus, si le mouvement estimé entre deux images de la transition est erroné (avec une amplitude très élevée), le médian permet d'être insensible à cette donnée aberrante. Lors de l'étape 5 (Fig. 6.2), nous traitons seulement les transitions de plus de trois images pour que le médian des déplacements ait un sens et nous leur appliquons la règle suivante : Si le mouvement de caméra est important, c'est-à-dire avec un médian des déplacements grand (G) et si la distance maximale entre les histogrammes couleur est moyenne (M), faible (F) ou très faible (TF) pour être sûr de ne pas traiter une transition instantanée alors la transition est supprimée. Le tableau 6.2 résume la règle d'implication et montre comment la combinaison des masses est réalisée.

L'étape 6 (Fig. 6.2) permet d'être robuste au flash ou au changement de luminosité. Pour se prémunir de ces fausses alarmes, la distance entre les histogrammes couleur du début et de fin de la transition est calculée. Parmi les transitions détectées, la règle d'implication est alors appliquée pour ne conserver que celles qui ont une distance grande (G) ou très grande (TG).

TAB. 6.2 – Attribution des masses en fonction de la distance $d_c(i)$ et du médian des déplacements de la transition.

		$max(d_c(t))$				
		TF	F	M	G	TG
$median(dp(t))$	G	$\overline{T}$	$\overline{T}$	$\overline{T}$	T	T
	F	T	T	T	T	T

Puis, le procédé est répété sur chaque transition en considérant l'histogramme couleur 3D flou local. L'utilisation d'un descripteur local permet d'être invariant aux mouvements des objets de grande taille. L'étape 7 consiste d'abord à effectuer l'étape 2 mais cette fois-ci, c'est la distance maximale entre histogrammes couleur locaux qui est utilisée. La règle est toujours la même : si la distance maximale est très faible (TF) alors la transition est supprimée ($\overline{T}$) sinon elle est conservée (T). Puis, les étapes 4 et 5 sont à nouveau effectuées en appliquant les distances entre histogrammes couleur locaux suivant les mêmes règles.

Finalement, à la fin de cette succession d'étapes, nous obtenons les transitions instantanées ou progressives des vidéos. Afin de juger de la performance de la méthode, une étude doit être effectuée.

6.3 Evaluation de la méthode de détection des changements de plan de la vidéo

La détection des changements de plan doit être évaluée pour vérifier l'efficacité de la méthode. Nous présentons successivement les vidéos utilisées, les mesures d'évaluation puis les résultats obtenus.

6.3.1 Base de vidéos

Des vidéos doivent être sélectionnées de manière à balayer les différentes catégories de vidéo fréquemment rencontrées. Nous avons donc choisi, pour la variété de leurs contenus, les quatre vidéos qui ont été présentées dans le chapitre 3 (Section 3.4.1.2).

- « Documentaire » avec 20 plans et 3272 images (13 transitions instantanées et 6 transitions progressives)
- « Série » avec 28 plans et 2496 images (27 transitions instantanées et 0 transition progressive)
- « Journal » avec 42 plans et 6872 images (36 transitions instantanées et 5 transitions progressives)
- « Générique » avec 44 plans et 3137 images (42 transitions instantanées et 1 transition progressive)

La vidéo « Générique » qui correspond au générique d'un épisode de la série « The Avengers » a la particularité de posséder beaucoup de plans avec de forts mouvements de caméra.

6.3.2 Mesures d'évaluation des méthodes de segmentation

L'évaluation des méthodes de segmentation en plans des vidéos consiste en général à établir manuellement une vérité terrain des transitions sur chaque vidéo. La comparaison entre

la vérité terrain et la méthode de segmentation est ensuite réalisée en mesurant le rappel et la précision. Ces mesures sont déterminées au niveau des images. Par exemple, si une transition instantanée est identifiée entre les instants t et $t+1$ et si la méthode de segmentation retourne les instants de $t-1$ jusqu'à $t+2$ alors le rappel et la précision pour cette transition valent respectivement 1 et 1/3. Ainsi cette technique d'évaluation permet de mesurer à la fois la détection des transitions et la précision sur la position des transitions, ce qui est une forte contrainte. En revanche, d'autres méthodes d'évaluation sont moins exigeantes avec la détection. Par exemple, dans [Sae04], une transition progressive est déclarée comme correcte si au moins 50% de la transition est détectée alors que dans [Bes05], la transition progressive est considérée correcte si la différence entre la détection et la vérité terrain est inférieure à 5 images.

D'autres critères peuvent être proposés pour évaluer la détection des transitions en privilégiant leur détection et en ne tenant pas compte de la précision. Dans les expérimentations de TREC Video [Tre05], une tâche concerne la détection des plans dans les vidéos. Le critère utilisé pour évaluer les méthodes est de considérer qu'une transition est correctement retrouvée s'il existe au moins une image qui se chevauche entre la transition soumise et la transition de référence. De plus, les transitions progressives soumises ne peuvent être regroupées qu'avec des transitions progressives de référence et les transitions instantanées soumises qu'avec des transitions instantanées de référence. Ils définissent également des transitions progressives courtes (short gradual transitions), transitions inférieures à un certain nombre d'images (5 images) qui sont traitées comme des transitions instantanées. Pour se prémunir des erreurs d'indice, ils autorisent aussi le regroupement entre des transitions instantanées soumises et des transitions instantanées de référence malgré un écart temporel de 5 images.

Par la suite, nous allons utiliser deux types d'évaluation : le rappel et la précision au niveau des images et les mesures proposée par TREC Video. La première requière la détection des transitions avec une bonne précision alors que la seconde demande seulement la détection.

6.3.3 Résultats

Les résultats de la détection des transitions sont présentés dans le tableau 6.3 où le rappel et la précision sont calculés au niveau des images pour chacune des étapes de la méthode. Nous pouvons observer que la méthode fournit de bonnes performances avec un rappel $\geq 98\%$ et une précision $\geq 77\%$ pour les quatre vidéos. La première étape permet une sélection d'images qui correspondent aux transitions. Pour les 4 vidéos, nous avons un rappel de 100% mais une précision faible. Les étapes successives vont au fur et à mesure permettre d'augmenter la précision tout en conservant un rappel élevé. De plus, pour les quatre vidéos, aucune fausse alarme n'est presentée. Cela signifie qu'il existe pour chaque transition détectée au moins une image de celle-ci qui appartient effectivement à une transition de la vérité terrain. Nous pouvons aussi constater que la vidéo « Journal » a les moins bons résultats. C'est en raison du dernier plan de cette vidéo qui est le générique de fin du journal télévisé et correspond à une succession d'images non continues et apparentées à des flashs.

TAB. 6.3 – Résultats de la méthode de détection des transitions avec un rappel et une précision calculés suivant chacune des étapes. (Commun = nombre d'images communes entre la vérité terrain et la méthode proposée, Vérité = nombre d'images situées dans des transitions, Détecté = nombre d'images soumises par la méthode et supposées appartenir à des transitions)

Vidéo	Etape	Commun	Vérité	Détecté	Précision	Rappel
Documentaire	1	77	77	449	0.17	1
	2	77	77	369	0.21	1
	3	77	77	322	0.24	1
	4	77	77	271	0.28	1
	5	77	77	125	0.62	1
	6	77	77	100	0.77	1
	7	77	77	90	0.86	1
Journal	1	52	52	410	0.13	1
	2	52	52	270	0.19	1
	3	51	51	217	0.24	1
	4	52	52	161	0.32	1
	5	52	52	97	0.54	1
	6	52	52	69	0.75	1
	7	51	52	66	0.77	0.98
Série	1	27	27	328	0.08	1
	2	27	27	235	0.11	1
	3	27	27	210	0.12	1
	4	27	27	166	0.16	1
	5	27	27	91	0.30	1
	6	27	27	71	0.38	1
	7	27	27	30	0.90	1
Générique	1	53	53	386	0.14	1
	2	53	53	254	0.21	1
	3	53	53	229	0.23	1
	4	53	53	155	0.34	1
	5	53	53	82	0.64	1
	6	53	53	74	0.71	1
	7	53	53	53	1	1

Afin de comparer la méthode de détection des transitions, nous développons une méthode de détection des transitions instantanées qui utilise les distances entre les images successives. Classiquement la comparaison des distances à un seuil permet de retrouver les transitions instantanées. La difficulté consiste à déterminer le seuil de manière optimale. La distance utilisée ici est soit la distance entre histogrammes couleur globaux soit la distance correspondant à la taille du support où l'estimation du mouvement dominant a eu lieu entre deux images successives. Ainsi si la taille du support est proche de la taille de l'image alors l'estimation du mouvement permet de faire concorder les deux images. En revanche, si la taille est faible, alors l'estimation ne permet pas une bonne mise en correspondance des images. Dans ce cas, un changement de plan a probablement lieu.

Le seuil pour signaler la présence des transitions instantanées est obtenu à partir d'un classifieur Bayesien. Nous considérons deux hypothèses (ou classes) : l'une correspond à l'absence de transition instantanée et l'autre à la présence d'une transition instantanée. En supposant les densités de probabilité gaussiennes, la règle bayésienne de décision permet de déterminer le seuil de séparation des deux classes. L'apprentissage a été effectué seulement avec les distances qui correspondent aux deux classes (transition instantanée et absence de transi-

tion). Les endroits où une transition progressive se produit n'ont pas été pris en compte pour l'apprentissage.

Le tableau 6.4 présente les résultats de l'apprentissage sur les 4 vidéos. La détection des transitions instantanées a été réalisée pour chaque vidéo avec l'apprentissage sur cette même vidéo. Il faut aussi préciser que ces valeurs correspondent uniquement à la détection de transitions instantanées et les endroits où se situe une transition progressive dans la vidéo ne sont pas évalués. Nous pouvons observer que la détection des transitions instantanées n'est pas très bonne puisque la précision n'est pas élevée, ce qui se traduit par de nombreuses fausses alarmes. La comparaison des histogrammes couleur n'est pas performante pour détecter des transitions instantanées, ce qui prouve que les vidéos choisies ont un contenu qui ne permet pas une détection élémentaire des transitions instantanées par simple seuillage. L'utilisation du descripteur mouvement est parfois plus efficace pour détecter les transitions instantanées. Par exemple, les résultats obtenus pour la vidéo « Générique » sont très bons avec un rappel et une précision de 100%, mais pour la vidéo « Documentaire » le rappel et la précision ne sont respectivement que de 100% et 35%. La combinaison des descripteurs couleur et mouvement n'améliore pas les résultats par rapport aux descripteurs pris séparément. Cependant, le classifieur est appliqué sur la même vidéo d'apprentissage, c'est-à-dire sur des données qui ont déjà été apprises.

Tab. 6.4 – Détection des transitions instantanées suivant différents descripteurs (couleur, mouvement ou combinaison des deux) avec un classifieur Bayesien. La détection est effectuée sur la même vidéo où l'apprentissage a eu lieu. (Commun = nombre d'images communes entre la vérité terrain et la méthode proposée, Vérité = nombre d'images situées dans des transitions, Détecté = nombre d'images soumises par la méthode et supposées appartenir à des transitions)

Vidéo	Descripteur	Commun	Vérité	Détecté	Précision	Rappel
Documentaire	Couleur	13	13	26	0.50	1
	Mouvement	13	13	37	0.35	1
	Combinaison	13	13	30	0.43	1
Journal	Couleur	34	35	52	0.65	0.97
	Mouvement	34	35	48	0.71	0.97
	Combinaison	34	35	50	0.68	0.97
Série	Couleur	27	27	48	0.56	1
	Mouvement	27	27	33	0.82	1
	Combinaison	27	27	35	0.77	1
Générique	Couleur	42	42	69	0.61	1
	Mouvement	42	42	42	1	1
	Combinaison	42	42	48	0.88	1

Le tableau 6.5 présente les résultats dans le cas où la détection est réalisée sur des vidéos différentes de l'apprentissage. L'apprentissage a été effectué sur la vidéo « Générique » et le classifieur a été réalisé sur les trois autres vidéos. Par exemple, la détection suivant le descripteur couleur est alors moins efficace que celle obtenue avec le classifieur appliqué sur la même vidéo d'apprentissage (Tab. 6.4).

Néanmoins, les résultats présentés jusqu'ici ne concernent que les transitions instantanées qui sont considérées comme les plus simples à détecter. Si les deux types de transitions sont considérés (instantanée et progressive) pour réaliser l'apprentissage, la détection des transi-

tions est alors moins performante qu'auparavant. Par exemple, si un apprentissage est réalisé sur la vidéo « Générique » où une seule transition (transition progressive) est rajoutée, la détection sur cette même vidéo diminue avec une précision de 43% et un rappel de 79% pour le descripteur mouvement alors que celui-ci était très performant en considérant seulement les transitions instantanées.

TAB. 6.5 – Détection des transitions instantanées suivant différents descripteurs (couleur, mouvement ou combinaison des deux) avec un classifieur Bayesien. L'apprentissage est réalisé sur la vidéo « Générique » puis le classifieur est appliqué sur ces trois vidéos. (Commun = nombre d'images communes entre la vérité terrain et la méthode proposée, Vérité = nombre d'images situées dans des transitions, Détecté = nombre d'images soumises par la méthode et supposées appartenir à des transitions)

Vidéo	Descripteur	Commun	Vérité	Détecté	Précision	Rappel
Documentaire	Couleur	13	13	69	0.19	1
	Motion	12	13	12	1	0.92
	Combinaison	13	13	20	0.65	1
Journal	Couleur	35	35	58	0.60	1
	Motion	34	35	48	0.71	0.97
	Combinaison	35	35	52	0.67	1
Série	Couleur	27	27	83	0.33	1
	Motion	27	27	34	0.79	1
	Combinaison	27	27	50	0.54	1

Les mesures utilisées jusqu'ici pour l'évaluation des transitions demandent à la fois de détecter des transitions et d'être précis sur le début et la fin des transitions. L'évaluation selon le protocole TREC Video est présentée dans le tableau 6.6. Cette évaluation privilégie la détection et ne tient pas compte de la précision de la détection. Si on compare la précision et le rappel du tableau 6.3 avec ceux du tableau 6.6, on s'aperçoit que les résultats sont meilleurs dans le second cas avec un rappel et une précision supérieurs à 94% pour la détection des transitions (intantanées et progressives) et pour les 4 vidéos. Cela confirme que cette évaluation privilégie la détection.

Dans le tableau 6.7, nous présentons des résultats sur 5 vidéos comportant des journaux TV, de la publicité et un épisode de la série « The Avengers ». La base de vidéos contient 159723 images. Nous comparons la méthode proposée avec une méthode basée sur des comparaisons d'histogrammes élaborée par le projet MOCA [Moc]. Nous pouvons constater que la méthode proposée fournit les meilleurs résultats. Néanmoins, la méthode du Projet MOCA qui repose sur les logiciels « Cutdet » et « Fadedet » réalise des comparaisons d'histogrammes avec des seuils fixés par défaut. Comme la méthode précédemment décrite (avec la comparaison d'histogrammes et sélection du seuil avec un classifieur Bayesien) le suggérait, il n'est pas surprenant que la méthode du Projet MOCA donne de moins bons résultats.

TAB. 6.6 – Evaluation de la détection des transitions selon le protocole TREC Video sur 4 vidéos.

		Documentaire	Journal	Série	Générique
Transition (instantanée et progressive)	Nombre d'images	3272	6872	2496	3137
	Transition de référence	19	41	27	43
	Transition insérée	1	1	0	0
	Transition supprimée	1	1	0	0
	Rappel	0.947	0.975	1.0	1.0
	Précision	0.947	0.975	1.0	1.0
Transition instantanée	Taille maximale (transit. prog. courtes)	5	5	5	5
	Transition de référence	17	41	27	42
	Transition insérée	0	0	0	0
	Transition supprimée	1	1	0	0
	Rappel	0.941	0.975	1.0	1.0
	Précision	1.0	1.0	1.0	1.0
Transition progressive	Transition de référence	2	0	0	1
	Transition insérée	1	1	0	0
	Transition supprimée	0	0	0	0
	Rappel	1.0	0.0	0.0	1.0
	Précision	0.666	0.0	0.0	1.0
	Rappel (images)	54/54	0.0	0.0	10/10
	Précision (images)	54/62	0.0	0.0	10/10

TAB. 6.7 – Evaluation de la détection des transitions selon le protocole TREC Video sur une base de vidéos comprenant 159723 images.

	Méthode proposée	Méthode MOCA
Rappel	1720/2064	1435/2064
Précision	1720/1873	1435/1824
Rappel (instantanée)	1663/1934	1425/1934
Précision (instantanée)	1663/1756	1425/1813
Rappel (progressive)	57/130	10/130
Précision (progressive)	57/117	10/11

6.4 Conclusion

Nous avons présenté une méthode de détection des transitions instantanées et progressives basée sur le Modèle des Croyances Transférables. Cette approche consiste à localiser dans la vidéo les changements de plan quel que soit le type de transition. Elle s'appuie sur une structure unifiée sous forme d'arbre divisé en 7 étapes. Chaque étape réalise un traitement qui repose sur un certain nombre de règles heuristiques. Les première et deuxième étapes (étapes 1 et 2) permettent une sélection exhaustive des transitions éventuelles. Celles-ci ont pour conséquence d'obtenir un rappel élevé (proche de 1) et une précision faible. Les étapes suivantes (étape 3 à 7) permettent une suppression des fausses alarmes afin d'augmenter la précision. Elles sont conçues pour résoudre différents problèmes comme la détection de transitions due aux mouvements de caméra, des objets ou à la présence de flashs.

Nous avons effectué une étude comparative entre plusieurs méthodes. Tout d'abord, une méthode classique basée sur des comparaisons d'histogrammes couleur avec sélection du seuil par un classifieur Bayesien a été effectuée pour détecter les transitions instantanées. Puis, la même étude a été réalisée en utilisant un descripteur mouvement avec sélection du seuil par le classifieur Bayesien. Enfin la combinaison entre l'histogramme couleur et le descripteur mouvement a été effectuée. Les résultats montrent combien il est difficile de retrouver les transitions instantanées. Cette étude prouve également que les vidéos choisies ont un contenu qui ne permet pas une détection élémentaire des transitions instantanées par simple seuillage. De plus, la combinaison des descripteurs n'améliore pas forcément les résultats par rapport aux descripteurs pris séparément. Les résultats obtenus en terme de rappel et de précision nous permettent de conclure que notre méthode basée sur le MCT fournit de bons résultats avec un rappel et une précision élevés.

La méthode proposée a ensuite été comparée avec une méthode basée sur des comparaisons d'histogrammes du projet MOCA. Notre méthode de détection a obtenu les meilleurs résultats avec un rappel et une précision supérieurs à ceux de la méthode du projet MOCA. Nous avons montré l'efficacité de notre méthode de détection des changements de plan.

Chapitre 7

Création de résumé de vidéo à partir de l'attention visuelle

Dans ce chapitre, nous nous intéressons à l'attention visuelle pour mettre en valeur les régions saillantes dans les images. L'objectif est de développer un modèle d'attention visuelle qui facilitera l'analyse du contenu des vidéos et qui sera exploité dans la création de résumé de vidéo.

Nous présentons le modèle spatio-temporel d'attention suivi d'une expérience psychophysique pour le valider. Puis, nous montrons comment ce modèle contribue à une nouvelle méthode de création de résumé de vidéo.

Sommaire

7.1 Introduction

Malgré les avancées récentes dans le traitement des vidéos et dans la détection des événements, l'extraction du contenu sémantique des vidéos est encore loin d'être réalisée. Pour réduire le fossé entre les caractéristiques de bas niveau et celles de plus haut niveau sémantique, l'attention visuelle a été introduite. Elle est destinée à faciliter l'analyse des vidéos en vue d'améliorer l'indexation du contenu.

Avant de discuter des modèles existants, nous devons donner une définition plus précise de l'attention. Comme Rousselet et al. l'expliquent dans [Rou03], la perception d'une scène est dirigée par deux types de modulations : les modulations attentionnelles descendantes dépendantes de la tâche (« top-down ») et les modulations attentionnelles ascendantes dépendantes de la saillance de certaines caractéristiques du stimulus (« bottom-up »). Par exemple, les modulations descendantes concernent les tâches de recherche, comme retrouver un objet donné dans une image naturelle. D'après Rousselet et al. [Rou03], ce type de tâche fait intervenir des modulations descendantes qui prennent leur origine dans des « structures hiérarchiques élevées de l'architecture cérébrale pour influencer les structures plus primaires du traitement de l'information ». De la même manière, les modulations attentionnelles ascendantes vont privilégier certains stimuli du champ visuel, et ces informations ascendantes correspondent à la saillance de l'objet. Or la saillance d'un objet est liée à ses caractéristiques physiques. Ainsi l'attention est portée automatiquement sur les objets les plus saillants de la scène. Cependant les deux systèmes de modulations interagissent constamment pour cibler les objets vers lesquels notre attention va être dirigée. Comme nous souhaitons connaître les régions où notre regard s'est porté durant le visionnage d'une vidéo, nous aborderons uniquement la modélisation de l'attention ascendante puisqu'aucune tâche spécifique est demandée. Elle est obtenue à partir de l'information bas niveau de l'image (intensité, orientation, ...) et dirigée par les attributs des stimuli dans une scène. L'étude de l'attention visuelle ascendante a pour but de fournir les régions saillantes dans les images et donc renseigner sur les extraits importants dans les vidéos. Par la suite, le terme « attention » fera référence à l'attention visuelle ascendante.

L'attention a d'abord été étudiée dans le traitement des images fixes où seule l'information spatiale a été considérée. Le premier modèle d'attention a été proposé par Koch et Ullman [Koc85] en 1985. Puis Itti et al. [Itt99] ont défini un modèle qui combine différentes cartes créées à partir de caractéristiques de bas niveau (orientation, intensité et couleur). D'autres approches ont créé des modèles plus élaborés. Chauvin et al. [Cha02] ont développé un modèle inspiré de la rétine et des fonctionnalités des cellules du cortex visuel primaire. Plus récemment, l'attention a été abordée dans le traitement des vidéos. Les méthodes ont alors exploité la composante temporelle et se sont basées sur l'estimation du mouvement [Ma02]. De nouveaux systèmes sont apparus en combinant des cartes d'attention visuelle statiques et dynamiques [Ho03, Cou03, Meu05b]. Dans [Che03a], les auteurs ont proposé un modèle d'attention basé sur des caractéristiques visuelles statiques mais également sur la détection des visages et du texte. Généralement, ces modèles informatiques ne sont pas inspirés des fonctionnalités des cellules du système visuel humain et les cartes de saillance obtenues sont directement employées dans diverses applications comme le résumé de vidéo, le codage, la surveillance. Enfin, quelques travaux [Mac02, Nav02, Oli03, Mac05] ont étudié plus spécifiquement l'attention descendante (« top-down »), mécanisme important dans les tâches de reconnaissance d'objets. Par exemple, des méthodes [Tor03, Gui05] proposent de modéliser les relations entre le contexte et l'objet et d'utiliser les informations de contexte pour faciliter la détection des objets. Comme déjà signalé, nous n'avons pas étudié ce mécanisme.

Dans ce chapitre, nous proposons un nouveau système pour modéliser l'attention visuelle ascendante. Il repose sur la fusion d'un modèle statique inspiré du système humain avec un modèle de détection d'objets en mouvement dans une scène. Le modèle statique est basé sur un filtrage rétinien suivi d'un banc de filtres de Gabor. La détection des objets en mouvement est effectuée en compensant le mouvement de caméra. Une fois le modèle spatio-temporel d'attention construit, nous avons proposé une expérience psychophysique pour tester et valider le modèle proposé. Une nouvelle méthode de résumé de vidéo a ensuite été conçue à partir de ce modèle d'attention.

La suite du chapitre est organisée de la façon suivante. Dans la section 7.2, le modèle d'attention est présenté. Une expérience psychophysique est exposée dans la section 7.3 pour juger le modèle d'attention. Enfin, une méthode de résumé est conçue à partir de ce modèle dans la section 7.4.

7.2 Modèle d'attention visuelle

Nous allons décrire le modèle d'attention qui va permettre d'extraire les régions saillantes des images de la vidéo. Il est divisé en deux parties : une partie statique et une partie dynamique. L'architecture du modèle est présentée dans la figure 7.1. La suite de cette section décrit chaque partie du modèle.

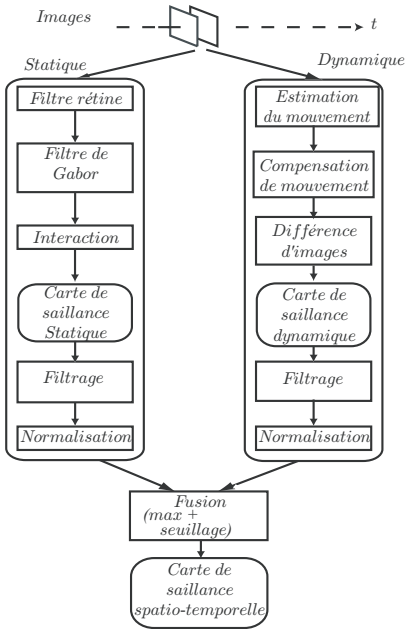


FIG. 7.1 – Architecture du modèle spatio-temporel d'attention.

7.2.1 Partie statique du modèle

Cette partie est inspirée de la biologie et des fonctionnalités des cellules du système visuel humain (de la rétine au cortex visuel primaire). Celle-ci sera appliquée à chaque image de la vidéo pour obtenir des cartes de saillance statiques. La construction de la carte de saillance

statique s'appuie sur les précédents travaux de thèse de A. Chauvin [Cha03] effectués au sein de notre laboratoire.

Filtrage rétinien

Au premier niveau de traitement des informations visuelles, les photorécepteurs rétiniens effectuent une compression adaptative de l'intensité lumineuse suivie d'un filtrage passe-haut [Bea96]. Ce procédé conduit à une égalisation du contraste de l'image, fournissant une insensibilité relative aux variations locales d'illumination. Ce prétraitement est intéressant pour extraire la saillance puisqu'il est insensible à certaines modifications, comme les variations de luminosité entre les images. Puis, la voie parvocellulaire est réalisée par un filtre spatial passe-haut (pour blanchir le spectre de l'image) qui compense le spectre d'amplitude de l'image en $1/f$. La figure 7.2 montre un exemple de filtrage rétinien. De plus amples explications sur l'implémentation de ce filtre peuvent être trouvées dans la thèse de N. Guyader [Guy04].

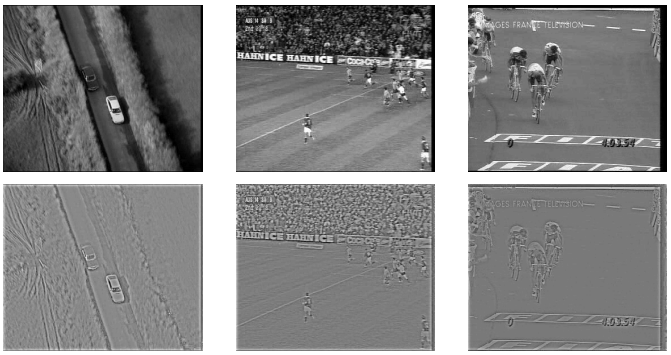


FIG. 7.2 – Exemple de filtrages rétiniens. En haut : Image de vidéos. En bas : Image après filtrage.

Cortex visuel primaire

Dans le cortex visuel primaire, des cellules sont sensibles aux orientations et aux fréquences spatiales du signal visuel. Ici, nous avons choisi de modéliser le champ récepteur des cellules simples. Ces cellules sont sensibles aux stimuli ayant une certaine orientation et une certaine fréquence avec une position spécifique dans le champ visuel, celles-ci sont modélisées par des filtres bidimensionnels de Gabor. Une fonction de Gabor est définie par une gaussienne d'écarts-type σ_x et σ_y, modulée par une exponentielle complexe de fréquence f dans la direction θ. Nous avons choisi 4 bandes de fréquence $f_m = \frac{0.3}{2^m}$ avec $m = [0 \cdots 3]$ et 8 orientations $\theta_n = \frac{n\pi}{8}$ avec $n = [0 \cdots 7]$.

$$\begin{aligned} g(x, y, f_m, \theta_n, \sigma_x, \sigma_y) &= \tfrac{1}{2\pi\sigma_x\sigma_y} e^{2\pi j f_m x'} e^{-(\frac{x'}{2\sigma_x^2} + \frac{y'}{2\sigma_y^2})} \\ &\begin{cases} x' = x \cdot \cos\theta_n + y \cdot \cos\theta_n \\ y' = y \cdot \cos\theta_n - x \cdot \cos\theta_n \end{cases} \end{aligned} \tag{7.1}$$

Nous avons effectué ce filtrage en multipliant directement l'image de sortie de la rétine avec le filtre de Gabor dans le domaine de Fourier. Avant de réaliser la transformée de Fourier, nous avons multiplié l'image par une fenêtre de Hanning pour réduire les effets de bord. Nous avons choisi ici de décomposer chaque image de la vidéo en utilisant 32 filtres de Gabor (quatre

fréquences spatiales et huit orientations). La figure 7.3 illustre les filtres de Gabor utilisés. Ainsi, nous obtenons 32 cartes (module de sortie des filtres de Gabor) notées $S(f_i, \theta_j)$, qui dépendent non seulement de la fréquence spatiale et de l'orientation, mais aussi de la position spatiale dans l'image.

Interactions

Un neurone est, par définition, une cellule nerveuse en contact avec d'autres neurones. Ainsi, la réponse d'une cellule est dépendante de la réponse des neurones voisines. L'activité du neurone est donc modulée par le voisinage du champ visuel selon des connexions latérales. En ce qui concerne les orientations, les interactions connectent préférentiellement les neurones de même orientation et permettent d'accentuer les contours. Ces interactions, qui symbolisent à la fois des connexions excitatrices et inhibitrices, sont modélisées par une combinaison linéaire des cellules simples :

$$S_{int}(f_i, \theta_j) = \sum_{k,l} w_{k,l} \cdot S(f_{i+k}, \theta_{j+l}), w = \begin{pmatrix} 0 & -0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & -0.5 & 0 \end{pmatrix} \tag{7.2}$$

La figure 7.3 illustre les interactions. Dans cet exemple, le filtre de Gabor avec le poids 1 dans la direction $\pi/4$ interagit avec ses voisins. Les filtres de Gabor avec la même direction symbolisent l'excitation et ceux avec des directions différentes représentent l'inhibition. Pour les filtres de Gabor n'ayant pas quatre voisins (par exemple les filtres à la fréquence $f = 0.3$ et aux différentes orientations), la matrice w est modifiée afin de le prendre en compte dans les interactions. La sortie de cette étape se compose de 32 cartes pour chaque image de la

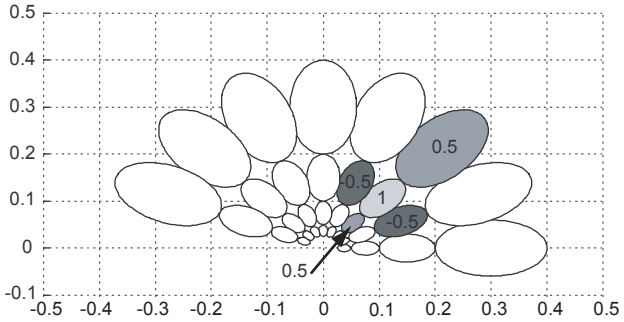


FIG. 7.3 – Exemple des interactions dans le domaine de Fourier. Le filtre de Gabor avec un poids de 1 dans la direction $\pi/4$ interagit avec ses voisins.

vidéo. Chaque carte met en évidence l'énergie se trouvant dans l'image en fonction de la fréquence spatiale et de l'orientation donnée, et prend en compte l'interaction entre les cartes d'orientation.

Carte de saillance statique

Nous obtenons une carte de saillance statique pour chaque image en effectuant la somme des 32 cartes d'énergie décrites ci-dessus :

$$S_f = \left| \sum_{i,j} S_{int}(f_i, \theta_j) \right| \tag{7.3}$$

Les régions ayant les énergies les plus élevées sont considérées comme saillantes. La figure 7.4 montre un exemple de cartes de saillance statiques. Le contenu des images est plutôt varié : un match de rugby, une course de voiture et une course de vélo. Nous pouvons observer que l'énergie est située sur les objets qui semblent être saillants.

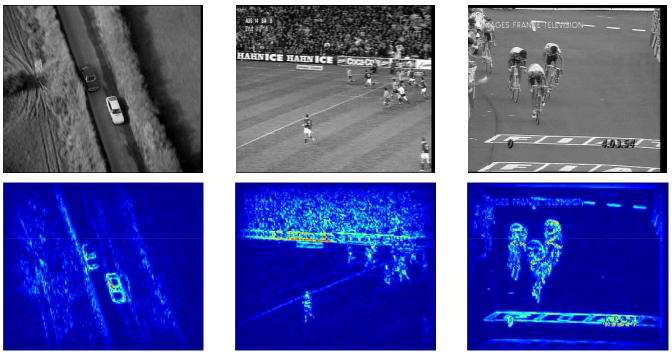


FIG. 7.4 – Exemple de cartes de saillance statiques. En haut : Images extraites de 3 vidéos. En bas : Cartes de saillance statiques correspondantes.

7.2.2 Partie dynamique du modèle

La partie dynamique du modèle détecte les objets en mouvement dans la scène. En fait, nous supposons que l'endroit où quelque chose bouge est saillant. Ainsi, il est nécessaire d'estimer le mouvement de caméra afin de ne conserver que le mouvement des objets. Nous utilisons l'algorithme d'estimation du mouvement décrit à la section 4.3. Cet algorithme fournit le mouvement dominant entre deux images successives. Nous avons choisi le modèle affine pour représenter le mouvement de caméra et la figure 7.5 montre un exemple de l'estimation du mouvement de caméra.

$$\begin{cases} v_x(p_i) = c_1 + a_1 \cdot x_i + a_2 \cdot y_i \\ v_y(p_i) = c_2 + a_3 \cdot x_i + a_4 \cdot y_i \end{cases} \tag{7.4}$$

Les coefficients $[c_1, \cdots, a_4]$ du modèle affine caractérisent le mouvement de caméra $v = (v_x(p_i), v_y(p_i))$ en fonction de la position du pixel $p_i = (x_i, y_i)$ entre deux images successives. Une fois les coefficients $[c_1, \cdots, a_4]$ estimés, nous réalisons une interpolation bilinéaire pour compenser le mouvement de l'image $I_c(x, y, t+1) = I(x+v_x, y+v_y, t+1)$. L'image précédente est ensuite soustraite de l'image compensée $I_c(x, y, t+1)$ pour générer la différence d'images déplacées (DFD) :

$$DFD(x, y, t) = I_c(x, y, t+1) - I(x, y, t) \tag{7.5}$$

Finalement, la valeur absolue de DFD informe sur les régions qui ne suivent pas le modèle de mouvement et correspond aux déplacements des objets. La figure 7.6 montre un exemple de détection d'objets. Nous pouvons voir que les joueurs de rugby sont correctement détectés de même que les voitures et les cyclistes.

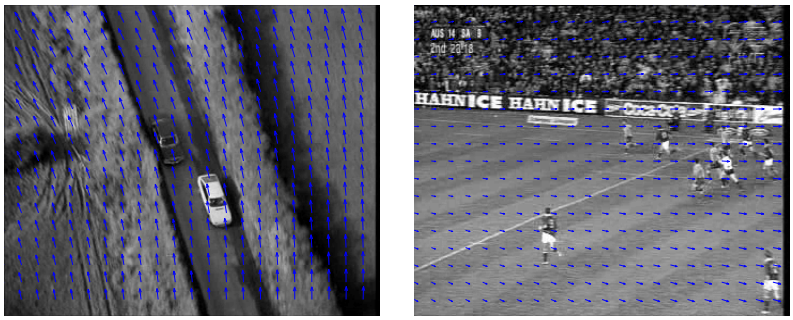


FIG. 7.5 – Exemple d'estimations du mouvement de caméra. Un modèle affine de mouvement est utilisé pour représenter le mouvement de caméra.

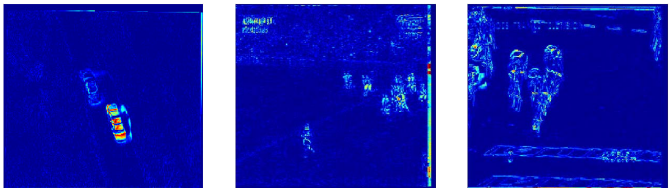


FIG. 7.6 – Exemple de détection d'objets en mouvement. Les images représentent la valeur absolue de la différence d'images déplacées.

7.2.3 Modèle spatio-temporel d'attention

Avant de combiner les cartes statique et dynamique, il est nécessaire d'effectuer un filtrage temporel de chacune des cartes. En effet, les cartes sont calculées localement, sur une image (pour le modèle statique) ou sur deux images successives (pour le modèle dynamique), et les régions saillantes doivent être temporellement cohérentes à l'intérieur d'une fenêtre de durée L. La continuité temporelle de la vidéo empêche l'apparition de régions saillantes sur une ou deux images seulement (deux images correspondent à $2/25 = 0.08s$). C'est pourquoi un filtre médian de longueur L est effectué. Dans notre expérience, la taille de la fenêtre L est égale à cinq images.

Une étape de normalisation est nécessaire avant de réaliser la fusion des cartes. Elle permet d'éviter des problèmes d'échelle en limitant les valeurs entre zéro et un. Cette étape est effectuée de la façon suivante :

$$S_n = \begin{cases} S/T_h & \text{if } S < T_h \\ 1 & \text{if } S \geq T_h \end{cases} \tag{7.6}$$

où S est une carte de saillance statique ou dynamique et T_h un seuil préfixé (25 dans les deux cas).

Une fois les cartes normalisées, une étape de fusion est réalisée en combinant les cartes statique et dynamique pour obtenir une carte finale de saillance. La fusion est effectuée en utilisant l'opérateur « max » qui peut être interprété comme un « ou » logique. Ainsi, la carte finale contient des informations statique et dynamique. Finalement, la carte obtenue est une image en niveau de gris avec une valeur élevée pour les zones saillantes.

En utilisant diverses techniques de traitement d'images, nous détectons les régions saillantes. Les étapes suivantes sont réalisées : seuillage, opération morphologique (fermeture et ouverture), sélection des régions. Dans la dernière étape, nous déterminons les régions selon le voisinage en 4-connexité. Les régions avec une aire inférieure à un seuil sont supprimées. Finalement, les régions restantes sont sélectionnées et définies comme masques. Nous supposons également qu'un sujet humain ne peut pas se focaliser sur plus de cinq régions simultanément. Par conséquent, si le nombre de masques excède cinq, nous gardons seulement les cinq plus grands masques. La figure 7.7 illustre la fusion des cartes et la sélection des masques d'attention : les voitures, foule et joueurs, et les cyclistes.

FIG. 7.7 – Exemple de masques spatio-temporels d'attention.

7.3 Expérience psychophysique

Afin de tester la validité du modèle, des expériences oculométriques sont en général proposées [Itt00, Cha03, Meu05a]. Néanmoins, n'ayant pas à disposition un oculomètre (appareil qui mesure le déplacement de l'oeil), nous avons proposé une expérience psychophysique qui se décline en deux versions, la deuxième étant une version améliorée de la première. Le but de cette expérience est de savoir si les régions définies comme saillantes par le modèle sont effectivement saillantes. Nous avons donc essayé de confronter notre modèle aux jugements de sujets humains. L'expérience que nous avons établie est basée sur le paradigme de choix forcé à deux alternatives. Dans ce type de paradigme, le sujet doit indiquer entre deux stimuli présentés, lequel lui parait le mieux répondre à la tâche demandée. Dans la suite, la première version de l'expérience sera présentée. Afin de palier les limites de cette première version, nous exposerons ensuite la deuxième.

7.3.1 Expérience : version I

L'expérience consiste à présenter une vidéo suivie de l'affichage de deux images cibles : l'une contient les masques du modèle d'attention, l'autre est la même image mais avec des masques conçues de façon aléatoire. Le sujet doit sélectionner l'image qui lui semble le mieux représenter la vidéo.

7.3.1.1 Méthode

Participants

Seize sujets naïfs ont passé l'expérience. Tous les sujets avaient une vue normale ou corrigée.

Stimuli

Les sujets ont vu neuf vidéos, affichées au milieu de l'écran à une fréquence de 25 images

par seconde. Les images de la vidéo ont une taille de 288x352 pixels codés en niveaux de gris (256 niveaux). Pour chacune des vidéos, deux images sont sélectionnées aléatoirement (2 images parmi 30). Nous associons à chaque image tirée sa carte d'attention spatio-temporelle fournie par notre modèle. Afin de tester si les régions saillantes fournies par le modèle sont en accord avec la perception visuelle, nous sélectionnons à nouveau la même image et les mêmes masques mais nous les positionnons aléatoirement dans l'image comme montré sur la figure 7.8. Le principe est le suivant : le masque du modèle ayant l'aire la plus élevée est aléatoirement déplacé dans l'image ; le deuxième masque ayant l'aire la plus élevée est alors déplacé sans recouvrement possible avec le masque précédent et ainsi de suite... Finalement, les deux images (Fig. 7.8) ont les mêmes masques mais à des positions spatiales différentes.

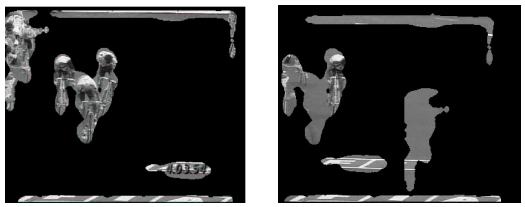


FIG. 7.8 – Exemple d'images cibles. L'image de gauche est la sortie du modèle. Celle de droite est la même image avec les mêmes masques calculés par le modèle mais placés à des positions aléatoires.

Procédure

L'expérience a été réalisée sur un ordinateur personnel. Les stimuli ont été présentés sur un écran 21" avec une résolution de 1024 par 768 pixels et une fréquence de rafraîchissement de 100 hertz. Les sujets sont placés approximativement à 50 centimètres de l'écran. La figure 7.9 décrit les événements pour un essai : un point de fixation apparaît (ici une petite croix noire) au milieu de l'écran pendant deux secondes, suivi d'une vidéo de 1.2 s toujours au milieu de l'écran. Puis, deux images sont présentées symétriquement au milieu de l'écran. Ces images appartiennent à la vidéo présentée et sont masquées de différentes façons : une suivant le modèle et l'autre avec les masques positionnés aléatoirement. Il est demandé aux sujets de choisir celle qui lui semble être la plus proche de la vidéo et de répondre le plus rapidement possible. L'image sélectionnée représentera le contenu de la vidéo. Finalement la réponse et le temps de réaction sont mesurés avec une boîte de réponse et le logiciel E-Prime.

Chaque vidéo apparaît quatre fois, avec deux images cibles différentes dans les deux positions possibles (droite-gauche et gauche-droite). Ainsi dans les deux cas, les mêmes images sont utilisées et dans un cas, l'image fournie par le modèle est affichée à droite de l'écran et dans l'autre cas, elle est affichée à gauche. Ceci nous permet d'avoir plus de réponses pour une condition et de vérifier que le sujet ne donne pas sa réponse aléatoirement. L'expérience est divisée en trois phases. Chaque phase contient trois vidéos et donc douze essais. Pendant une expérience, chaque sujet donne 36 réponses.

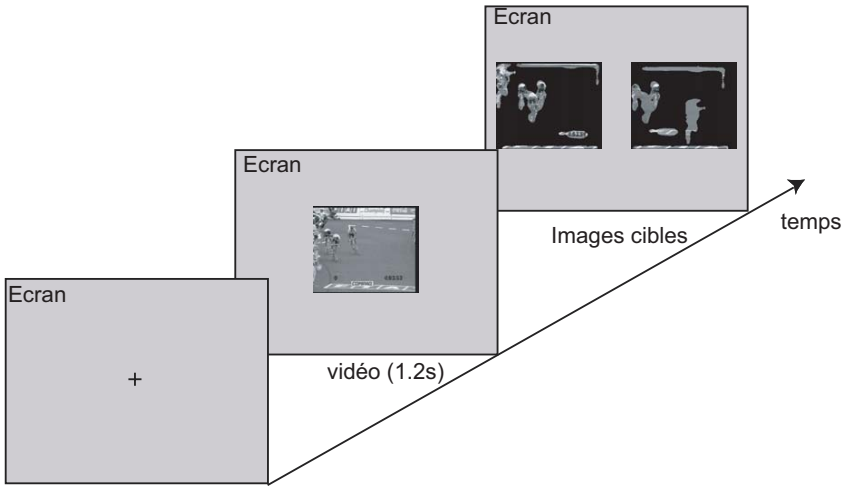


FIG. 7.9 – Déroulement d'un essai. D'abord, une vidéo apparaît pendant 1.2s, puis deux images sont présentées au milieu de l'écran. La tâche est de choisir l'image qui est la plus proche de la vidéo.

7.3.1.2 Résultats

Nous allons présenter les résultats concernant le pourcentage de réponses correctes par sujet. Une réponse est considérée correcte si le sujet a choisi le masque du modèle. En considérant les seize sujets, la moyenne des réponses correctes par sujet est de 84% avec un écart type de 11%. Comme attendu, pour tous les sujets, les masques du modèle, décrivent davantage la vidéo que les masques aléatoires.

Nous pouvons affiner les résultats. En effet, pour certains masques, il y a un chevauchement entre les masques du modèle et les masques aléatoires. La figure 7.10 montre une illustration du chevauchement entre deux masques. Ainsi nous devrions trouver moins de réponses correctes lorsque les masques du modèle se recouvrent avec les masques aléatoires que lorsque les masques sont séparés. Nous avons ajouté une condition : les masques ont soit plus de 50% de recouvrement soit moins de 50%. Les pourcentages de réponses correctes en fonction de cette condition sont présentés dans le tableau 7.1.

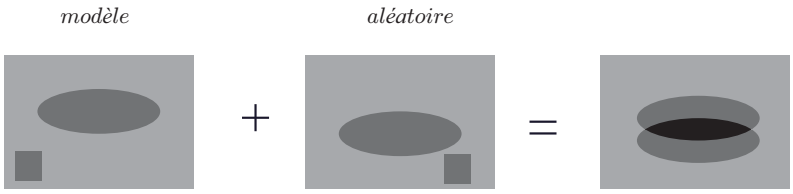


FIG. 7.10 – Illustration du recouvrement entre deux masques de stimuli. La zone noir correspond au chevauchement entre les images cibles.

Une analyse statistique a été réalisée sur le pourcentage de réponses correctes en fonction du recouvrement. Nous utilisons la « méthode des couples » qui consiste à associer, pour chaque sujet, la différence entre les pourcentages de réponses correctes suivant le recouvrement. Nous vérifions que la distribution des différences suit une loi normale (test de Lilliefors). Un test d'hypothèse est ensuite effectué sachant que la distribution des différences réduit par l'écart

type suit une loi de Student. L'hypothèse nulle est qu'il n'y a pas de différence des réponses moyennes suivant le recouvrement et l'hypothèse alternative est qu'il y a une différence significative. De plus amples explications sur ce test d'hypothèse peuvent être trouvées dans [Bai89]. Finalement, le test est significatif (au seuil de signification 5%) et donc le pourcentage des réponses correctes est plus faible quand les masques du modèle et les masques aléatoires se recouvrent, ce qui consolide le modèle.

TAB. 7.1 – Effet du chevauchement entre les images cibles (Expérience : version I).

	$> 50\%$	$< 50\%$
moyenne	0.77	0.87
écart type	0.17	0.12

Nous confirmons ainsi que notre modèle est cohérent avec la perception. Néanmoins, les masques aléatoires dépendent du modèle d'attention, puisqu'ils proviennent des masques du modèle qui ont simplement été positionnés aléatoirement. Or, la forme des masques peut être un critère de sélection des images pour les sujets. De plus, bien que la réponse soit demandée d'être la plus rapide possible, le sujet peut prendre le temps d'analyser en détail le contenu des images et de les comparer. Devant les limites de cette expérience, nous avons proposé une deuxième version.

7.3.2 Expérience : version II

Cette expérience est similaire à la précédente. La principale différence se situe dans la création des masques aléatoires où la forme et la position sont aléatoires.

7.3.2.1 Méthode

Participants

Trente nouveaux sujets ont passé l'expérience. Tous les sujets avaient une vue normale ou corrigée.

Stimuli

Les stimuli sont construits presque de la même manière que ceux de l'expérience I. Cependant, nous avons choisi des vidéos différentes de durée de 3 secondes. Pour chaque vidéo, deux images sont tirées aléatoirement (2 images parmi 75). Comme pour l'expérience I, à chaque image tirée, nous associons sa carte de saillance spatio-temporelle. En revanche, les masques aléatoires associés à chaque image sélectionnée sont créés différemment. Nous souhaitons concevoir des masques de forme aléatoire où leur surface est équivalente à celle des masques du modèle. La construction s'effectue de cette manière. L'aire la plus grande des masques du modèle est sélectionnée. Un germe est alors choisi (un germe est un pixel sélectionné), puis nous définissons les positions éventuelles du prochain germe en conservant la 4-connexité. Une de ces positions est alors sélectionnée aléatoirement, et nous recommençons jusqu'à obtenir l'aire souhaitée. Pour être dans les mêmes conditions entre les deux masques (aléatoire et du modèle), nous réalisons un filtrage par des opérations morphologiques (fermeture et ouverture). Si le filtrage engendre un masque avec une aire inférieure à celle voulue, nous lui réappliquons une sélection de germes pour augmenter sa surface jusqu'à atteindre la

valeur attendue, puis le filtrage est réeffectué. Par conséquent, le masque créé a une forme aléatoire et sa surface est supérieure ou égale à celle du masque du modèle. Finalement, il est positionné aléatoirement dans l'image. Ce procédé est répété pour le deuxième masque ayant l'aire la plus élevée et ainsi de suite.... La figure 7.11 montre un exemple d'images cibles. L'image de droite est l'image avec les masques aléatoires alors que celle de gauche est celle du modèle.

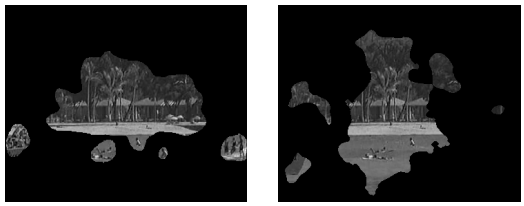


FIG. 7.11 – Exemple d'images cibles. L'image de gauche est la sortie du modèle. Celle de droite est la même image mais avec des masques de forme aléatoire.

Procédure

La procédure est identique à celle de l'expérience I. Un essai comporte toujours un point de fixation de deux secondes suivi d'une vidéo de trois secondes. Puis deux images sont présentées au milieu de l'écran. Néanmoins, pour éviter que les sujets analysent en détail les images, nous avons simplement limité l'affichage des deux images à deux secondes au maximum.

Comme nous avons choisi 27 vidéos à présenter aux sujets, chacune d'elles n'apparaîtra que deux fois, la position des images cibles à l'écran est choisie aléatoirement et deux images cibles différentes sont choisies pour chaque vidéo. L'expérience est ici effectuée en une seule phase (contrairement à celle de l'expérience I) et chaque sujet répond à 54 (27*2) essais.

7.3.2.2 Résultats

Similairement à l'expérience I, c'est le pourcentage de réponses correctes par sujet qui va nous intéresser. Pour les 30 sujets, la moyenne des réponses correctes par sujet est de 86% avec un écart type de 8%. Ainsi les masques du modèle qui représentent les zones saillantes de l'image sont largement choisis par les sujets. En ce qui concerne le recouvrement entre les masques, nous avons effectué à nouveau une analyse statistique et celui-ci a été à nouveau significatif (au seuil de signification 5%). Les pourcentages de réponses correctes en fonction du recouvrement (supérieur à 50% ou inférieur à 50%) sont présentés dans le tableau 7.2.

TAB. 7.2 – Effet du chevauchement entre les images cibles (Expérience : version II).

	$> 50\%$	$< 50\%$
moyenne	0.77	0.88
écart type	0.11	0.09

Nous avons vérifié s'il y a un effet de la surface des masques du modèle. La condition est la suivante : les masques du modèle sont soit inférieurs à 20% de la taille de l'image soit supérieurs à 20%. Les pourcentages de réponses correctes en fonction de cette condition sont présentés dans le tableau 7.3. L'analyse statistique a montré que le test est significatif et donc

le pourcentage des réponses correctes est plus élevé quand les masques du modèle sont petits (inférieurs à 20%).

TAB. 7.3 – Effet de la surface des masques du modèle (Expérience : version II).

	$> 20\%$	$< 20\%$
moyenne	0.79	0.89
écart type	0.10	0.10

En résumé, nous avons présenté un modèle spatio-temporel d'attention. Il repose sur la fusion d'un modèle statique inspiré du système humain avec un modèle de détection d'objets en mouvement. Nous avons également établi une expérience pour juger de l'efficacité du modèle. Les résultats obtenus sont bons avec une précision supérieure à 84% pour les deux versions de l'expérience. De plus, le modèle peut être utilisé dans beaucoup d'applications comme l'indexation, la surveillance, la compression... Dans la section suivante, nous l'allons employer pour concevoir une méthode de résumé de vidéo.

7.4 Création de résumé de vidéo à partir du modèle d'attention visuelle

Le modèle d'attention fournit des régions où le regard se porte durant le visionnage d'une vidéo, et renseigne sur la quantité d'informations contenues dans les images. Nous allons l'exploiter pour concevoir la méthode de résumé. Le modèle d'attention nous permet de passer des images de la vidéo à des cartes de saillance en niveaux de gris. Nous ne nous intéressons plus à l'image initiale mais aux cartes de saillance et à leur évolution au cours du temps. Afin d'étudier la meilleure manière de créer un résumé, nous avons distingué trois cartes de saillance pour le résumé : cartes statique, dynamique, spatio-temporelles. Suivant le contenu des plans, différentes stratégies pourront être appliquées sur ces cartes pour sélectionner les images clés. La figure 7.12 montre un exemple de cartes de saillance. La difficulté est le passage des cartes de saillance à la sélection d'images clés pour représenter le contenu de la vidéo.

7.4.1 Méthode de résumé de vidéo

Nous allons maintenant présenter la méthode de résumé de vidéo. Elle va consister à déterminer les changements au cours du temps des cartes de saillance afin de les sélectionner pour le résumé de vidéo. Cette méthode a été développée avec la participation de Sophie Marat [Mar06] dans le cadre d'un stage de Master. Après avoir exposé la méthode de résumé, nous utiliserons la même méthode d'évaluation que celle proposée dans le chapitre 5 concernant le résumé de vidéo à partir du mouvement de caméra (Section 5.3.2) pour mesurer sa qualité.

7.4.1.1 Passage de la carte de saillance à la courbe d'attention

Pour faciliter l'analyse des images, Ma et al. [Ma05a] ont proposé d'attribuer à chacune des cartes une valeur correspondant à la moyenne des niveaux de gris afin de définir une *courbe temporelle d'attention* à partir de ces valeurs. La figure 7.13 présente un plan issu de la vidéo « Documentaire » qui est l'interview d'un skieur par un présentateur avec de gros plans sur le

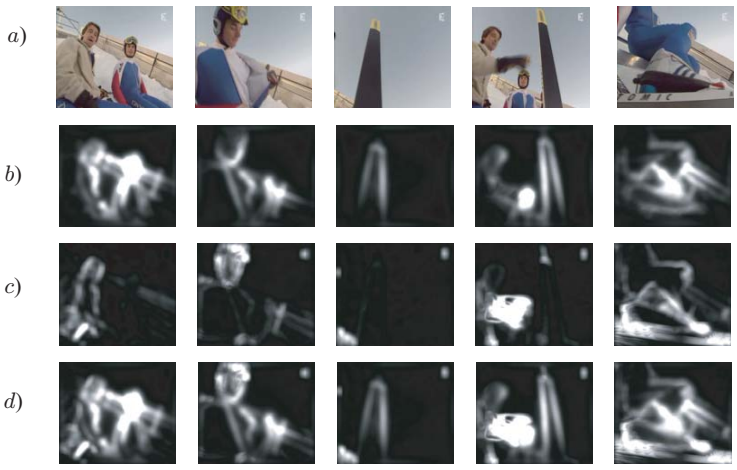


FIG. 7.12 – Exemple de cartes de saillance. a) images extraites de la vidéo « Documentaire », cartes de saillance : b) statiques, c) dynamiques, d) spatio-temporelles.

skieur et son équipement. Plusieurs phases dans cette vidéo peuvent être distinguées et quatre phases sont illustrées dans la figure 7.13 :

- les deux personnages sont presque immobiles
- zoom sur la tenue du skieur
- retour sur les deux personnages
- zoom sur un ski

Les différentes phases du plan peuvent effectivement se retrouver sur la courbe d'attention (Fig. 7.13.c). Par exemple la première phase est constituée par les 330 premières images. En effet, sur cette phase, la courbe ne présente pas de grandes variations car il n'y a pas beaucoup de changement de contenu dans cette partie du plan. Au contraire lorsqu'un gros plan est effectué sur le skieur puis sur son ski, on peut voir une variation significative des valeurs de la courbe d'attention statique.

Plus les images possèdent des zones saillantes et plus les valeurs sur la courbe d'attention sont élevées. Ma et al. [Ma05a] ont proposé de sélectionner les images clés en déterminant les maximums locaux sur la courbe d'attention afin de leur associer une image clé. Néanmoins, l'extraction des maximums ne permet pas de prendre en compte la nouveauté. Bien que les images fournies puissent contenir des informations intéressantes pour le résumé, des phases de la séquence peuvent être oubliées. Par exemple, si un objet vient à disparaître, l'image aura une carte de saillance avec un niveau de gris moins important donc elle ne pourra pas être retenue par l'extraction des maximums. Or la disparition de l'objet est aussi intéressante que son apparition. Dans la figure 7.13, nous pouvons voir que les cartes ayant des niveaux de gris plus faibles comme celles correspondant aux images 375 et 1000 ont un contenu moins riche que celles où les deux personnages sont présents (images 200 et 600). Les images 375 et 1000 sont des gros plans sur un personnage ou un objet et ont un réel intérêt pour le résumé en apportant des informations supplémentaires à celles données par les images avec les deux personnages. Elles doivent donc aussi apparaître dans le résumé.

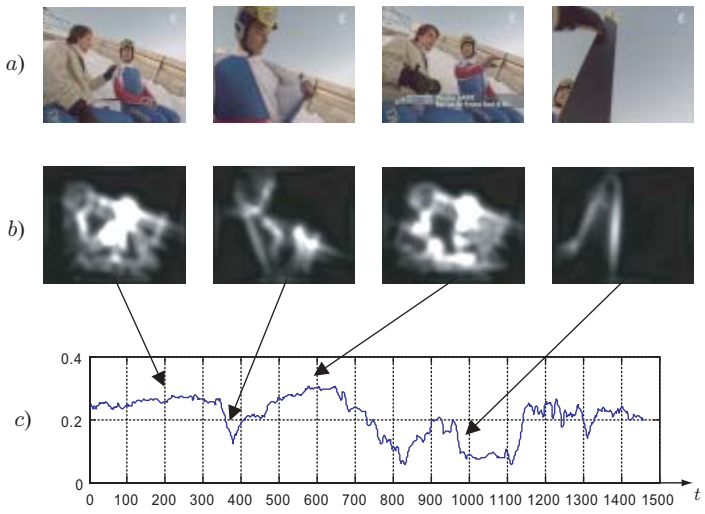


FIG. 7.13 – Courbe d'attention du premier plan de la vidéo « Documentaire ». a) images extraites de la vidéo, b) cartes de saillance statiques correspondantes, c) courbe d'attention statique (chaque point correspond à la valeur moyenne de la carte de saillance).

7.4.1.2 Résumé par détection des changements

Notre objectif est d'obtenir des courbes d'attention moins bruitées présentant des pics bien marqués indiquant les zones de changement dans les images du plan au cours du temps. Il s'agit de détecter les changements dans les cartes de saillance afin de révéler des événements importants. Nous définissons de nouvelles cartes de saillance qui s'éclaircissent sur les zones de changements et qui s'obscurcissent avec le temps lorsqu'aucun changement n'apparaît dans le plan. Les changements sont mis en évidence par une simple différence d'images :

$$|M_k - M_{k-i}| \tag{7.7}$$

où M_k est le masque (i.e. la carte de saillance) correspondant à l'image k et i est un paramètre qui définit l'écart entre le masque courant et le masque précédent. Si i est trop grand, les changements sont bien détectés mais les masques mettent du temps à s'obscurcir. Au contraire si i est petit, c'est le changement qui aura du mal à être détecté du fait de la trop grande proximité des images. Dans notre expérimentation, ce paramètre a été fixé à $i = 10$, ce qui nous a paru être un bon compromis entre l'assombrissement des zones sans changement et l'éclaircissement des zones comportant des changements.

Comme nous le voyons dans la figure 7.14, entre l'image 810 et l'image 820, la tête du skieur a disparu. Cela se remarque aussi sur les cartes de saillance avec détection de changement et donc sur la carte de saillance avec détection de changement de l'image 820. Il y a une zone blanche où se trouvait la tête du skieur.

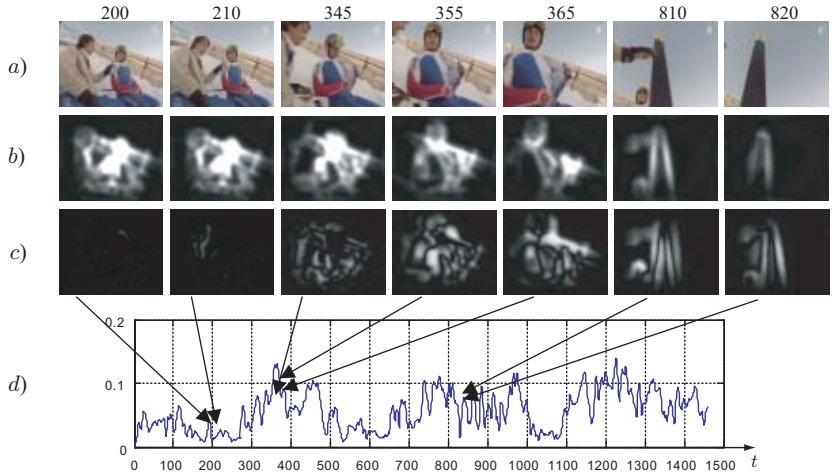


FIG. 7.14 – Exemple d'images et de cartes de saillance correspondant à un extrait du premier plan de la vidéo « Documentaire ». a) images extraites de la vidéo, b) cartes de saillance statiques, c) carte de saillance avec détection de changement, d) courbe d'attention avec détection de changement

Les nouveaux masques obtenus en appliquant la détection de changement vont donner à leur tour des courbes d'attention qui, comme nous le voyons sur la figure 7.14, ne dépendent plus du contenu des images mais uniquement des zones de changements. La première phase du plan allant jusqu'à l'image 330 ne présente pas beaucoup de changements, ce qui se traduit bien sur la courbe d'attention avec détection de changement par des valeurs faibles sur cette portion. Par contre, lors du gros plan sur le skieur, nous voyons que les valeurs sur la courbe d'attention ont considérablement augmenté. Nous avons donc bien atteint notre but en obtenant une courbe plus facile à exploiter où les événements importants correspondent aux pics.

7.4.1.3 Sélection des images clés

Le modèle de détection de changement nous a permis d'obtenir des nouvelles courbes d'attention où les pics représentent le changement et donc les zones intéressantes que nous cherchons à obtenir.

Pour extraire ces pics, nous calculons un seuil adaptatif, variable au cours du temps et les instants sur la courbe d'attention supérieurs au seuil sont choisis comme images clés (Fig. 7.15). Le seuil est donné par $\mu \pm 2.9\sigma$ où μ représente la moyenne glissante sur une fenêtre précédant l'image et σ l'écart type sur cette même fenêtre. Dans notre expérimentation, la taille de la fenêtre a été fixée de manière empirique à 50 (2 secondes de vidéo). Quant au paramètre 2.9, nous l'avons fait varier entre 3 et 2.7 et les résultats obtenus sont très proches pour toutes ces valeurs. Si on suppose la distribution de moyenne normale μ et d'écart type σ alors il est attendu que la prochaine valeur soit, à environ 99% de chance, dans cet intervalle $\mu \pm 2.9\sigma$. Dans le cas où plusieurs images consécutives (ou espacées de moins de 5 images) dépasseraient ce seuil, seule l'image ayant la plus grande valeur sur la courbe d'attention est retenue comme étant représentative du groupe. Ceci permet de ne pas retenir des images trop proches temporellement et donc proches du point de vue de leur contenu. Un exemple de résumé obtenu à partir de cette méthode est donné dans la figure 7.15. Ce résumé nous donne bien les différentes phases du plan : l'interview du skieur avec les deux personnages et les gros

plans sur les différentes parties de son équipement.

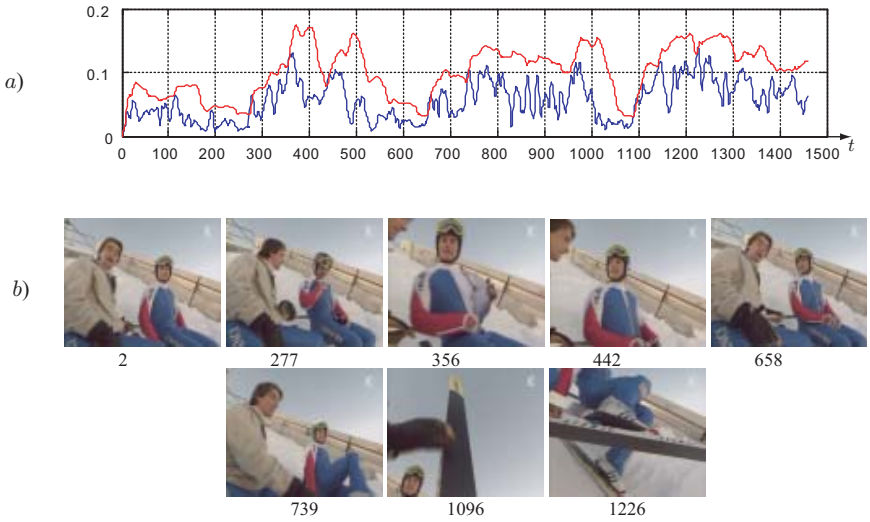


FIG. 7.15 – Résumé statique du premier plan de la vidéo « Documentaire ». a) courbe d'attention statique obtenue avec les cartes de saillance statiques avec détection de changement (courbe bleue) et seuil de décision (courbe rouge), b) images clés retenues pour constituer le résumé

Un cas particulier a été envisagé pour traiter les plans courts. Ils sont définis comme contenant au maximum 100 images dans le plan, ce qui correspond à 4s de vidéo. Durant un plan court, comme le contenu varie généralement peu, il n'est pas utile d'appliquer la méthode de détection de changement et une image suffit pour le résumé. Cette image est alors choisie en prenant le maximum de la courbe d'attention sans détection de changement.

7.4.1.4 Elimination des images redondantes

Afin de limiter la redondance des images clés du résumé, un post traitement permettant d'enlever les images trop semblables a été ajouté. Il s'agit de comparer l'image sélectionnée avec l'image précédemment sélectionnée en faisant une différence absolue entre les masques statiques correspondant (Fig. 7.16). Si cette différence est inférieure à un certain seuil, les images sont considérées comme trop ressemblantes et seule l'image correspondant au masque statique avec le niveau de gris moyen le plus élevé est retenue. Dans notre expérimentation, le seuil a été fixé à 0.11.

7.4.1.5 Choix de la carte de saillance

Nous avons maintenant une méthode d'extraction des images clés rapide et sans redondance. Pour obtenir un résumé optimal, nous avons testé deux approches : la première utilise les cartes de saillance statiques et dynamiques considérées séparément et la deuxième les cartes statiques et dynamiques combinées, c'est à dire les cartes spatio-temporelles.

Cartes statiques et dynamiques prises séparément

Après avoir réalisé la méthode de sélection des images suivant les cartes statiques et dynamiques sur une dizaine de plans, nous avons constaté que les résultats sont différents suivant

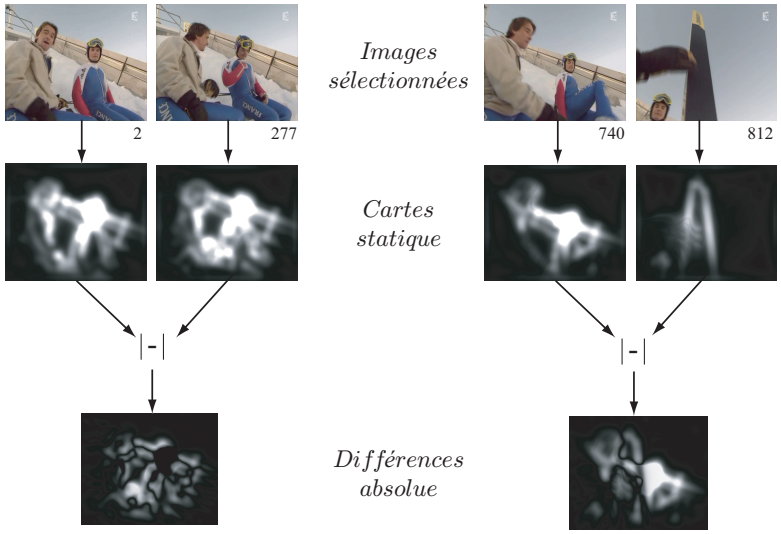


FIG. 7.16 – Application du post-traitement permettant d'enlever les images redondantes. A gauche, les masques se ressemblent et leur différence est inférieure à 0.11 alors que dans l'exemple de droite, les masques sont différents et leur différence est supérieure à 0.11.

la carte considérée. En fonction du contenu des plans (plutôt statique ou plutôt dynamique), il s'avère que les résultats sont meilleurs suivant la carte étudiée. Nous avons alors caractérisé les plans comme dynamique ou statique afin d'utiliser la carte de saillance statique (resp. dynamique) pour les plans statiques (resp. dynamiques). Pour cela, nous comparons les écarts types des courbes d'attention obtenues avec les cartes de saillance statiques et dynamiques sans détection de changement. La courbe donnant l'écart type le plus grand est choisie pour caractériser le plan et la sélection des images (Fig. 7.17) sera réalisée à partir des cartes correspondantes (statique ou dynamique avec détection de changement).

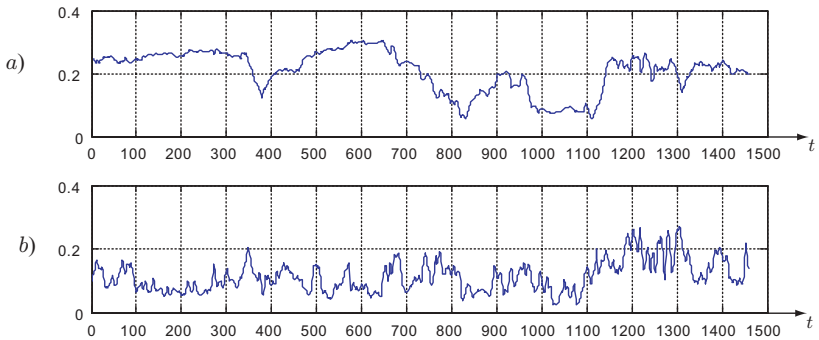


FIG. 7.17 – Courbe d'attention permettant de caractériser le premier plan de la vidéo « Documentaire ». a) courbe d'attention statique (écart type de 0.065), b) courbe d'attention dynamique (écart type de 0.049). Le plan est caractérisé par la courbe qui a l'écart type le plus grand, ici le plan est statique.

La figure 7.18 montre les résumés obtenus avec les cartes statiques et dynamiques d'un plan de la série « The Avengers ». Durant ce plan, nous voyons deux personnages dans un couloir qui essayent d'écouter aux différentes portes à l'aide d'un stéthoscope. Ce plan est caractérisé

comme statique, et c'est bien le résumé utilisant les cartes statiques qui donne le meilleur résultat. En effet, celui-ci donne des informations plus complètes (les deux personnages sont bien présents) que le résumé utilisant les cartes dynamiques et tout cela en retenant moins d'images.

FIG. 7.18 – Résumés d'un plan statique utilisant les cartes : a) statiques, b) dynamiques.

Dans la figure 7.19, nous pouvons voir les résumés d'un autre plan de la série « The Avengers ». Le plan est ici détecté comme dynamique et le résumé obtenu avec les cartes dynamiques donne le meilleur résultat. Dans ce plan, nous voyons deux voitures se suivre et croiser un camion sur une route étroite. Cette fois-ci, le résumé utilisant les cartes statiques ne retient qu'une image ne contenant qu'une voiture vue de loin, il ne permet donc pas de retrouver le contexte du plan alors que le résumé utilisant les cartes dynamiques va contenir deux images dont l'une montre bien les 3 véhicules.

FIG. 7.19 – Résumés d'un plan dynamique utilisant les cartes : a) statiques, b) dynamiques.

En ce qui concerne les plans courts (inférieurs à 4 secondes), comme en général le contenu varie peu, nous avons utilisé la courbe d'attention statique sans détection de changement pour déterminer l'image clé du plan court (en prenant le maximum de la courbe d'attention statique).

Finalement, nous avons obtenu une méthode permettant d'aboutir à des résumés prenant en compte la longueur du plan. Les plans longs sont caractérisés comme statiques ou dynamiques et les images clés sont sélectionnées en utilisant la courbe d'attention correspondant aux cartes de saillance avec détection de changement. Pour les plans courts nous utilisons systématiquement la courbe d'attention des cartes de saillance statiques sans détection de changement et une seule image est retenue.

Cartes spatio-temporelles

Pour éviter de caractériser le plan comme statique ou dynamique, nous utilisons directement les cartes spatio-temporelles qui regroupent les informations des cartes statiques et

dynamiques. Les cartes de saillance spatio-temporelles sont la fusion des cartes de saillance statiques et dynamiques par l'opérateur logique « ou ». Les cartes de saillance statiques donnent une information sur le contenu spatial des images de la vidéo alors que les cartes dynamiques détectent les objets en mouvement au cours du temps et donnent donc une information temporelle sur ces images. Le seuil adaptatif est défini de la même manière que précédemment. Dans cette approche (comme dans l'approche statique ou dynamique), la méthode de résumé sur les plans longs est effectuée sur les cartes de saillance avec détection de changement. Cette méthode donne des résultats aussi bons que ceux donnés par l'approche statique ou dynamique comme le montre la figure 7.20. En effet, le résumé (Fig. 7.20.a) utilisant les cartes spatio-temporelles est visuellement équivalent au résumé utilisant les cartes statiques et il est meilleur que le résumé utilisant les cartes dynamiques, car il montre bien les deux personnages. De même, le résumé (Fig. 7.20.b) par l'approche spatio-temporelle est semblable à celui obtenu en utilisant les cartes de saillance dynamiques et il est meilleur que le résumé utilisant les cartes de saillance statiques.

FIG. 7.20 – Résumé par l'approche spatio-temporelle (a) d'un plan détecté statique et (b) d'un plan détecté dynamique.

Pour les plans courts, il n'y a plus qu'une seule manière d'obtenir un résumé : en utilisant la courbe d'attention spatio-temporelle. Le résumé est alors obtenu en choisissant le maximum de la courbe d'attention spatio-temporelle sans détection de changement. L'approche spatio-temporelle donne aussi des résultats satisfaisants. De plus, elle est moins lourde en calcul car elle n'utilise qu'une carte de saillance, même si les cartes statiques et dynamiques doivent être générées au préalable.

Les résumés obtenus dans les deux cas (statique ou dynamique, spatio-temporelle) semblent convenables mais il nous reste à les évaluer objectivement avant de se prononcer quant à leur qualité.

7.4.2 Evaluation

Afin d'évaluer cette méthode de résumé, nous réalisons la même méthode d'évaluation proposée dans le chapitre sur le résumé de vidéo à partir du mouvement de caméra (Section 5.3.2). Elle consiste à comparer les résumés proposés par la méthode aux résumés établis par des sujets humains.

Les résultats sont présentés dans le tableau 7.4. Nous pouvons voir que les deux méthodes (résumé suivant les cartes statiques et dynamiques prises séparément et résumés suivant les cartes spatio-temporelles) donnent des résultats relativement similaires. Le rappel est toujours

supérieur ou égal à 50%, ce qui signifie que les méthodes retiennent au moins la moitié des images du résumé de référence. La précision est assez bonne, ce qui montre que les méthodes ne retiennent pas trop d'images par rapport à celles des sujets.

TAB. 7.4 – Résultats des méthodes de résumé pour les trois vidéos. Le seuil δ_s de regroupement entre deux images est fixé à 0.3. (R : Rappel, P : Précision, F_1)

Résumé	Cartes statiques et dynamiques prises séparément			Cartes spatio-temporelles		
	R (%)	P (%)	F_1 (%)	R (%)	P (%)	F_1 (%)
Documentaire	(15/24) 62.5	(15/29) 51.7	56.6	(14/24) 58.3	(14/27) 51.9	54.9
Journal	(43/55) 78.2	(43/61) 70.5	74.1	(40/55) 72.7	(40/56) 71.4	72.1
Série	(24/30) 80.0	(24/32) 75.0	77.4	(19/30) 63.3	(19/33) 57.5	60.3

Pourtant, les résultats sont ici moins bons que ceux présentés dans le chapitre sur le résumé à partir du mouvement de caméra. L'évaluation est fournie sur des vidéos ayant un script (scripted content) où les mouvements de caméra apportent une information pertinente pour le résumé. Les plans et les mouvements de caméra sont effectivement réfléchis par le réalisateur afin de transmettre le message.

Néanmoins, la méthode de résumé à partir de l'attention nous semble complémentaire de celle qui utilise le mouvement de caméra. Si on considère un plan long tourné à caméra fixe où de nombreux événements se produisent, alors la méthode de résumé qui exploite le mouvement de caméra ne sera pas efficace pour le résumé puisqu'une seule image au milieu du plan sera sélectionnée. La méthode de résumé à partir de la saillance détecte le changement, donc le résumé devrait être plus performant que la méthode basée sur le mouvement de caméra.

7.5 Conclusion

Nous avons présenté un modèle spatio-temporel d'attention ascendante qui consiste à extraire des zones saillantes dans les images à partir des caractéristiques physiques des stimuli (caractéristiques de bas niveau). Il repose sur la fusion d'un modèle statique inspiré du système visuel humain avec un modèle de détection des objets en mouvement. Une expérience déclinée en deux versions, a ensuite été conçue pour valider le modèle. Les résultats obtenus sont satisfaisants avec une précision supérieure à 84% pour les deux versions de l'expérience.

Nous avons ensuite proposé une méthode de résumé de vidéo qui s'appuie sur le modèle d'attention visuelle. Elle consiste à étudier les cartes de saillance au cours du temps et à détecter les variations temporelles de zones de saillance afin de les extraire pour le résumé. La méthode de résumé a été développée suivant deux variantes : une variante qui utilise les cartes statiques et dynamiques prises séparément, puis une variante qui exploite les cartes spatio-temporelles. Nous avons ainsi obtenu deux résumés pour chaque plan.

La méthode d'évaluation développée au chapitre 5 nous a permis de comparer les deux résumés de manière quantitative et objective. Les deux variantes de résumé basées sur le modèle d'attention visuelle, donnent des résultats relativement proches et correctes. Le modèle d'attention visuelle utilisé est donc performant pour la création automatique de résumé de vidéo.

Des améliorations peuvent cependant être apportées au modèle d'attention. Par exemple,

la carte de saillance statique ne prend pas en compte la couleur dans les images. D'autres caractéristiques peuvent également être introduites comme la détection des visages. En ce qui concerne le résumé de vidéo, une étude pourrait être menée afin de combiner la méthode de résumé utilisant le mouvement de caméra avec celle qui emploie le modèle d'attention. Ces deux méthodes paraissent complémentaires et une combinaison astucieuse devrait permettre la création de résumé performant sur une plus large gamme de vidéos (vidéos réalisées avec ou sans script).

Chapitre 8

Classification des vidéos

Nous proposons d'étudier une application du résumé, la classification des vidéos. A partir des images clés du résumé, un système d'annotation est développé pour classer les plans de la vidéo suivant différents concepts prédéfinis (par exemple, bateau, plage ou basketball...). Cette étude a été réalisée dans le cadre des expérimentations de TREC Video 2004.

Sommaire

8.1 Introduction

Devant l'augmentation des données audiovisuelles, de nouveaux systèmes d'aide à la navigation ont émergé. La nécessité d'organiser de manière efficace une grande base de données est apparue pour faciliter la recherche par le contenu afin de mieux satisfaire les demandes des utilisateurs. L'organisation des données est une tâche délicate qui peut être résolue en réalisant une classification automatique des vidéos. En effet, l'annotation des vidéos par mots clés est une tâche longue et fastidieuse qui s'accroît avec le volume de données. L'objectif de ce chapitre est donc de fournir un système d'annotation des plans suivant des concepts prédéfinis. Un concept correspond ici à une classe sémantique comme bateau, plage ou basketball.

Les récentes avancées dans l'analyse du contenu ont permis le développement de systèmes d'annotation de vidéo. Cependant, comme déjà précisé, la difficulté consiste à combler le fossé

entre les caractéristiques de bas niveau et les concepts sémantiques. Les techniques de classification des vidéos sont divisées en deux catégories : *(i)* des approches basées sur des règles qui exploitent une connaissance *a priori* sur un domaine particulier (comme par exemple des vidéos de sport) pour extraire des concepts [Shi03a, Zho00], *(ii)* des approches statistiques qui essayent de réaliser des annotations suivant une analyse indépendante du type de vidéos. Les méthodes statistiques exigent généralement un apprentissage pour classer automatiquement les scènes de la vidéo à partir de caractéristiques de bas niveau. Différents classifieurs peuvent être proposés, par exemple, les réseaux Bayésiens [Shi03a] ou le classifieur de type Machine à Vecteur de Support (SVM) [Nap04]. Dans ce dernier cas, différentes combinaisons de caractéristiques (seulement descripteur couleur, concaténation des descripteurs couleur et texture,...) sont considérées pour créer le modèle SVM et la combinaison qui donne les meilleurs résultats est choisie comme combinaison optimale. Souvannavong et al. [Sou05] proposent d'utiliser un algorithme génétique pour trouver la meilleure combinaison des sorties des classifieurs. Dans [AlA02], plusieurs caractéristiques sont extraites et chacune est employée pour entraîner un réseau de neurones artificiels (RNA). De ces classifieurs, une méthode de combinaison est réalisée en s'appuyant sur la théorie de Dempster-Shafer. Cependant cette combinaison demande à nouveau un apprentissage pour minimiser l'erreur quadratique moyenne entre la sortie combinée et la sortie voulue sur un ensemble d'apprentissage.

Dans la plupart des approches, les images sont représentées par des caractéristiques globales de bas niveau (histogramme couleur, descripteur de texture...). Néanmoins, certains auteurs décrivent localement les images. Par exemple, dans [Sou04], les images sont considérées comme un ensemble de régions qui contribuent individuellement au contenu sémantique des images. Leur méthode consiste à segmenter les images en régions et à regrouper les régions en types de régions. Ils construisent ensuite deux dictionnaires suivant les types de régions (l'un avec les caractéristiques couleur et l'autre avec les caractéristiques texture). Un document (plan) est alors décrit par le nombre d'occurrences de types de régions. Par une décomposition en valeur singulière de la matrice des occurrences, ils obtiennent un modèle LSI (Latent Semantic Indexing) qui permet de définir une similarité entre deux documents. Ce système est ensuite employé dans la tâche de classification de TREC Video 2003. Ce type d'approche est approprié pour rechercher des objets. Cependant, lorsque les concepts recherchés sont plage ou basketball, la description globale des images peut suffire pour les identifier.

Dans ce chapitre, nous allons décrire un système d'annotation sémantique de vidéos. Nous avons appliqué notre système en utilisant le protocole de TREC Video 2004 [Sme04]. L'intérêt de travailler avec la base de vidéos TREC réside dans la grande quantité de données (segmentation de vidéo de référence, vérité terrain pour l'extraction de concepts de haut-niveau...). De plus, le résumé de chaque plan des vidéos est également fourni. Ainsi une ou plusieurs images clés sont sélectionnées pour chaque plan et elles sont utilisées comme entrée de notre système d'annotation. L'architecture du système est illustrée sur la figure 8.1. Nous distinguons trois étapes : *capteurs*, *fusion de capteurs* et *fusion de concepts*. La première étape consiste à créer un ensemble de capteurs à partir de descripteurs de bas niveau (basés sur la couleur et la texture) et un apprentissage d'une Machine à Vecteurs de Support (SVM) pour permettre de prédire un concept donné (par exemple, bateau, plage ou basketball...). Le capteur permet ainsi à partir d'un apprentissage sur un descripteur donné de signaler si le concept est présent ou absent. La deuxième étape correspond à la combinaison des capteurs pour un concept donné. La troisième étape modélise l'interaction entre les concepts et permet de modifier leur prédiction. La méthode de fusion est réalisée en utilisant le Modèle des Croyances Transmissibles(MCT). Elle est appropriée pour modéliser l'incertitude des sources

et combiner les informations des différents capteurs.

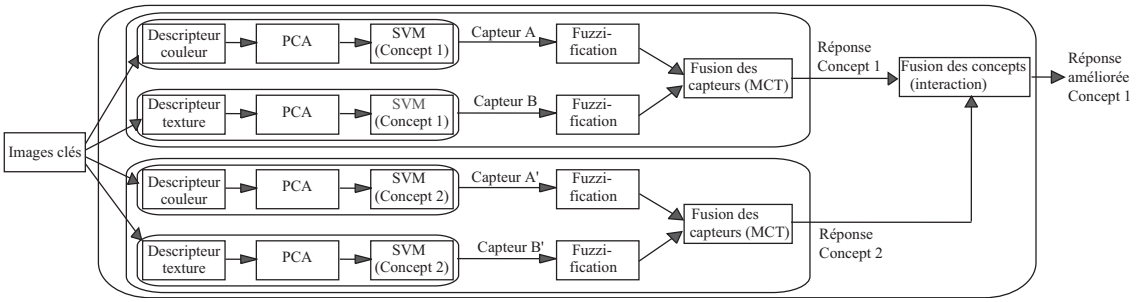


FIG. 8.1 – Architecture du système de classification.

Dans la suite de ce chapitre, nous présenterons les descripteurs, puis nous détaillerons la méthode d'extraction des concepts. Les résultats de la méthode seront ensuite exposés.

8.2 Descripteurs de bas niveau

Nous avons choisi d'étudier les caractéristiques de bas niveau suivantes : couleur et texture. Ces caractéristiques sont souvent utilisées pour la classification des images ou la recherche par l'exemple.

Parmi les descripteurs couleur, nous avons retenu l'histogramme couleur qui capture la distribution globale de couleur dans une image. L'histogramme couleur flou 3D développé dans la section 6.2.1.1 a été utilisé pour représenter les images et correspond à un vecteur caractéristique de 8x8x8 composantes.

La texture a aussi été très étudiée dans les tâches d'identification. Bien que beaucoup d'approches numériques aient été proposées pour représenter la texture de l'image, nous avons choisi de concevoir un descripteur inspiré de la perception humaine et adapté à la description du contenu des vidéos. Nous avons repris la partie statique du modèle d'attention, filtrage rétinien suivi d'un banc de filtres de Gabor (Section 7.2.1) pour créer le descripteur texture. Similairement aux travaux de N. Guyader [Guy01], nous avons considéré la sortie des filtres de Gabor selon 7 bandes de fréquences $f_m = \frac{0.3}{2^m}$ (avec $m = [0 \cdots 6]$) et 7 orientations $\theta_n = \frac{n\pi}{8}$ (avec $n = [0 \cdots 6]$). Pour chaque sortie de filtre de fréquence f_m et d'orientation θ_n, l'énergie totale $E(f_m, \theta_n)$ est obtenue en additionnant les carrés du module de sortie du filtre. La figure 8.2 montre le banc de filtres de Gabor utilisés dans notre expérimentation.

Une normalisation [Guy01] est alors appliquée pour être invariant à la netteté des images. Le flou est modélisé par une fonction $G(f)$ isotrope de la fréquence et la normalisation réalisée par bande de fréquence supprime ce terme.

$$E(f_k, \theta_i) = \frac{E(f_k, \theta_i)G(f_k)}{\sum_j E(f_k, \theta_j)G(f_k)} = \frac{E(f_k, \theta_i)}{\sum_j E(f_k, \theta_j)} \tag{8.1}$$

Chaque image clé est caractérisée par un vecteur caractéristique de texture à 49 (7x7) composantes où chaque composante correspond à l'énergie pour une orientation et une fréquence données. Ce type de descripteur peut également être retrouvé dans [Den02]. La description d'une image est réalisée en appliquant un banc de filtres de Gabor (4 orientations et 5 fréquences) et en considérant les énergies moyennes sur les images filtrées comme vecteur

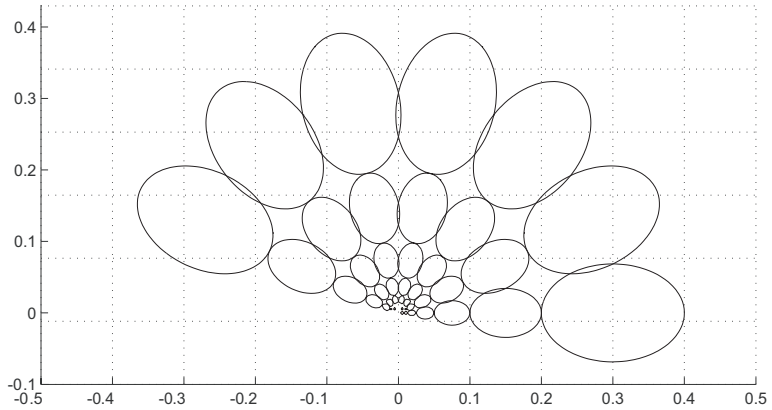


FIG. 8.2 – Banc de 49 filtres de Gabor

caractéristique de l'image. Puis la classification des images est effectuée à partir d'un réseau de fonctions à base radiale.

8.3 Méthode de classification des vidéos

Après avoir présenté les capteurs des différents concepts, une méthode de fusion est proposée dans le but de combiner la réponse des différents capteurs, puis de modéliser l'interaction entre les concepts afin d'améliorer la classification. .

8.3.1 Capteurs

Les capteurs sont créés à partir de descripteurs de bas niveau (couleur ou texture). Après une réduction de la dimension des descripteurs, un classifieur est utilisé pour identifier chaque concept. Une analyse en composante principale (ACP) a été exécutée pour réduire la dimension des caractéristiques. Les données TREC de 2004 se divisent en deux, une base de développement et une base de test. L'apprentissage est seulement appliqué sur les images clés de la base TREC de développement de 2004. Pour chaque descripteur, la dimension sélectionnée est choisie pour retenir au moins 99% de la variance. Dans notre expérimentation, nous passons de 512 à 64 composantes pour le descripteur couleur et de 49 à 32 pour le descripteur texture.

Comme le classifieur de type Machine à Vecteur de Support (SVM) a été utilisé avec succès dans de nombreuses tâches d'identification, nous l'avons employé pour apprendre les concepts de TREC Video 2004. Nous allons rapidement décrire le principe des SVM, une description plus détaillée pourra être trouvée dans [Sch98].

Soit $\{x_1 \cdots x_n\}$ un ensemble de données représentant des vecteurs caractéristiques d'images labelisées et leurs étiquettes $\{y_1 \cdots y_n\}$ où $y_i \in \{-1, 1\}$. Le problème consiste à approximer la fonction inconnue g suivante :

$$g(x) = \sum_{i=1}^{L} y_i \cdot w_i \cdot K(x_i, x) + b \tag{8.2}$$

où $K(\,,\,)$ est une fonction noyau (kernel function), x_i sont appelés vecteurs du support et sont déterminés à partir des données d'apprentissage, L est le nombre de vecteurs de support,

y_i est l'étiquette associée à chaque x_i, et w_i, b sont des constantes déterminées à partir de l'apprentissage. Dans cette étude, nous avons choisi la fonction à base radiale (Radial Basis Function RBF) gaussienne qui est couramment utilisée comme fonction noyau $K(\ ,\ )$.

Le classifieur SVM détermine l'hyperplan qui sépare les données d'apprentissage avec une marge maximale. La classification d'un nouveau vecteur x est donnée par le signe de la fonction de décision g. Néanmoins nous préférons attribuer une mesure de confiance sur le vecteur x à classer. La distance perpendiculaire de l'hyperplan au vecteur x est employée comme mesure de confiance. Nous appliquons le programme SVM_light développé par Thorsten Joachims [Sch99] avec les paramètres par défaut.

Une vérité terrain de chaque concept est donc déterminée sur la base de développement de 2004 et a permis de créer un modèle SVM pour chaque concept. Les classifieurs sont ensuite appliqués sur les images clés des données de test. La chose importante ici n'est pas le classifieur employé, mais la façon dont nous allons les combiner. Finalement, pour chaque concept, deux capteurs sont créés à partir des descripteurs couleur et texture.

8.3.2 Étape de fusion de capteurs

La méthode de fusion est basée sur le Modèle des Croyances Transférables. Cette théorie est adaptée pour combiner des informations imprécises et une description de celle-ci a été donnée à la section 4.4.

8.3.2.1 Définition de BBA

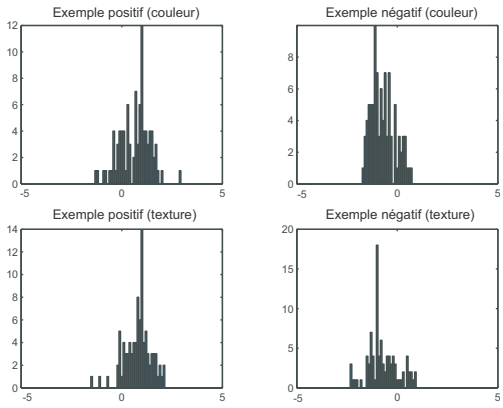
En l'absence de connaissance statistique sur les données, une mesure de confiance d'un concept est attribuée suivant la sortie $g(x)$ du SVM correspondant. Des fonctions d'appartenance sont définies et permettent d'attribuer des masses de croyance pour un concept donné en fonction de la sortie du SVM. La figure 8.3 montre la modélisation des fonctions de masses en fonction de la sortie $g(x)$ du SVM du concept « Plage ». La distribution de la sortie du SVM est montrée sur la figure 8.3.a. La valeur absolue de $g(x)$ correspond à la distance perpendiculaire entre l'hyperplan du modèle SVM et le vecteur caractéristique de l'image, et le signe révèle la classe à laquelle appartient l'image. Si $g(x)$ est positif alors l'image est associée à la classe « Plage » sinon $g(x)$ est négatif et l'image n'appartient pas à cette classe. Plus la distance est proche de zéro, plus le classifieur commet des erreurs et inversement. Nous avons alors défini les fonctions de masse (BBA) comme montré sur la figure 8.3.b.

Si la sortie du concept « Plage » est considérée, l'ensemble des hypothèses est $\Omega = \{H_1 = Plage, \overline{H_1} = \overline{Plage}\}$. La masse $m(H_1)$ représente la confiance qui est assignée au concept « plage », $m(H_1 \cup \overline{H_1})$ est le doute sur le concept « Plage » et $m(\overline{H_1})$ est la confiance pour que ce ne soit pas le concept « Plage ».

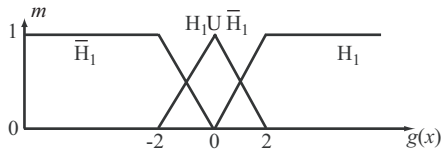
8.3.2.2 Fusion des capteurs

Le but de la fusion des capteurs est de modéliser la fusion de plusieurs capteurs sur le même concept. Chaque capteur réalise une observation et assigne sa confiance sur l'ensemble Ω. Dans notre expérimentation, deux capteurs sont définis à partir des descripteurs couleur et texture. Le tableau 8.1 illustre la combinaison des deux capteurs qui décrivent le même concept H_i.

Une masse peut être assignée à l'ensemble vide. Celui-ci est interprété comme un conflit entre les capteurs. Le conflit signifie qu'un des capteurs fait une erreur mais on ne sait pas



(a)



(b)

FIG. 8.3 – (a) Exemple de la distribution de la sortie SVM sur l'ensemble d'apprentissage du concept « Plage ». (b) Définition des BBA à partir de la sortie SVM.

TAB. 8.1 – Interaction de deux capteurs m_1 et m_2 pour le même concept.

		m_2		
		H_i	$H_i \cup \overline{H_i}$	$\overline{H_i}$
m_1	H_i	H_i	H_i	$H_i \cup \overline{H_i}$
	$H_i \cup \overline{H_i}$	H_i	$H_i \cup \overline{H_i}$	$\overline{H_i}$
	$\overline{H_i}$	$H_i \cup \overline{H_i}$	$\overline{H_i}$	$\overline{H_i}$

lequel. La masse de conflit $H_i \cap \overline{H_i} = \emptyset$ est alors transférée sur l'ignorance $H_i \cup \overline{H_i}$. Cela évite de prendre une décision. Néanmoins, la combinaison avec d'autres capteurs n'empêchera pas de déterminer si l'image appartient ou non à ce concept.

8.3.2.3 Décision

Nous employons le protocole TREC pour évaluer la classification. Il demande la soumission d'une liste ordonnée de plans par concept. Il s'agit de trier les plans par ordre décroissant, du plus plausible au moins plausible. Les règles de la décision sont effectuées sur les hypothèses singletons H_i. La masse $m(H_i)$ est choisie pour la décision finale et représente le degré de confiance placé exactement sur le concept H_i. Un plan peut contenir une ou plusieurs images clés et le passage des images clés au plan est réalisé de la façon suivante :

$$v_i = \max_{I_k \in S_j} (m(H_i)\{I_k\}) \tag{8.3}$$

où $m(H_i)\{I_k\}$ représente la masse associée à l'hypothèse H_i pour l'image I_k du plan S_j et, v_i est le degré de confiance du concept H_i placé sur le plan S_j.

8.3.3 Étape de fusion de concepts

L'étape de fusion de concepts modélise l'interaction entre les concepts. Le principe consiste à combiner un concept avec un autre concept ayant une bonne fiabilité pour améliorer la classification. La BBA sur un ensemble $\Omega_1 = \{H_1, \overline{H_1}\}$ peut également être combinée avec la BBA d'un autre ensemble $\Omega_2 = \{H_2, \overline{H_2}\}$ si une relation existe entre H_1 et H_2. Par exemple, si les hypothèses sont exclusives, le tableau 8.2 montre comment la combinaison est effectuée. Plusieurs stratégies peuvent être adoptées pour traiter l'ensemble vide. Comme vu précédemment, la masse est transférée sur l'union $H_1 \cup \overline{H_1}$.

Tab. 8.2 – Combinaison des masses m_1 d'un concept avec les masses m_2 d'un autre concept ayant une grande fiabilité.

		m_2		
		$H_2 = \overline{H_1}$	$H_2 \cup \overline{H_2} = H_1 \cup \overline{H_1}$	$\overline{H_2} = H_1 \cup \overline{H_1}$
m_1	H_1	$H_1 \cup \overline{H_1}$	H_1	H_1
	$H_1 \cup \overline{H_1}$	$\overline{H_1}$	$H_1 \cup \overline{H_1}$	$H_1 \cup \overline{H_1}$
	$\overline{H_1}$	$\overline{H_1}$	$\overline{H_1}$	$\overline{H_1}$

8.4 Expérimentation de TREC Video 2004

Notre but est de juger de l'efficacité de la méthode de fusion précédemment décrite. Nous présentons successivement les données TREC puis les résultats obtenus.

8.4.1 Données

Nous avons considéré la tâche d'extraction de caractéristique de haut niveau sur un ensemble d'images clés de la base TREC Video 2004. L'unité élémentaire dans le contexte de TREC est un plan qui peut contenir une ou plusieurs images clés. Ces données sont fournies

par l'Institut National des Normes et de la Technologie (NIST). Le base TREC de 2004 se divise en deux parties : une base de développement avec 254 vidéos comprenant 138823 images clés et une base de test avec 128 vidéos comprenant 48818 images clés. Cette collection de vidéos contient des journaux télévisés de CNN ou ABC, des publicités...

8.4.2 Résultats

L'évaluation de la liste ordonnée de plans est réalisée par la précision moyenne (Average Precision) et le nombre total de plans pertinents (Number of Relevevant Documents) retournés pour un concept donné. La précision moyenne est obtenue en calculant la moyenne de la précision obtenue après chaque plan pertinent retrouvé. Le tableau 8.3 illustre la méthode de fusion. Les résultats montrent que le nombre de plans trouvés en combinant les capteurs couleur et de texture (Fusion des capteurs) est supérieur ou égal au nombre de plans trouvés indépendamment par les capteurs. Évidemment, les résultats dépendent du type de descripteur utilisé, des données d'apprentissage et du classifieur employé. Si les capteurs pris séparément ne sont pas très efficaces, leur combinaison ne permettra pas de retrouver tous les plans. Le point important est que la combinaison des capteurs améliore le nombre de plans trouvés.

Tab. 8.3 – Résultats de la fusion des capteurs et des concepts. La première ligne correspond au nombre total de plans trouvés et la seconde est la précision moyenne.

Concept	Nombre total de plans pertinents	Capteur texture	Capteur couleur	Fusion des capteurs	Fusion des concepts
Bateau	441	62	120	120	120
Navire		0.0054	0.0185	0.0182	0.0191
Plage	374	84	139	139	145
		0.0143	0.0358	0.036	0.0381
Panier de	103	11	22	34	34
basket marqué		0.0006	0.0049	0.0065	0.0071
Avion qui	62	18	14	20	28
décolle		0.0045	0.002	0.0049	0.0327
Personnes	1695	153	167	171	191
qui marchent		0.009	0.0084	0.0081	0.0102
Violence	292	28	41	50	54
physique		0.0016	0.0024	0.0036	0.0048
Route	938	205	162	243	279
		0.0418	0.0128	0.0322	0.0429

Un concept étudié peut être mis en compétition avec un autre concept de meilleure qualité afin d'améliorer la précision. Cela permet de supprimer les fausses alarmes de la liste ordonnée des plans. Un nouveau concept est alors défini et contient des images mono-couleur et peu texturées parce que ce type d'images est souvent inséré entre différents sujets à la télévision (par exemple entre deux publicités). Les concepts interagissent avec le concept « Mono-couleur » et les résultats sont montrés dans le tableau 8.3 (Fusion de concept). Les résultats montrent que le nombre de documents trouvés et la précision moyenne augmentent.

Ce processus est itératif et la sortie de la fusion des concepts peut être combinée avec d'autres concepts. Par exemple, si un concept « Paysages naturels » est créé, nous pouvons le combiner avec le concept « Panier de Basket marqué » et non avec les autres concepts parce qu'ils ne sont pas exclusifs avec les paysages. Cette combinaison améliore encore les résultats

avec 37 plans pertinents trouvés contre 34 et une précision moyenne de 0.0117 contre 0.0065. La difficulté consiste à trouver de nouveaux concepts qui suppriment le plus possible les fausses alarmes.

8.5 Conclusion

Nous avons présenté une méthode d'extraction de concepts de haut niveau. Cette approche est divisée en trois étapes. D'abord, les capteurs sont créés pour chaque concept à partir des descripteurs couleur ou de texture, et de l'apprentissage d'un modèle SVM. Puis, la fusion de capteurs est réalisée pour chaque concept afin d'améliorer la classification. Finalement, la fusion de concepts modélise l'interaction entre les concepts. La méthode de fusion est basée sur le Modèle des Croyances Transférables. Cet outil est adapté pour modéliser et combiner des informations imprécises. Nous avons appliqué notre méthode dans le cadre des expérimentations de TREC Video 2004. Les résultats obtenus sur les données TREC Video 2004 démontrent l'amélioration fournie par une telle combinaison, comparée aux résultats des capteurs pris séparément. La méthode de fusion peut être appliquée avec d'autres capteurs (information sonore) et peut également être employée pour modéliser d'autres interactions entre les concepts.

Les résumés des différentes vidéos ont été obtenus par une annotation fournie par des sujets (TREC Video). Néanmoins, il serait intéressant d'appliquer cette méthode de classification des vidéos à partir de nos méthodes de résumé présentées dans les chapitres précédents. Une étude comparative de la classification pourrait ainsi être réalisée suivant le résumé considéré.

Chapitre 9

Conclusion et perspectives

9.1 Conclusion

Durant cette thèse, nous nous sommes intéressés à la création de résumé de vidéo. L'objectif du résumé est d'extraire les images (ou les extraits) les plus représentatives du contenu de la vidéo afin de donner un aperçu rapide à l'utilisateur. Le résumé de vidéo peut également être utilisé dans de nombreuses applications comme la classification, la recherche par l'exemple et la navigation dans une base de vidéos.

A partir des informations visuelles des images de la vidéo, nous avons proposé trois méthodes de résumé de vidéo qui reposent sur des caractéristiques différentes.

Nous avons d'abord créé une méthode de résumé hiérarchique à partir de caractéristiques de bas niveau. Trois nouveaux descripteurs compacts et flous ont été développés (couleur, orientation et mouvement) et leur combinaison (linéaire ou suivant un système d'inférence floue) a été particulièrement étudiée. Elle a permis de découper les vidéos en segments homogènes suivant un ou plusieurs descripteurs pour former le premier niveau du résumé. Un algorithme de regroupement par similarité avec contrainte temporelle a ensuite été élaboré pour construire les différents niveaux de résolution du résumé et s'adapter à la demande de l'utilisateur. Une application du résumé, la recherche par l'exemple, a montré l'intérêt de cette méthode de résumé et la combinaison des descripteurs a confirmé l'amélioration des résultats pour la recherche par l'exemple.

Nous avons ensuite conçu deux autres méthodes qui reposent sur des caractéristiques de plus haut niveau. La première s'appuie sur l'information apportée par le mouvement de caméra et la deuxième sur l'attention visuelle. Le mouvement de caméra nous a paru être une information pertinente pour décrire efficacement les vidéos. Par exemple, un zoom avant informe sur l'importance d'un passage de vidéo. Un système basé sur le Modèle des Croyances Transférables a été développé pour extraire et classer différents types de mouvements de caméra. Les résultats obtenus montrent la performance de la classification. A partir des différents mouvements extraits et en supposant les plans connus, nous avons construit une méthode de résumé suivant l'enchaînement et l'amplitude des mouvements selon des règles heuristiques.

Une méthode d'évaluation a alors été établie pour juger de la pertinence du résumé proposé. Une expérience psychophysique a permis à plusieurs sujets de créer leur propre résumé. Puis

une étude a été menée pour fusionner les résumés et obtenir le résumé de référence. Les comparaisons entre le résumé de référence et différents résumés proposés ont montré que la méthode de résumé suivant le mouvement de caméra donne de bons résultats.

Pour avoir une méthode entièrement automatique, nous avons également conçu un système de détection des changements de plan. Elle s'appuie sur la combinaison de descripteurs selon le Modèle des Croyances Transférables pour être insensible aux changements liés aux mouvements de caméra et des objets. Cette méthode a été évaluée et a fourni des résultats satisfaisants.

Ainsi les différents maillons de la chaîne ont été abordés : segmentation de vidéo en plans, identification des mouvements de caméra, et résumé de chacun des plans à partir des mouvements identifiés.

Nous avons enfin développé une méthode de résumé qui s'appuie sur un modèle d'attention visuelle. Effectivement, l'attention portée par un observateur lors du visionnage d'une vidéo fournit des informations susceptibles d'être efficaces pour créer le résumé.

Pour cela, nous avons proposé un modèle spatio-temporel d'attention. Il s'appuie sur la fusion d'un modèle statique inspiré du système humain et d'un modèle de détection d'objets en mouvement. Le modèle d'attention a ensuite été évalué par une expérience psychophysique pour être validé.

Nous l'avons ensuite utilisé pour détecter le changement de contenu au cours du temps afin de sélectionner des images clés. L'évaluation de cette méthode de résumé par l'expérience qui a fourni le résumé de référence a montré des résultats moins performants que la méthode de résumé créée à partir du mouvement de caméra. Cependant, la méthode d'évaluation repose sur des vidéos ayant un script (scripted video). Pour des vidéos sans script, cette méthode de résumé est certainement plus performante que celle utilisant le mouvement de caméra.

Nous avons également participé aux expérimentations de TREC Video 2004 et avons étudié plus particulièrement une application du résumé : la classification des vidéos. Dans le cadre de ces expérimentations, les images clés pour chaque plan sont fournies ainsi que la base de vidéos. Nous avons conçu un système d'annotation des concepts comme plage ou basketball. L'originalité du système repose sur la combinaison des sorties de différents classifieurs qui permet d'améliorer la classification.

9.2 Perspectives

La poursuite de ce travail de thèse à court terme va concerner la fusion des méthodes de résumé à partir du mouvement de caméra et du modèle d'attention visuelle. Les deux méthodes apparaissent complémentaires : l'une est efficace pour les vidéos avec script et l'autre pour les vidéos sans script. Leur combinaison permettra de traiter un plus large éventail de vidéos (vidéos avec ou sans script). Elle fournira une sélection d'images clés pour représenter les différents plans des vidéos. Pour réduire le nombre d'images clés et répondre aux demandes des utilisateurs, l'algorithme de regroupement par similarité avec contrainte temporelle de la première méthode de résumé pourrait également être appliqué afin de former un résumé hiérarchique.

De façon générale, l'influence des paramètres des méthodes de résumés pourrait être approfondie. Par exemple, pour la première méthode de résumé, d'autres descripteurs pourraient être étudiés comme ceux développés par le groupe MPEG7. Les différentes combinaisons abor-

dées dans la thèse reposent sur des transformations numériques-symboliques qui ont été fixées par notre expertise. Or celles-ci pourraient certainement être définies à partir d'une étude statistique sur la vidéo à traiter.

Une amélioration à apporter à nos méthodes pourrait être l'introduction d'une phase d'apprentissage. Les systèmes neuro-flous, qui permettent de combiner les systèmes flous et la capacité d'apprentissage des réseaux de neurones, pourraient être étudiés. Ce type de système permet d'apprendre les règles floues et les fonctions d'appartenance des variables d'entrées et de sorties en partant d'un ensemble représentatif. Pour cela, la base de vidéos devra être étendue pour permettre un apprentissage sur des données conséquentes. De même, la classification des mouvements de caméra, la détection des changements de plan, la classification des vidéos qui reposent sur le Modèle des Croyances Transférables pourraient demander une phase d'apprentissage pour déterminer les règles et les fonctions d'appartenance. Nous pourrions même effectuer une comparaison entre les modèles ayant un apprentissage et ceux établis par notre expertise.

Les trois méthodes de résumé fournissent une interactivité réduite avec l'utilisateur. Il serait intéressant d'introduire les préférences de l'utilisateur pour construire un résumé qui lui soit spécifique. Une investigation devra être menée pour concevoir ces interactions. Un processus de rétroaction pourrait aussi être envisagé afin d'affiner les résumés proposés. A partir d'un résumé automatique, le sujet pourrait intervenir en signalant les images qui l'intéressent. Un nouveau résumé pourrait alors lui être proposé.

Le modèle d'attention visuelle pourrait être complété en considérant d'autres caractéristiques comme la couleur, la détection de visages... Une première méthode basée sur le modèle d'attention a été étudiée par Sophie Marat [Mar06] pour simuler la scrutation d'un sujet humain lors d'un visionnage de vidéo.

Nous avons étudié des caractéristiques seulement visuelles. Cependant, d'autres caractéristiques pourraient être considérées comme la bande son ou le texte. Par exemple, la bande son apporte de nombreuses informations de haut niveau sur le contenu des vidéos. Une étude sur la combinaison des caractéristiques visuelles et sonores pourrait enrichir les méthodes de résumé et toutes les applications qui en découlent.

La technique d'évaluation que nous avons développée dans le chapitre 5 a permis de mesurer les performances des résumés au niveau des plans. Celle-ci peut également être employée pour traiter les résumés hiérarchiques. Mais la méthode de résumé hiérarchique (Chap. 3) n'a pas été testée par cette technique. Il serait intéressant, à plus long terme, de l'évaluer non seulement au niveau des plans mais également au niveau de la hiérarchie. Une étude comparative et détaillée entre les trois méthodes de résumé pourrait être menée. Il serait également souhaitable d'intégrer les méthodes de résumé proposées dans une chaîne complète de documentation pour se confronter à des problèmes de grande échelle.

Bibliographie

[Agh03] Z. Aghbari and A. Makinouchi. Semantic approach to image database classification and retrieval. *National Institute of Informatic (NII)*, 0(7), November 2003.

[AlA02] A. Al-Ani and M. Deriche. A new technique for combining multiple classifiers using the dempster-shafer theory of evidence. In *Artificial Intelligence*, volume 17, pages 333–361, 2002.

[AlH03] A.S. Al-Hammadi and A.M. Dawood. Intelligent technique for scene cut detection from mpeg video. In *Proceedings of International Conference on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS'03)*, pages 73–77, Lugano, Switzerland, July 2003.

[Ane04] A. Aner-Wolfa and J. R. Kenderb. Video summaries and cross-referencing through mosaic-based representation. *Journal of Computer Vision and Image Understanding*, 95(2) :201–237, August 2004.

[Ang02] J. Angulo and J. Serra. Morphological color size distributions for image classification and retrieval. In *Proceedings of Advanced Concepts for Intelligent Vision Systems Graphics (ACIVS'02)*, Ghent, Belgium, September 2002.

[Bai89] G. Baillargeon. *Probabilités statistiques et techniques de régression*. Edition SMG, 1989.

[Bea96] W. H. A. Beaudot. Sensory coding in the vertebrate retina : towards an adaptive control of visual sensitivity. *Journal Network : Computation in Neural Systems*, 7(2) :317–323, May 1996.

[Ben06] S. Benini, A. Bianchetti, and R. Leonardi. Using content analysis for video compression. In *Picture Coding Symposium*, Beijing, China, April 2006.

[Bes05] J. Bescos, G. Cisneros, J.M. Martinez, J.M. Menendez, and J. Cabrera. A unified model for techniques on video-shot transition detection. *IEEE Transactions on Multimedia*, 7(2) :293 – 307, August 2005.

[Blo03] I. Bloch. *Théorie des ensembles flous et des possiblités*. Fusion d'informations en traitement du signal et des images, Chap. 7, Traité IC2, Hermès, 2003.

[Bor00] J. Boreczky, A. Girgensohn, G. Golovchinsky, and S. Uchihashi. An interactive comic book presentation for exploring video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 185–192, The Hague, The Netherlands, April 2000.

[Bou99] P. Bouthemy, M. Gelgon, and F. Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Transactions Circuits and Systems Video Technology*, 9(7) :1033–1044, October 1999.

[Bru01a] E. Bruno. *De l'estimation locale à l'estimation globale du mouvement dans les images*. PhD thesis, Thèse de doctorat, Université Joseph Fourier Grenoble, LIS, 2001.

[Bru01b] E. Bruno and D. Pellerin. Global motion model based on b-spline wavelets : application to motion estimation and video indexing. In *2nd Int. Symposium. on Image and Signal Processing and Analysis (ISPA'01)*, pages 289–294, Pula, Croatia, June 2001.

[Cai04] C. Cai, K. M. Lam, and Z. Tan. An efficient scene break detection based on linear prediction. In *Proceedings of the Intelligent Multimedia, Video and Speech Processing (ISIMP'04)*, pages 555 – 558, Hong Kong, October 2004.

[Cal02] J. Calic. *New perspectives of video indexing and retrieval*. PhD thesis, M. Phil, Queen Mary, University of London, 2002.

[Cap02] A. S. Capelle, O. Colot, and C. Fernandez-Maloigne. Segmentation of multimodality MR images by means of evidence theory for 3d reconstruction of brain tumors. In *IEEE International Conference on Image Processing (ICIP'02)*, Rochester, New York, September 2002.

[Cap04] A. S. Capelle, C. Fernandez-Maloigne, and O. Colot. Segmentation of brain tumors by evidence theory : on the use of the conflict information. In *International Conference on Information Fusion*, pages 264–271, Stockholm, Sweden, June 2004.

[Cha02] A. Chauvin, J. Herault, C. Marendaz, and C. Peyrin. Natural scene perception : visual attractors and image processing. *Connectionist Models of Cognition and Perception, Proceedings of the Seventh Neural Computation and Psychology Workshop, World Scientific Press*, pages 236–245, 2002.

[Cha03] A. Chauvin. *Perception des scènes naturelles : étude et simulation du rôle de l'amplitude et de la saillance dans la catégorisation et l'exploration des scènes naturelles*. Thèse de Doctorat de l'Université Pierre Mendès-France (Grenobre), 2003.

[Che94] M. Cherfaoui and C.Bertin. Two-stage strategy for indexing and presenting video. In *Storage and Retrieval for Image and Video Databases II, Proc. SPIE 2185*, pages 174–184, San Jose, CA, USA, February 1994.

[Che98] F. Chevrie and F. Guély. La logique floue. Technical report, Cahier Technique n°191, Groupe Schneider, March 1998.

[Che03a] L.-Q. Chen, X. Xie, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou. Image adaptation based on attention model for small form factor devices. In *the 9th International Conference on Multimedia Modeling*, pages 340–343, Taipei, Taiwan, January 2003.

[Che03b] W.-G. Cheng and D. Xu. An approach to generating two-level video abstraction. In *IEEE International Conference on Machine Learning and Cybernetics*, volume 5, pages 2896–2900, Xi-an, China, November 2003.

[Che04] C.L.P Chen, C. Bhumireddy, and P. K. Darvemula. Camera motion classification using a genetic functional-link neural network. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS'04)*, volume 3, pages 2343 – 2348, Sandai, Japon, 28 Sept - 2 Oct 2004.

[Che05a] J.-C. Chen, J.-H. Yeh, W.-T. Chu, J.-H. Kuo, and J.-L. Wu. Improvement of commercial boundary detection using audiovisual features. In *6th Pacific-Rim Conference Multimedia (PCM'05)*, pages 776–786, Jeju Island, Corea, November 2005.

[Che05b] W.-H. Chend, W.-.T Chu, and J.-L. Wu. A visual attention based region-of-interest determination framework for video sequences. *IEICE Transactions on Information and Systems Journal*, E88-D(7) :1578–1586, October 2005.

[Cio05] G. Ciocca and R. Schettini. Dynamic key-frame extraction for video summarization. In *Proceedings of SPIE Internet Imaging VI*, pages 137–142, January 2005.

[Cio06] G. Ciocca and R. Schettini. An innovative algorithm for key frame extraction in video summarization. In *Journal of Real-Time Image Processing (in Print)*, 2006.

[Cor04] S. Corchs, G. Ciocca, and R. Schettini. Video summarization using a neurodynamical model of visual attention. In *IEEE 6th Workshop on Multimedia Signal Processing*, pages 71–74, Sienne, Italy, October 2004.

[Cou03] N. Courty, E. Marchand, and B. Arnaldi. A new application for saliency maps : Synthetic vision of autonomous actors. In *International Conference on Image Processing (ICIP'03)*, volume 3, pages 1065–1068, Barcelona, Spain, September 2003.

[Den02] N. Denquive and P. Tarroux. Multi-resolution codes for scene categorization. In *European Symposium on Artificial Neural Networks (ESANN'02)*, pages 281–287, Bruges, Belgium, April 2002.

[Dou00] A. D. Doulamis, N. Doulamis, and S. Kollas. Non-sequential video object representation using temporal variation of feature vectors. *IEEE Transactions on Consumer Electronics*, 46(3) :758–768, August 2000.

[Dua04] L.-Y. Duan, M. Xu, Q. Tian, and C.-S. Xu. Mean shift based nonparametric motion characterization. In *Proceedings of International Conference on Image Processing(ICIP'04)*, volume 3, pages 1597– 1600, Singapore, October 2004.

[Eki03] A. Ekin, A. M. Tekalp, and R. Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7) :796–807, August 2003.

[Far02] D. Farin, W. Effelsberg, and P. H. N. de With. Robust clustering-based video-summarization with integration of domain-knowledge. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'02)*, volume 1, pages 89–92, Lausanne, Switzerland, 2002.

[Fau04] B. Fauvet, P. Bouthemy, P. Gros, and F. Spindler. A geometrical key-frame selection method exploiting dominant motion estimation in video. In *3rd International Conference on Image and Video Retrieval*, pages 419–427, Dublin, Ireland, July 2004.

[Fer03] A. M. Ferman and A. M. Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Transactions on Multimedia*, 5(2) :244–256, June 2003.

[Gar00] U. Gargi, R. Kasturi, and S. H. Strayer. Performance charaterization of video-shot-change detection methods. *IEEE transactions on Circuits and Systems for Video Technology*, 10(1) :1–13, February 2000.

[Gil04] W. J. Gillespie and D. T. Nguyen. Robust estimation of camera motion in MPEG domain. In *Proceedings of Conference on Analog and Digital Techniques in Electrical Engineering (TENCON'04)*, volume 1, pages 395–398, Chiang Mai, Thailand, 21-24 nov 2004.

[Gir00] A. Girgensohn and J. Boreczky. Time-constrained keyframe selection technique. *Multimedia Tools and Applications*, 11(3) :347–358, August 2000.

[Gir03] A. Girgensohn. A fast layout algorithm for visual video summaries. In *IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages 77–80, Baltimore, Maryland, July 2003.

[Gir05] V. Girondel, L. Bonnaud, A. Caplier, and M. Rombaut. Static human body postures recognition in video sequences using the belief theory. In *IEEE International Conference on Image Processing (ICIP'05)*, volume 2, pages 45–48, Genoa, Italy, September 2005.

[Gon00] Y. Gong and X. Liu. Generating optimal video summaries. In *IEEE International Conference on Multimedia and Expo (ICME'00)*, volume 3, pages 1559–1562, New York, USA, July/August 2000.

[Gon01] Y. Gong and X. Liu. Video summarization with minimal visual content redundancies. In *IEEE International Conference on Image Processing (ICIP'01)*, volume 3, pages 362–365, Thessalonique, Grèce, October 2001.

[Gui05] H. Guillaume, N. Denquive, and P. Tarroux. Contextual priming for artifical visual perception. In *Europena Symposium on artifical neural networks*, pages 545–550, Bruges, Belgium, April 2005.

[Gun97] B. Gunsel, Y. Fu, and A. M. Tekalp. Hierarchical temporal video segmentation and content characterization. In *Proc SPIE Multimedia Storage and Archiving Systems II*, volume 3229, pages 46–56, 1997.

[Guy01] N. Guyader and J. Herault. Representation espace-frequence pour la categorisation d'images. In *Proc. GRETSI 2001*, Toulouse, France, September 2001.

[Guy02] N. Guyader, H. Le Borgne, J. Hérault, and A. Guérin-Dugué. Towards the introduction of human perception in a natural scene classification system. In *IEEE International workshop on Neural Network for Signal Processing (NNSP'02)*, pages 470–473, Martigny Valais, Switzerland, September 2002.

[Guy04] N. Guyader. *Scènes visuelles : Catégorisation basée sur des modèles de perception.* Thèse de Doctorat de l'Université Joseph Fourier (Grenobre), 2004.

[Hak97] Y. Hakoula. *Apprentissage des modèles linguistiques flous par de règles pondérés.* PhD thesis, Thèse de doctorat, Université de Savoie, LISTIC, 1997.

[Ham05] Z. Hammal, A. Caplier, and M. Rombaut. A fusion process based on belief theory for classification of facial basic emotions. In *Proc. Fusion'2005 the 8th International Conference on Information fusion (ISIF'05)*, Philadelphia, PA, USA, 2005.

[Ho03] C.-C. Ho, W.-H. Cheng, T.-J. Pan, and J.-L. Wu. A user-attention based focus detection framework and its application. In *Proceedings of the fourth International Conference on Information, Communications and Signal Processing and Fourth Pacific-Rim Conference on Multimedia (ICICS-PCM'2003)*, volume 3, pages 1315–1319, Singapour, December 2003.

[Hoa61] C. A. R. Hoare. Algorithm 63 (partition) ; 64 (quicksort) ; 65 (find). *Comm of ACM*, 4(7) :321–322, July 1961.

[Hua04] M. Huang, A. B. Mahajan, and D. Dementhon. Automatic performance evaluation for video summarization. Technical Report LAMP-TR-114, CAR-TR-998,CS-TR-4605,UMIACS-TR-2004-47, University of Maryland, College Park, June 2004.

[Itt99] L. Itti and C. Koch. Target detection using saliency-based attention. In *Proc. RTO/SCI-12 Workshop on Search and Target Acquisition (NATO Unclassified)*, pages 3.1–3.10, Utrecht, The Netherlands, Jun 1999.

[Itt00] L. Itti. *Models of bottom-up and top-town visual attention.* PhD Thesis, California Institute of technology, 2000.

[Jea01] S. Jeannin and A. Divakaran. Mpeg-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6) :720–724, June 2001.

[Kau02] A. Kaup, S. Treetasanatavorn, U. Rauschenbach, and J. Heuer. Video analysis for universal multimedia messaging. In *5th IEEE Southwest symposium on image analysis and interpolation*, pages 211–215, Sante Fe, USA, April 2002.

[Kim04] J.-G. Kim, H. S. Chang, J. Kim, , and H.-M. Kim. Threshold-based camera motion characterization of mpeg video. *ETRI Journal*, 26(3) :269–272, June 2004.

[Kir02] S. Kiranyaz, K. Caglar, B. Cramariuc, and M. Gabbouj. Unsupervised scene change detection techniques in feature domain via clustering and elimination. In *Proceedings of the IWDC 2002 Conference on Advanced Methods for Multimedia Signal Processing*, Capri, Italy, September 2002.

[Koc85] C. Koch and S. Ullman. Shifts in selective visual attention : Towards the underlying neural circuitry. *In Human Neurobiology, Springer-Verlag*, pages 219–227, 1985.

[Kop04] S. Kopf, T. Haenselmann, D. Farin, and W. Effelsberg. Automatic generation of video summaries for historical films. In *IEEE International Conference on Multimedia and Expo 2004 (ICME'04)*, volume 3, pages 2067–2070, Taipei, Taiwan, June 2004.

[Kos03] J. Kosecka, L. Zhou, P. Barber, and Z. Duric. Qualitative image based localization in indoors environments. In *Computer Vision and Pattern Recognition (CVPR'03)*, volume 2, pages 3–8, Madison, Wisconsin, June 2003.

[Laz02] M. Lazarescu, S. Venkatesh, and G. West. On the automatic indexing of cricket using camera motion parameters. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02)*, volume 1, pages 809–813, Lausanne, Switzerland, August 2002.

[Laz03] M. Lazarescu and S. Venkatesh. Using camera motion to identify different types of american football plays. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages 181–184, Baltimore, USA, July 2003.

[Lee02] S. Lee and M. Hayes. Real-time camera motion classification for content-based indexing and retrieval using templates. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'02)*, pages 3664–3667, Orlando, Florida, 13-17 May 2002.

[Li01] Y. Li, T. Zhang, and D. Tretter. An overview of video abstraction techniques. Technical report, HP Laboratory Technical Report, HPL-2001-191, July 2001.

[Li04] S. Li. *Content Analysis and Summarization for Video Documents.* PhD thesis, Research Associate, VIEW Lab, The Chinese University of Hong Kong, Department of Computer Science and Engineering, December 2004.

[Li05] Z. Li, G. M. Schuster, and A. K. Katsaggelos. Minmax optimal video summarization. *IEEE Transactions on circuits and systems for video video technology*, 15(10) :1245–1256, October 2005.

[Liu02] M. Liu and C. Wan. Content-based audio classification and retrieval using fuzzy logic system : towards multimedia search engines. *A fusion of Foundations, Methodologies and Applications*, 6(5) :357–364, August 2002.

[Lu04] S. Lu, M. R. Lyu, and I. King. Video summarization by spatial-temporal graph optimization. In *Proceedings of the International Symposium on Circuits and Systems (ISCAS'04)*, volume 2, pages 197–200, Vancouver, Canada, May 2004.

[Lu05] S. Lu, I. King, and M. R. Lyu. A novel video summarization framework for document preparation and archival applications. In *IEEE Aerospace Conference*, Big Sky, Montana, USA, March 2005.

[Ma02] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the 10th ACM Internaltional Conference on Multimedia (ACM'02)*, pages 533–542, Juan Les Pins, France, December 2002.

[Ma05a] Y. Ma, X. Hua, and H. Zhang. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 7(5) :907–919, October 2005.

[Ma05b] Y.-F. Ma and H.-J. Zhang. Video snapshot : A bird view of video sequence. In *Proceedings of the 11th International Multimedia Modelling Conference (MMM'05)*, pages 94–101, Melbourne, Australia, January 2005.

[Mac02] J. Machrouh. *Perception attentive et vision en intelligence artificelle*. Thèse de Doctorat, Université Paris-sud, December 2002.

[Mac05] J. Machrouh and P. Tarroux. Attentional mechanisms for interactive image exploration. *Journal of applied signal proccessing*, 14 :2391–2396, 2005.

[Mam75] E. H. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of Man-Machine Studies*, 7(1) :1–13, 1975.

[Man01] B. S. Manjunath, J. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *Transactions on Circuits and Systems for Video Technology*, 11(6) :703–715, June 2001.

[Mar06] S. Marat. Utilisation d'un modèle d'attention visuelle pour la création automatique de résumés et la scrutation de vidéos. Technical report, Rapport de Master de l'Université Joseph Fourier (Grenoble), 2006.

[Meg05] N. Megherbi, S. Ambellouis, O. Colot, and F. Cabestaing. Data association in multi-target tracking using belief theory : Handling target emergence and disappearance issue. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 517–521, Como, Italy, September 2005.

[Meu05a] O. Le Meur. *Analyse et caractérisation de séquence vidéo : mise en oeuvre dans le cadre de la compression vidéo*. Thèse de Doctorat, R&D France Télécom, site de Rennes, 2005.

[Meu05b] O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba. A spatio-temporal model of the selective human visual attention. In *Proceedings of IEEE International Conference on Image Processing (ICIP'05)*, volume 3, pages 1188–1191, September 2005.

[Mil92] M. Mills, J. Cohen, and Y. Y. Wong. A magnifier tool for video data. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 93–98, Monterey, California, USA, May 1992.

[Moc] The moca project, automatic movie content analysis. http ://www.informatik.uni-mannheim.de/pi4/projects/Moca.

[Nag91] A. Nagasaka and Y. Tanaka. Automatic video indexing and full-video search for object appearances. In *Visual Database Systems*, pages 113–127, Budapest, Hungary, September/October 1991.

[Nag02] K. Nagao, S. Ohira, and M. Yoneoka. Annotation-based multimedia summarization and translation. In *Proceedings of the 19th international conference on Computational linguistics*, volume 1, pages 1–7, Taipei, Taiwan, August/September 2002.

[Nam99] J. Nam and A. H. Tewfik. Video abstract of video. In *Proceedings of IEEE 3rd Workshop on Multimedia Signal Processing*, pages 117–122, Copenhagen, Denmark, September 1999.

[Nap04] M. R. Naphade. On supervision and statistical learning for semantic multimedia analysis. *Journal of Visual Communication and Image Representation*, 15(3) :348–369, July 2004.

[Nav02] V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *British Machine Vision Conference (BMCV'02)*, pages 453–461, 2002.

[Ngo03] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Automatic video summarization by graph modeling. In *Proceedings of 9th International Conference on Computer Vision (ICCV'03)*, volume 1, pages 104–109, Nice, France, October 2003.

[Odo95] J. M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4) :348–365, December 1995.

[Odo03] J.-M. Odobez, D. Gatica-Perez, and M. Guillemot. Video shot clustering using spectral methods. In *Third International Workshop on Content-Based Multimedia Indexing*, Rennes, France, September 2003.

[Oh04] J. H. Oh, Q. Wen, S. Hwang, and J. Lee. *Video Abstraction*. Video Data Management and Information Retrieval (A book edited by Sagarmay Deb). Idea Group Inc. and IRM, Press 2004, 2004.

[Ohb04] R. Ohbuchi and C.-S. Hung. Combining multiresolution shape descriptors for 3D model retrieval. In *Proceedings of 14th International Conference in Central Europe on Computer Graphics, Vizualization and Computer Vision (WSCG'06)*, Plzen, Czech Republic, January/February 2004.

[Oli03] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson. Top-down control of visual attention in object detection. In *Proceedings of IEEE International Conference on Image Processing (ICIP'03)*, volume 1, pages 253–256, 2003.

[Par04] V. Parshin and L. Chen. Video summarization based on user-defined constraints and preferences. In *Recherche d'information assistée par ordinateur RIAO*, Avignon, France, April 2004.

[Pek03] K.A. Peker and A. Divakaran. An extended framework for adaptive playback-based video summarization. In *SPIE Internet Multimedia Management Systems IV*, pages 26–33, Orlando, USA, September 2003.

[Pfe96] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg. Abstracting digital movies automatically. *Journal of Visual Communication and Image Representation*, 7(4) :345–353, December 1996.

[Pon99] D. Ponceleon, A. Amir, , S. Srinivasan, T. Syeda-Mahmood, and D. Petkovic. Cuevideo : Automated multimedia indexing and retrieval. In *Proceedings of the 7th ACM international conference on Multimedia*, volume 2, Orlando, Florida, US, October/November 1999.

[Por01] S. Porter. Detection and classification of shot transitions. In *Proceedings of the 12th British Machine Vision Conference(BMVC'01)*, pages 73–82, Manchester, England, September 2001.

[Por03] S. V. Porter, M. Mirmehdi, and B. T. Thomas. A shortest path representation for video summarisation. In *12th International Conference on Image Analysis and Processing (ICIAP'03)*, pages 460–465, Mantova, Italy, September 2003.

[Qi03a] Y. Qi, A. Hauptmann, and T. Liu. Sports video categorizing method using camera motion parameters. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages 689–692, Baltimore, USA, 6-9 July 2003.

[Qi03b] Y. Qi, A. Hauptmann, and T. Liu. Supervised classification for video segmentation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages 689–692, Baltimore, USA, July 2003.

[Ram06] E. Ramasso, M. Rombaut, and Denis Pellerin. A temporal belief filter improving human action recognition in videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, volume 2, pages 141–144, Toulouse, France, May 2006.

[Rav05] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg. Dynamosaics : Video mosaics with non-chronological time. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 58–65, San Diego, California, USA, June 2005.

[Ron04] J. Rong, Y.-F. Ma, and L. Wu. Gradual transition detection using em curve fitting. In *Proceedings of International Conference on Multimedia Modelling(MMM'05)*, pages 364 – 369, Melbourne, Australia, January 2004.

[Rou03] G. A. Rousselet and M. Fabre-Thorpe. *Les mécanismes de l'attention.* L'attention : aspects théoriques, Psychologie française, 2003.

[Rui98] Y. Rui, T. S. Huang, and S. Mehrotra. Exploring video structure beyond the shots. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, pages 237–240, Austin, Texas, USA, June/July 1998.

[Rui99] Y. Rui, T. S. Huang, and S. Mehrotra. Constructing table-of-content for videos. *Multimedia systems*, 7(5) :359–368, September 1999.

[Rui00] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the 8th ACM international conference on Multimedia*, pages 105–115, Marina del Rey, California, US, October/November 2000.

[Rui04] Y. Rui, Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang. A unified framework for video summarization, browsing and retrieval. Technical report, MERL Technical Report, September 2004.

[Sae04] E. Saez, J.I. Benavides, and N. Guil. Combining luminance and edge based metrics for robust temporal video segmentation. In *Proceedings of International Conference on Image Processing(ICIP'04)*, volume 4, pages 2231–2234, Singapore, October 2004.

[San99] S. Santini and R. Jain. Similarity mesures. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 21(9) :871–883, 1999.

[Sch98] B. Schölkopf. Svms - a practical consequence of learning theory. *Journal IEEE Intelligent Systems*, 13(4) :18–21, 1998.

[Sch99] T. Joachims. Advances in kernel methods - support vector learning. *Chapter Making large-scale svm learning practical, MIT Press*, 1999.

[Sha76] G. Shafer. A mathematical theory of evidence. *Princeton University Press, Princeton*, 1976.

[Sha04] X. Shao, C. S. Xa, and M. S. Kankanhalli. A new approach to automatic music video summarization. In *IEEE International Conference on Image Processing (ICIP'04)*, volume 1, pages 625–628, October 2004.

[Shi03a] H. Shih and C.Huang. Semantic network modeling for understanding baseball video. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 5, pages 820–823, Hong-Kong, April 2003.

[Shi03b] F. Shipman, A. Girgensohn, and L. Wilcox. Generation of interactive multi-level video summaries. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 392–401, Berkeley, California, USA, November 2003.

[Sme94] Ph. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2) :191–234, December 1994.

[Sme04] A Smeaton, W. Kraaij, and P. Over. Trecvid 2004 an introduction. In *13th Text Retrieval Conference*, USA, 2004.

[Sou04] F. Souvannavong, B. Mérialdo, and B Huet. Latent semantic indexing for semantic content detection of video shots. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'04)*, pages 1783–1786, Taipei, Taiwan, June 2004.

[Sou05] F. Souvannavong, B. Merialdo, and B. Huet. Multi-modal classifier fusion for video shot content retrieval. In *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, April 2005.

[Sun00a] X. Sun and M. S. Kankanhalli. Video summarization using r-sequences. *Real-Time Imaging*, 6(6) :449–459, December 2000.

[Sun00b] H. Sundaram and S.-F. Chang. Determining computable scenes in films and their structures using audio-visual memory models. In *Proceedings of the 8th ACM international conference on Multimedia*, pages 95–104, Marina del Rey, California, US, October/November 2000.

[Sun03] H. Sundaram and S.-F. Chang. *Video Analysis and Summarization at Structural and Semantic Levels*. Book chapter, Multimedia information retrieval and management : Technological Fundamentals and Applications, March 2003.

[Sur02] S. Sural, G. Qian, and S. Pramanik. A histogram with perceptually smooth color transition for image retrieval. In *Fourth International Conference on Computer Vision, Pattern Recognition and Image Processing*, pages 664–667, Durham, North Carolina, March 2002.

[Tak03] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga. Sports video categorizing method using camera motion parameters. In *Visual Communications and Image Processing (VCIP'03)*, volume 5150, pages 2082–2088, Lugano, Switzerland, 2003.

[Tan95] Y. Taniguchi, A. Akutsu, Y. Tonomura, and H. Hamada. An intuitive and efficient access interface to real-time incoming video based on automatic indexing. In *Proceedings of the 3rd ACM international conference on Multimedia*, pages 25–33, San Francisco, California, USA, November 1995.

[Tan97] Y. Taniguchi, A. Akutsu, and Y. Tonomura. Panoramaexcerpts : Extracting and packing panoramas for video browsing. In *Proceedings of the 5th ACM Internaltional Conference on Multimedia (ACM'97)*, pages 427–436, Seattle, WA, USA, November 1997.

[Tan00] Y.-P. Tan, D.D. Saur, S.R. Kulkarni, and P.J. Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions Circuits Systems and Video Technology*, 10(1) :133–146, February 2000.

[Tar05] G. Tardini, C. Grana, and R. Cucchiara. Shot detection and motion analysis for automatic mpeg-7 annotation of sports videos. In *13th International Conference on Image Analysis and Processing (ICIAP'05)*, pages 131–140, Cagliari, Italy, September 2005.

[Tok00] C. Toklu, S.-P. Liou, and M. Das. Video abstract : A hybrid approach to generate semantically meaningful video summaries. In *IEEE International Conference on Multimedia and Expo (ICME'00)*, volume 3, pages 1333–1336, New York, USA, July/August 2000.

[Ton93] Y. Tonomura, A. Akustsu, K. Otsuji, and T. Sadakata. Videomap and video spaceicon : Tools for anatomizing video content. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 131–136, April 1993.

[Tor03] A. Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2) :169–191, 2003.

[Tre05] The nist trec video retrieval evaluation. http ://www-nlpir.nist.gov/projects /tv2005/tv2005.html, 2005.

[Uch99] S. Uchihashi, J. Foote, A. Girgensohn, , and J. Boreczky. Video manga : Generating semantically meaningful video summaries. In *Proceedings of the 7th ACM international conference on Multimedia*, pages 383–392, Orlando, Florida, USA, October/November 1999.

[Ued91] H. Ueda, T. Miyatake, and S. Yoshizawa. Impact : An interactive natural-motion-picture dedicated multimedia authoring system. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 343–350, April/May 1991.

[Van99a] P. Vannoorenberghe, O. Colot, and D. Brucq. Color image segmentation using dempster-shafer's theory. In *IEEE International Conference on Image Processing (ICIP'99)*, volume 4, pages 300–303, 1999.

[Van99b] P. Vannoorenberghe, O. Colot, and D. Brucq. Dempster-shafer's theory as an aid to color information processing application to melanoma detection in dermatology. In *International Conference on Image Analysis and Processing (ICIAP'99)*, pages 774–779, 1999.

[Ven02] E. Veneau. *Macro-segmentation multi-critère et classification de séquences par le contenu dynamique pour l'indexation vidéo*. Thèse de doctorat, Université de Rennes I, IRISA, February 2002.

[Wan04] Y. M. Wang and H. Zhang. Detecting image orientation based on low-level visual content. *Computer Vision and Image Understanding (CVIU'04)*, 93(3) :328–346, March 2004.

[Whi03] A. Whitehead, J. Bose, and R. Laganière. Feature based cut detection with automatic threshold selection. In *Proceedings of the IEEE International Conf. on Image and Video Retrieval (CIVR'04)*, pages 410–418, Dublin, Ireland, July 2003.

[Wol96] W. Wolf. Key frame selection by motion analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'96)*, volume 2, pages 1228–1231, Atlanta, Georgia, May 1996.

[Wu04] P. Wu. A semi-automatic approach to detect highlights for home video annotation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Montreal, Canada, May 2004.

[Yac05] B. Yacine. *Détermination des conditions d'ionisation caractérisant le seuil de claquage de l'air par la logique floue*. PhD thesis, Magister en Electronique, Option Matériaux Electrotechniques, Université de Batna, 2005.

[Yah01a] I. Yahiaoui, B. Mérialdo, and B. Huet. Automatic video summarization. In *Multimedia Content-based Indexing and Retrieval (MMCBIR'O1)*, Rocquencourt, France, September 2001.

[Yah01b] I. Yahiaoui, B. Mérialdo, and B. Huet. Generating summaries of multi-episode video. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'01)*, pages 611–614, August 2001.

[Yan03] H. Yang, L. Chaisorn, Y. Zhao, S. Y. Neo, and T. S. Chua. VideoQA : Question answering on news video. In *Proceedings of the 2003 ACM Conference on Multimedia*, pages 623–641, Berkeley, CA, USA, November 2003.

[Yeu96] M. M. Yeung, B.-L. Yeo, and B. Liu. Extracting story units from long programs for video browsing and navigation. In *Proceedings of the 1996 International Conference on Multimedia Computing and Systems*, pages 296–305, Hiroshima, Japan, June 1996.

[You03] S. Youm and W. Kim. Dynamic threshold method for scene change detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages 337–340, Baltimore, USA, 6-9 July 2003.

[Yu03] J. C. S. Yu, M. S. Kankanhalli, and P. Mulhem. Semantic video summarization in compressed domain mpeg video. In *IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 3, pages 329–332, Baltimore, Maryland, July 2003.

[Yu04] X.-D. Yu, L. Wang, Q. Tian, and P. Xue. Multi-level video representation with application to keyframe extraction. In *Proceedings of the 10th International Multimedia Modelling Conference*, pages 117–123, Brisbane, Australia, January 2004.

[Zad65] L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3) :338–353, June 1965.

[Zha95] S. W. Smoliar D. Zhong H. J. Zhang, C. Y. Low. Video parsing, retrieval and browsing : An integrated and content-based solution. In *Proceedings of the 3rd ACM*

International Conference on Multimedia, pages 15–24, San Francisco, California, USA, November 1995.

[Zha00] L. Zhao, W. Qi, S. Z. Li, S.-Q. Yang, and H. J. Zhang. Key-frame extraction and shot retrieval using nearest feature line (nfl). In *Proceedings of the ACM Workshops on Multimedia*, pages 217–220, Los Angeles, California, USA, October/November 2000.

[Zha01] D. Zhang and D. Ellis. Detecting sound events in basketball video archive. Technical report, Tecnical Report, Dept. of Electrical Eng., Columbia University, 2001.

[Zho96] D. Zhong, H. J. Zhang, and S.-F. Chang. Clustering methods for video browsing and annotation. In *SPIE Storage and Retrieval for Image and Video Databases IV*, pages 239–246, 1996.

[Zho00] W. Zhou, A. Vellaikal, and C. C. Jay Kuo. Rule-based video classification system for basketball video indexing. In *Proceedings of the 13th ACM International Conference on Multimedia*, pages 213–216, San Franscico, California, USA, June 2000.

[Zho05] J. Zhou and X-P Zhang. Video shot boundary detection using independent component analysis. In *Proceedings of of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'05)*, pages 541–544, Philadelphia, USA, March 2005.

[Zhu98] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *IEEE International Conference on Image Processing (ICIP'98)*, pages 886–870, Chicago,USA, October 1998.

[Zhu03a] X. Zhu, W. G. Aref, J. Fan, A. C. Catlin, and A. K. Elmagarmid. Medical video mining for efficient database indexing, management and access. In *Proceedings of the 19th International Conference on Data Engineering (ICDE'03)*, pages 569–580, Bangalore, India, March 2003.

[Zhu03b] X. Zhu, J. Fan, A. K. Elmagarmid, and X. Wu. Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Systems*, 9(1) :31–53, July 2003.

[Zhu03c] X. Zhu and X. Wu. Sequential association mining for video summarization. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'03)*, volume 3, pages 33–336, Baltimore, Maryland, July 2003.

[Zhu04] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Systems*, 10(2) :98–115, April 2004.

[Zhu05] X. Zhu, A. K. Elmagarmid, X. Xue, L. Wu, and A. C. Catlin. Insightvideo : toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 07(4) :648– 666, August 2005.

Table des figures

Liste des tableaux

Publications

M. Guironnet, D. Pellerin, M. Rombaut, Camera motion classification based on transferable belief model, *European Signal Processing Conference (EUSIPCO'2006)*, Florence, Italy, September 2006.

M. Guironnet, N. Guyader , D. Pellerin and P. Ladret, Spatio-temporal attention model for video content analysis, *IEEE International Conference on Image Processing (ICIP'2005)*, Genova, Italy, September 2005.

M. Guironnet, D. Pellerin, M. Rombaut, Video classification based on low-level feature fusion model, *European Signal Processing Conference (EUSIPCO'2005)*, Antalya, Turkey, September 2005.

M. Guironnet, N. Guyader , D. Pellerin and P. Ladret , Static and dynamic feature based visual attention model : comparison to human judment, *European Signal Processing Conference (EUSIPCO'2005)*, Antalya, Turkey, Septembre 2005.

M. Guironnet, D. Pellerin, and P. Ladret, Video summarization using fuzzy descriptors and a temporal segmentation, *International conference on Advanced Concepts for Intelligent Vision Systems (ACVIS'2004)*, Brussels, Belgium, September 2004.

M. Guironnet, D. Pellerin, and P. Ladret, Combinaison de descripteurs flous de couleur et d'activité pour le résumé de vidéos, *14ème congrès de Reconnaissance des Formes et Intelligence Artificielle (RFIA'2004)*, Toulouse, January 2004.

M. Guironnet, D. Pellerin and P. Ladret, Création de résumés de vidéos appliquée à la recherche par l'exemple, *19ème colloque sur le traitement du signal et des images (GRETSI'03)*, Paris, September 2003.

G. Quénot, D. Moraru, S. Ayache , M. Charhad , M. Guironnet , L. Carminati , P. Mulhem, J. Gensel, D. Pellerin , L. Besacier, CLIPS-LIS-LSR-LABRI experiments at TRECVID 2004, *TREC Video Retrieval Evaluation*, November 2004.

Résumé

Le volume grandissant de vidéos a suscité le besoin de nouveaux outils d'aide à l'indexation. Un des outils possibles est le résumé de vidéo qui permet de fournir un aperçu rapide à l'usager. L'objectif de cette thèse est d'extraire, à partir d'informations visuelles, un résumé de vidéo contenant le « message » de la vidéo. Nous avons choisi d'étudier trois nouvelles méthodes de résumé de vidéo utilisant différentes informations visuelles.

La première méthode de résumé repose sur des caractéristiques de bas niveau (couleur, orientation et mouvement). La combinaison de ces index qui s'appuie sur un système d'inférence floue a permis de construire un résumé hiérarchique. Nous avons montré l'intérêt d'un tel résumé dans une application de la recherche par l'exemple.

La deuxième méthode de résumé est construite à partir du mouvement de caméra. Cette caractéristique de plus haut niveau sémantique est réfléchie par le réalisateur et induit une information sur le contenu. Une méthode de classification des mouvements basée sur le Modèle des Croyances Transférables est élaborée. La méthode de résumé est alors établie selon des règles sur l'amplitude et l'enchaînement des mouvements de caméra identifiés.

La troisième méthode de résumé est développée à partir de l'attention visuelle. Connaître les endroits où le regard se porte lors du visionnage de la vidéo est une information de plus haut niveau sémantique et pertinente pour créer le résumé. Un modèle spatio-temporel d'attention visuelle est proposé, puis utilisé pour détecter le changement de contenu au cours du temps afin de construire le résumé.

Mots clés : résumé de vidéo, mouvement de caméra, attention visuelle, détection de plans, classification de vidéos, Modèle des Croyances Transférables (MCT), système d'inférence floue

Abstract

The growing volume of video leads to the need of new tools for indexing. One of the possible tools is video summary which provides a fast overview to the user. The objective of this thesis is to extract from visual information, a summary containing the "message" of video. We chose to study three new methods of video summary using different types of visual features.

The first method of summary rests on low level features (color, orientation and motion). The combination of these features which is based on a fuzzy inference system allows a hierarchical summary to be built. We show the interest of such a summary in an application of query by example.

The second method of summary is built from camera motion. This higher level feature is thought by the filmmaker and so induces information on the content. A method of camera motion classification based on Transferable Belief Model is achieved. The method of summary is elaborated according to rules about the magnitude and the chain of the identified motions.

The third method of summary is developed from visual attention. To know the places where the glance is directed during the video playback is higher level information and relevant to create the summary. A spatio-temporal attention model is proposed, and then used to detect the change of content in time in order to build the summary.

Key words : video summary, camera motion, visual attention, shot detection, video classification, Transferable Belief Model (TBM), fuzzy inference system