N° d'ordre : 3888

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX I

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Claire MORAND**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Segmentation spatio-temporelle et indexation vidéo dans le domaine des représentations hiérarchiques

---

Soutenue le : 25 Novembre 2009

Après avis des rapporteurs :

Mme Christine Fernandez-Maloigne	Professeur, Université de Poitiers
M. Alan Hanjalic	Professeur, Delft University, Pays-Bas

Devant la commission d'examen composée de :

Mme Jenny Benois-Pineau	Professeur, Université Bordeaux 1	Directrice de thèse
M. Michel Crucianu	Professeur, CNAM, Paris	Président
M. Eric Debreuve	Chercheur CNRS, Université Nice-Sophia Antipolis	Examinateur
M. Jean-Philippe Domenger	Professeur, Université Bordeaux 1	Codirecteur de thèse
Mme Christine Fernandez-Maloigne	Professeur, Université de Poitiers	Rapporteur
M. Alan Hanjalic	Professeur, Delft University, Pays-Bas	Rapporteur

– 2009 –

Remerciements

Ce travail a été réalisé au sein du thème Analyse et Indexation Vidéo au Laboratoire Bordelais de Recherche en Informatique, sous la direction du Professeur Jenny Benois-Pineau. Je tiens à la remercier pour m'avoir accueillie au sein de son équipe pendant ces trois ans et pour la confiance qu'elle m'a accordée en m'offrant l'opportunité de participer à plusieurs congrès. Je souhaite également remercier le Professeur Jean-Philippe Domenger qui a accepter de co-encadrer cette thèse pour ces précieux conseils et sa bonne humeur.

Je remercie les Professeurs Christine Fernandez-Maloigne et Alan Hanjalic de m'avoir fait l'honneur d'accepter de juger ce travail en qualité de rapporteurs. Vos remarques ont été précieuses et les discussions passionantes et enrichissantes. Un très grand merci au Professeur Michel Crucianu pour avoir accepté de présider le jury de recherche et au Chargé de Recherche Eric Debreuve pour avoir accepté de prendre part au jury en tant que rapporteur de soutenance.

Merci aux membres du projet ANR ICOS-HD avec lesquels j'ai eu la chance de collaborer. Merci particulièrement aux "porteurs" du projet Christine Guillemot, Michel Barlaud et Jenny Benois-Pineau de m'avoir intégrée dans des discussions toujours plus enrichissantes.

Je tiens à remercier les membres du LaBRI, scientifiques et administratifs, que j'ai cotoyés tout au long de cette thèse pour leur amabilité et leur disponibilité, en groupe de travail, en enseignement ou près de la machine à café. Je ne saurai pas en faire la liste ici. C'est une grande chance pour moi d'avoir pu les rencontrer.
Merci aux "bons camarades" de bureau, de repas et de soirées, ils ont supporté quotidiennement mes piques et mes raleries. En ordre chronologique plus qu'approximatif : Nicolas/"Maïkeul" (désolée mais ce surnom unique te restera toujours), Petra, Eliana (comment va ton "chourse" ?), Daniel, Chris, Ronan (Super Animateur) et Stéphanie, Alex, Daniel (moi, "la méchante", vraiment ?), Aurore et Elric (qui m'ont supportée en tant qu'encadrante de leur stage), Svebor, Hugo, Luc, Jérôme, François (rassembleur du midi), Emilie, Anne-Laure, Nicolas, Frédéric. Et le dernier, et non des moindres, Rémi (félicitations à toi et Gaëlle), avec qui j'ai partagé le même bureau pendant près de quatre ans, ce qui a été pour moi une véritable chance tant sur le plan humain ("c'est Rémi !") que scientifique ("j'arrive pas à compiler...").

Une pensée toute particulière pour les amies au long cours, toujours présentes même après un long silence radio : Laëtitia et Marjorie.

Enfin, un énorme merci à ma soeur Elisabeth, toujours là pour les conseils éclairés dans les moments de blues. Et à mes parents, pour leur soutien non seulement inconditionnel et indéfectible, mais qu'ils ont aussi rendu d'une évidence immuable.

Segmentation spatio-temporelle et indexation vidéo dans le domaine des représentations hiérarchiques

Résumé : L'objectif de cette thèse est de proposer une solution d'indexation "scalable" et basée objet de flux vidéo HD compressés avec Motion JPEG2000. Dans ce contexte, d'une part, nous travaillons dans le domaine transformé hiérarchique des ondelettes 9/7 de Daubechies et, d'autre part, la représentation "scalable" nécessite des méthodes en multirésolution, de basse résolution vers haute résolution. La première partie de ce manuscrit est dédiée à la définition d'une méthode d'extraction automatique des objets en mouvement. Elle repose sur la combinaison d'une estimation du mouvement global robuste et d'une segmentation morphologique couleur à basse résolution. Le résultat est ensuite affiné en suivant l'ordre des données dans le flux scalable. La deuxième partie est consacrée à la définition d'un descripteur sur les objets précédemment extraits, à partir des histogrammes en multirésolution des coefficients d'ondelettes.
Enfin, les performances de la méthode d'indexation proposée sont évaluées dans le contexte de requêtes scalables de recherche de vidéos par le contenu.

Mots clés : Indexation, extraction d'objets, JPEG2000, ondelettes, histogrammes des coefficients d'ondelettes.

Discipline : Informatique

LaBRI (UMR CNRS 5800)
Université Bordeaux 1
351, cours de la Libération
33405 Talence Cedex (FRANCE)

Spatio-temporal Segmentation and Video Indexing in the Domain of Hierarchical Representations

Abstract: This thesis aims at proposing a solution of scalable object-based indexing of HD video flow compressed by MJPEG2000. In this context, on the one hand, we work in the hierarchical transform domain of the 9/7 Daubechies' wavelets and, on the other hand, the scalable representation implies to search for multiscale methods, from low to high resolution. The first part of this manuscript is dedicated to the definition of a method for automatic extraction of objects having their own motion. It is based on a combination of a robust global motion estimation with a morphological color segmentation at low resolution. The obtained result is then refined following the data order of the scalable flow. The second part is the definition of an object descriptor which is based on the multiscale histograms of the wavelet coefficients. Finally, the performances of the proposed method are evaluated in the context of scalable content-based queries.

Keywords: Indexing, Object Extraction, JPEG2000, Wavelets, Histogram of Wavelet Coefficients

Discipline: Computer Science

LaBRI (UMR CNRS 5800)
Université Bordeaux 1
351, cours de la Libération
33405 Talence Cedex (FRANCE)

Table des matières

Table des figures

Liste des Abréviations

La liste des abréviations utilisées dans ce manuscrit est donnée ci-après. Certains sigles, couramment utilisés dans la littérature, sont les sigles anglais. Dans ce cas, les formulations en langue anglaise et française sont précisées.

BD Base de Données

BF Basse Fréquence

BR Basse Résolution

CBVR Recherche de vidéos par le contenu (Content-Based Video Retrieval)

DFD Displaced Frame Difference

EMG Estimateur du Mouvement Global

HD Haute Définition

HF Haute Fréquence

HH Sous-bande de la TOD résultant de l'application du filtre passe-haut sur les lignes et les colonnes

HL Sous-bande de la TOD résultant de l'application du filtre passe-haut sur les lignes et du filtre passe-bas sur les colonnes

HR Haute Résolution

LL Nom générique des sous-bandes de BF de la TOD

LH Sous-bande de la TOD résultant de l'application du filtre passe-bas sur les lignes et du filtre passe-haut sur les colonnes

LPE Ligne de Partage des Eaux

MAD Moyenne des Différences des Valeurs Absolues (Mean of Absolute Difference

MAP Maximum A Posteriori

MCB Mise en Correspondance de Blocs

MRF Champs de Markov aléatoire (Markov Random Fields)

MSE Erreur Quadratique Moyenne (Mean Square Error

PDF Fonction de Densité de Probabilité (Probability Density Function)

PSNR Pic Rapport Signal à Bruit (Peak Signal to Noise Ratio)

ROI Région d'intérêt (Region Of Interest)

SD Définition Standard (Standard Definition)

TCD Transformée en Cosinus Dicrète

TOD Transformée en Ondelettes Discrète

TODC Transformée en Ondelettes Discrète Complète

VOD Video a la Demande (Video On Demand)

Introduction

Depuis quelques années, la vidéo numérique prend une place de plus en plus importante dans notre quotidien. Entre autres domaines d'applications, citons la Télévision Numérique Terreste (TNT) française (qui va remplacer totalement la diffusion télévisuelle analogique à l'horizon novembre 2011), le cinéma numérique, la diffusion par ADSL, la Vidéo à la Demande (VOD) et le Web 2.0 collaboratif où chacun peut distribuer ses propres vidéos. La variété des applications se traduit par une masse importante de données, stockées sous forme de Bases de Données (BD) qu'il faut gérer. La croissance de ces volumes de données est telle que leur indexation ne peut se faire que par des méthodes automatiques. Au début des années 1990, l'idée est apparue que le contenu des vidéos lui-même pouvait être utilisé comme base de l'indexation en opposition à l'utilisation d'annotations textuelles. Très vite, il s'est avéré intéressant en terme de coût calculatoire d'utiliser directement les informations contenues dans le flux compressé (forme sous laquelle les vidéos sont stockées) pour définir les méthodes d'indexation. De cette idée est née le Paradigme de l'Indexation Primaire [Man04].

A ce jour, bien que de nombreuses techniques aient été proposées, l'indexation et la recherche par le contenu des vidéos reste un problème ouvert [Han08]. De plus, la mutation numérique n'étant pas achevée, les données à traiter ne cessent d'évoluer et la recherche de nouvelles techniques d'indexation doit se poursuivre. D'une part, la qualité des vidéos s'améliore avec le passage de la Définition Standard (SD) à la Haute Définition (HD). Ce changement a pour conséquence notamment d'augmenter encore la taille des données à traiter et de fournir des vidéos de caractéristiques physiques, telles que le bruit d'acquisition, différentes. D'autre part, de nouveaux standards de compression dits "scalables"[1] (graduels) ont été développés. Ils proposent une nouvelle structuration, plus souple, du flux vidéo. Il faut alors adapter, voire re-penser, les méthodes d'indexation déjà développées à ces nouvelles informations. C'est un des buts du projet ANR "ICOS-HD" [ICO] auquel le présent travail de thèse est associé.

[1] Bien que le terme "scalable" soit un anglicisme, nous avons décidé de l'adopter dans ce manuscrit. Des précisions sur sa signification sont données dans le chapitre 1

L'objectif de ce travail de thèse est de proposer une solution d'indexation automatique des vidéos qui s'adapte à la fois aux contraintes de la vidéo HD et aux contraintes du standard scalable utilisé pour la compression. Nous avons choisi de nous intéresser au standard JPEG2000, utilisé dans des applications telles que le cinéma numérique ou les archives vidéo numériques. Ce standard se distingue de ses prédécesseurs non seulement par sa propriété de scalabilité, mais aussi par l'utilisation de la Transformée en Ondelettes Discrète (TOD) pour effectuer la réduction des données. Grâce à cette transformation, chaque trame de la vidéo est représentée sous une forme hiérarchique qui peut être ré-utilisée pour la définition d'un descripteur scalable. Ce descripteur, nous choisissons de le construire uniquement sur les objets de la scène ayant leur propre mouvement. Nous pensons qu'une telle approche est au moins aussi prometteuse que les techniques les plus populaires d'indexation actuelles qui sont réalisées par des descripteurs locaux. En effet, utiliser les objets, tout en réduisant la quantité de données sur lesquelles définir le descripteur, apporte une information sémantique de haut niveau. Nous proposons de définir notre indice à l'aide d'histogrammes d'ondelettes en multirésolution.
Le préalable à la construction de notre descripteur est donc de disposer d'une méthode d'extraction automatique des objets en mouvement propre dans une vidéo. Cette thématique a largement été étudiée dans la littérature. Cependant, comme les problèmes d'extraction d'objets et de segmentation sont mal posés [Ber88], il n'en existe pas de solution unique et la recherche dans ce domaine reste d'actualité. Notre travail consiste alors en la définition d'une telle méthode dans le domaine compressé JPEG2000. En nous appuyant sur les techniques établies de l'état de l'art, nous nous proposons de combiner une estimation robuste du mouvement global et une segmentation couleur morphologique à Basse Résolution (BR). Le résultat de cette extraction est ensuite affiné en résolution en suivant l'ordre des données dans le flux scalable. Notre principal apport est de tirer parti de l'information de Haute Fréquence (HF) fournie par la pyramide d'ondelettes.
Enfin, la qualité d'un descripteur se définit par sa capacité à apporter des réponses satisfaisantes à des tâches d'indexation. Une partie de notre travail porte ainsi sur la définition de requêtes par similarité dans le contexte de la scalabilité qui nous permettent d'évaluer les performances du descripteur proposé.

Le présent manuscrit est organisé comme suit.

- Le **chapitre 1** poursuit l'introduction en expliquant le terme "scalable" du point de vue de la compression et de l'indexation. Les objectifs de la thèse par rapport à ces notions et l'originalité des méthodes proposées sont alors présentés.

- Les **chapitres 2 et 3** constituent la **1ère partie** du manuscrit consacrée à l'état de l'art. Le **chapitre 2** présente le standard JPEG2000, cadre de travail dans lequel sont définies nos méthodes. Le **chapitre 3** indique les grands axes de recherche dans

le domaine de l'indexation des vidéos d'une part et dans le domaine de l'extraction d'objets d'autre part.

- Les **chapitres 4 et 5** forment la **2ème partie** du manuscrit consacrée à la méthode d'extraction d'objets dans le domaine compressé JPEG2000. Le **chapitre 4** présente notre méthode d'estimation de mouvement hiérarchique. Le **chapitre 5** décrit notre méthode d'extraction hiérarchique d'objets.

- Les **chapitres 6 et 7** constituent la **3ème partie** du manuscrit. Le **chapitre 6** présente notre descripteur d'objets construit à partir des histogrammes d'ondelettes. Le **chapitre 7** est consacré à l'évaluation des performances du descripteur proposé.

- Le **chapitre 8** conclut le manuscrit et présente les perspectives de poursuite de notre travail.

- Les **annexes A et B** présentent les détails de mise en oeuvre des algorithmes. L'**annexe A** est consacrée aux logiciels d'encodage des flux par JPEG2000. L'**annexe B** présente les méthodes de détermination de seuils pour la méthode d'estimation de mouvement.

Chapitre 1

Le contexte de la scalabilité

La définition d'un descripteur pour l'indexation et la recherche par similarité peut se faire directement sur les données du flux compressé. Cela permet de ne pas décoder complètement la vidéo, ce qui serait prohibitif en temps de calcul. Cette approche est décrite par le *Paradigme de l'Indexation Primaire*, introduit par Manerba, Benois-Pineau et al. [Man04] pour les vidéos encodées en MPEG2. Cependant l'utilisation d'un tel standard de compression n'est plus acceptable pour l'encodage des vidéos HD et des solutions de compression scalable (section 1.1) qui propose un flux plus malléable lui sont préférées. L'extension du Paradigme de l'Indexation Primaire à ces standards de compression "nouvelle génération" est un objectif sous-jacent du projet ICOS-HD. D'autre part, il est possible de définir par analogie avec la compression le concept d'indexation scalable (section 1.2). Le travail présenté dans ce manuscrit de thèse propose une solution d'indexation scalable des vidéos HD (section 1.4).

1.1 La scalabilité : une nouvelle fonctionnalité des standards de compression vidéo

Le terme de *scalabilité* n'existe pas officiellement dans le vocabulaire français. C'est un anglicisme dérivé du mot "scalability" que l'on peut traduire par "la capacité d'être échelonnable, graduel". Cependant, une telle traduction ne recouvre pas toute la définition, donnée dans la suite de ce paragraphe, de ce mot. C'est pourquoi nous nous permettons d'utiliser le terme de scalabilité dans la suite de ce manuscrit.

La scalabilité est la propriété de la représentation compressée d'une image ou d'une vidéo de pouvoir être décompressée de différentes manières. Ces différentes décompressions aboutissent à différentes versions de l'image ou de la vidéo d'origine. La représentation codée de l'image (la vidéo) compressée, ou train binaire, est ainsi formée de sous-ensembles.

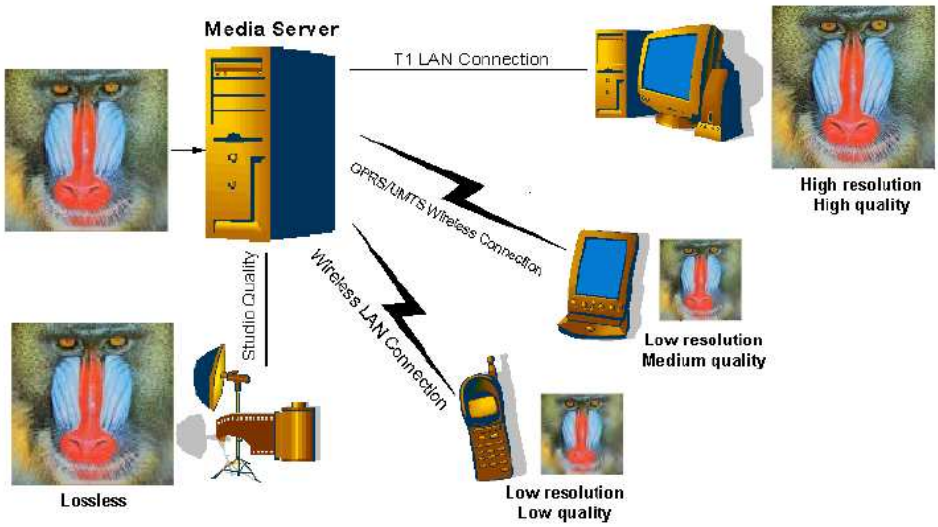


Fig. 1.1 – Exemple d'utilisation de la scalabilité : diffusion multi-terminaux, d'après [Gra07]

Chacun de ces sous-ensembles représente une compression efficace de l'image (la vidéo) originale à une résolution et/ou une qualité réduite. Ce train binaire peut alors être tronqué et, une fois décodé, conduire à une image (une vidéo) de moindre qualité. Pour être plus précis, les sous-ensembles sont organisés en couches. La "couche de base", conduisant à la version de résolution et/ou qualité la plus faible, est utilisée dans tous les schémas de décompression. Elle peut être combinée avec une ou plusieurs autres couches dites d'amélioration qui contiennent des informations de détails plus ou moins fins. Un des avantages de la compression scalable est que le taux d'échantillonnage ou la résolution de reconstruction nécessaires à une application donnée n'ont pas besoin d'être connus au moment de la compression de l'image (de la vidéo).

La propriété de scalabilité peut porter sur différentes caractéristiques d'un flux vidéo : qualité, résolution, localisation spatiale, composante, échantillonnage temporel. Les deux scalabilités les plus importantes, résolution et qualité, ont été évoquées dans le début du paragraphe.

Actuellement, deux formats de compression vidéo qui proposent la propriété de scalabilité ont fait l'objet d'une norme : JPEG2000 dans sa version Motion-JPEG2000 [ISO07] et H.264/SVC [ISO09].

Exemples d'utilisation de la compression scalable

Diffusion multi-terminaux. Avec la multiplication des réseaux de communications, plusieurs terminaux de capacités différentes (résolution d'écran, capacités de calcul, bande passante,...) peuvent se connecter à la même BD vidéo. C'est le cas, par exemple, des

applications de VOD. Dans ces conditions, le serveur de la BD doit être en mesure de fournir à un terminal spécifique une version de la vidéo adaptée à ses capacités afin de réduire la bande passante nécessaire et le temps de décompression. Sans flux scalable, il faut soit transcoder en temps réel la vidéo soit avoir sauvegardé plusieurs versions de cette vidéo dans la BD. Grâce à la compression scalable, une seule version de la vidéo est nécessaire, le serveur ne transmettant que la partie pertinente du flux compressé (Figure 1.1).

Navigation dans les Bases de Données. Qu'elles soient médicales ou cinématographiques, les archives vidéo représentent des masses de données importantes, souvent disséminées sur un large réseau. La navigation dans ces BD peut être résumée en trois phases. D'abord, la vue d'ensemble par le biais de vignettes de faible qualité permet une première sélection des données. Ensuite, la sélection peut être affichée à une qualité intermédiaire, ce qui permet à l'utilisateur de mieux appréhender le contenu et d'affiner son choix. Enfin, les données ainsi collectées sont transmises en intégralité à l'utilisateur. Les scalabilités en résolution et en qualité sont parfaitement adaptées à ce type de scénario.

Post-production cinématographique. De même que dans le scénario précédent, deux phases peuvent être définies dans la post-production cinématographique. Pendant l'édition et pour la sélection des vidéos, seule une faible résolution est nécessaire. Pour tester les effets, les images-clés peuvent être transmises à la station d'édition en haute qualité et résolution.

1.2 Indexation Scalable des Vidéos

La notion d'indexation scalable des contenus vidéo est un domaine de recherche jeune et les définitions de cette notion sont, à notre connaissance, très disparates dans la littérature. Nous pensons intéressant de présenter ici notre propre définition de l'indexation scalable par rapport à nos attentes et aux définitions trouvées dans l'état de l'art.

Notre point de vue sur l'indexation scalable est de la définir par analogie et par complémentarité avec la compression scalable[1]. Un descripteur est scalable s'il est adapté aux flux compressés scalables. Deux cas peuvent être distingués, suivant que le descripteur est calculé du côté de l'encodeur ou du décodeur. Du côté de l'encodeur, le but est d'embarquer le descripteur dans le flux codé. Le descripteur doit alors être structuré de telle sorte que sa représentation soit formée de sous-ensembles, chacun de ces sous-ensembles permettant de décrire la portion du flux scalable de la vidéo à laquelle il est associé. Une solution de codage du descripteur dans le flux scalable a récemment été proposée par Adami et al

[1]Une autre acceptation du sens d'indexation scalable est par exemple par analogie avec la notion de "montée en charge" des architectures logicielles et d'électricité. Dans ce sens, comme l'expriment Lui et Izquierdo [Lui03], la scalabilité est la capacité d'un système de maintenir un temps quasi-constant et indépendant de la taille de la BD pour répondre à une requête . Ce type de problème est hors du sujet du travail présenté ici.

dans [Ada09]. Du côté du décodeur, le but est de construire un descripteur avec seulement les données disponibles, c'est-à-dire la part transmise du flux. Le descripteur doit alors suivre la hiérarchie scalable du flux encodé pour permettre la comparaison entre vidéos transmises à des résolutions différentes. Comme Piro et al [Pir09] l'expliquent, plusieurs utilisateurs ayant chacun une version différente de la même vidéo doivent obtenir les mêmes réponses à la même requête par similarité sur cette vidéo pour une BD donnée.
Enfin, la scalabilité de l'indexation peut être vue du point de vue de la qualité des résultats retournés par un outil de recherche par similarité. Ainsi Albuz et al [Alb01] constatent que si les descripteurs sont différents à basse résolution, alors ils sont différents à pleine résolution. La comparaison peut alors s'effectuer dans un premier temps sur la basse résolution uniquement.

Dans les trois cas évoqués (côté encodeur, côté décodeur et côté requête) le descripteur possède toujours la même structure empruntée à la compression scalable. Il est composé d'une couche de base représentant le descripteur de qualité et/ou résolution minimale et de couches d'améliorations porteuses d'informations sur les détails et permettant d'ajuster la qualité et/ou la résolution du descripteur en fonction des besoins.

Exemple d'utilisation de l'indexation scalable
La *Recherche par le Contenu dans des BD Vidéo* peut se faire soit avec une vidéo requête à BR, soit avec une vidéo requête à HR. En fait, la résolution de la requête est déterminée par le terminal qui en fait la demande. Suivant ses capacités, il ne peut pas afficher de vidéos en pleine résolution HD (c'est le cas des téléphones portables par exemple). Il n'a donc qu'une version BR de la vidéo et ne peut calculer que le descripteur de BR associé. La BD est, elle, disponible en HD. La comparaison entre vidéos peut soit s'effectuer seulement à la résolution de la requête pour une réponse rapide, soit proposer d'affiner les résultats en comparant à la pleine résolution pour obtenir une réponse plus précise.

1.3 Extraction d'objets scalable

Nous l'avons évoqué dans l'introduction, nous nous proposons de définir une indexation basée objet des vidéos de HD compressées en JPEG2000. Il nous faut donc définir une méthode d'extraction des objets d'intérêt respectant le principe de scalabilité du contexte. Comme pour l'indexation, nous définissons le principe d'extraction d'objet scalable par analogie avec la compression scalable.

L'extraction d'objet est scalable si elle est adaptée aux flux compressés scalables. Là encore, deux cas peuvent être distingués suivant que l'extraction d'objet s'effectue du côté de l'encodeur ou du côté du décodeur. Du côté de l'encodeur, le but est d'extraire l'objet du fond, le plus souvent sur une pyramide en multi-résolution. L'idée est d'utiliser cette

information pour définir des Régions d'Intérêt (ROI) qui seront encodées avec plus de précision. La forme de l'objet extrait doit alors être similaire à tous les niveaux de résolution pour que différents utilisateurs, pourvus de terminaux ayant des résolutions d'affichage différentes, puissent visualiser avec une qualité correcte la même ROI [Akh05]. Pour obtenir cette cohérence de forme sur la pyramide, l'extraction de l'objet à un niveau de résolution se fait en fonction du résultat de l'extraction au niveau de résolution supérieure et du niveau de résolution inférieure. Du côté du décodeur, l'extraction doit se faire avec seulement les données transmises du flux. L'extraction est dans ce cas une étape préalable à une application donnée ; dans notre manuscrit il s'agit de l'utiliser comme base pour construire un descripteur scalable et faire de la recherche par similarité dans les bases de données. Il est alors important de suivre l'ordre de la hiérarchie scalable du flux encodé afin de permettre la comparaison entre vidéos transmises à des résolutions différentes. L'extraction de l'objet à un niveau de résolution se fait en fonction uniquement du résultat de l'extraction au niveau de résolution inférieure.

Dans les deux cas (côté encodeur et côté décodeur), l'objet extrait peut être représenté sous forme scalable, c'est-à-dire en une couche de base représentant l'objet extrait à basse résolution et des couches d'améliorations représentant les détails de l'objet. Cependant, les résultats d'une extraction d'objets côté encodeur seront différents de ceux côté décodeur du fait que dans le premier cas, l'information de HR peut être utilisée pour affiner l'extraction à BR. Ces deux cas ne sont pourtant pas incompatibles. En effet, une fois l'encodage des ROI effectué pour construire le flux compressé, il n'est pas besoin d'appliquer une nouvelle méthode d'extraction d'objets, les lieux des objets étant directement accessibles dans le flux. Notons que dans le cadre de cette thèse, nous supposons que les flux sont encodés sans information de ROI et que, dans ce cas, il nous faut extraire les objets.

1.4 Objectifs et Problématique de l'étude

Le travail présenté dans cette thèse a pu être confronté et intégré au projet ANR ICOS-HD. Avant d'indiquer la problématique et les objectifs de ce manuscrit, nous présentons d'abord rapidement le projet ICOS-HD.

Le projet ICOSHD [ICO]
Le projet "Indexation et COmpression Scalables et conjointes pour la gestion des contenus vidéo de Haute Définition" (ICOS-HD, [ICO]), soutenu par l'ANR, a débuté le 1er janvier 2007. Son but "est de proposer de nouvelles solutions de description scalable des contenus vidéo HD facilitant leur édition, diffusion et accès dans des infrastructures (réseaux, terminaux) hétérogènes." Le projet se découpe en quatre objectifs majeurs :

1. Développer des méthodes d'extraction de descripteurs spatio-temporels scalables dans le flux compressé généré suivant un standard scalable,

2. Etudier de nouvelles transformations générant des représentations hiérarchiques des séquences et adaptées à la fois aux besoins de compression et d'indexation,

3. Développer des méthodes d'extraction de descripteurs dans le domaine pixel et adaptées à la problématique du point précédent,

4. Définir des scénarios applicatifs de ces descriptions scalables permettant d'en évaluer la pertinence.

Les méthodes et résultats présentés dans ce manuscrit s'inscrivent dans les objectifs 1 et 4 de ce projet.

Objectifs de la thèse
Ce travail de thèse propose une solution d'indexation automatique, basée objet, des vidéos et qui s'adapte à la fois aux contraintes de la vidéo HD et aux contraintes du standard scalable JPEG2000 utilisé pour la compression. Les trois objectifs majeurs développés dans ce manuscrit sont :

1. **Proposer une méthode d'extraction spatio-temporelle des objets en mouvement dans le domaine JPEG2000**. Le problème d'extraction d'objet a largement été étudié dans la littérature. Bien que ce problème n'ait pas de solution générique satisfaisante du fait de sa nature de problème mal posé au sens de Hadamard [Ber88], des solutions offrant des résultats très satisfaisants existent. Notre objectif ici n'est pas d'inventer une nouvelle méthode d'extraction d'objet, mais de tirer parti des méthodes existantes dans un schéma multi-résolution d'extraction d'objet. Nous proposons d'ajouter des contraintes liées à l'utilisation des coefficients de HF associés à la représentation en ondelettes du flux JPEG2000 afin d'améliorer les résultats d'extraction.

2. **Proposer un descripteur scalable du contenu vidéo uniquement sur les objets précédemment extraits**. Nous pensons que travailler sur des objets permet non seulement de réduire la quantité d'information à décrire de façon pertinente par rapport au contenu de la vidéo mais aussi d'apporter une information supplémentaire (celle d'objet) de niveau sémantique élevé. L'objectif est là aussi de définir un descripteur à partir de descripteurs ayant fait la preuve de leur efficacité. L'originalité de notre approche porte sur l'adaptation du descripteur à une représentation en objet scalable, à la structuration du descripteur en un descripteur scalable et à la prise en compte des coefficients de HF dans le calcul du descripteur.

3. **Evaluer les performances du descripteur dans le cadre de requêtes par similarité scalables**. Nous définissons un ensemble de scénarios de requêtes permettant de tenir compte de la propriété de scalabilité du contexte et offrant ainsi plus de souplesse à l'utilisateur dans sa recherche. Suivant le scénario envisagé, tout ou seulement une portion du descripteur scalable est utilisé pour effectuer la tâche demandée. Nous évaluons les performances du descripteur et la portion minimum du descripteur pertinente pour une tâche donnée.

Première partie

Etat de l'art

Chapitre 2

Le standard JPEG2000

Un des objectifs de ce manuscrit est de proposer une solution d'indexation des vidéos HD à partir du flux compressé JPEG2000. Le présent chapitre est consacré à cette norme de compression. Nous présentons d'abord les objectifs de ce standard par rapport aux standards précédemment existants et quelques uns de ses cas les plus courants d'utilisation. Nous montrons ainsi la place importante qu'il occupe dans le paysage du stockage des données vidéo HD. Une description plus technique en est ensuite proposée sous la forme d'une présentation générale de la structure de l'encodeur JPEG2000. Les étapes majeures de cette compression sont présentées. Notre objectif ici n'est pas d'expliquer de façon exhaustive tous les méchanismes intervenants dans la compression mais de présenter la nature des données à disposition dans le flux et que nous utilisons pour faire de l'indexation. Ainsi, nous porterons une attention plus soutenue à la TOD, coeur et originalité du standard.

2.1 Présentation

MJPEG2000 [ISO07] (encore appelé Motion-JPEG2000) est l'un des deux standards actuels de codage scalable (cf section 1.1) des vidéos, l'autre étant H.264 [ISO09]. C'est l'extension pour les vidéos du standard de compression d'images fixes JPEG2000 [ISO04a] : chaque trame du flux vidéo est considérée individuellement et compressée en JPEG2000. la corrélation temporelle du flux vidéo. C'est pourquoi, dans la suite, nous présentons le standard JPEG2000 et nous utiliserons aussi bien les termes MJPEG2000 et JPEG2000 pour la compression de vidéos.

Initié en mars 1997 et normalisé dès décembre 2000, le standard JPEG2000 [ISO04a] a été développé pour être un complément - et non un remplaçant - des standards de compression d'images existants. Sa principale caractéristique est de proposer une seule architecture de compression adaptée à diverses applications. En comparaison, le standard JPEG pos-

sède près de 44 modes différents, la plupart non reconnus par la majorité des décodeurs. JPEG2000 permet ainsi la compression de différents types d'images (bi-niveaux, à niveaux de gris, couleur, hyper-spectrales) avec des caractéristiques différentes (naturelles, scientifiques, de télédétection, graphiques, composées...) et pouvant être de très grande taille (supérieure à 64K x 64K pixels). Entre autres propriétés liées à la scalabilité, on peut citer la possibilité d'embarquer dans le même train binaire les compressions avec et sans pertes, la conservation de la qualité visuelle d'une image à de faibles taux d'échantillonage et le codage de ROI.

Cas pratiques d'utilisation

Les domaines d'utilisation possibles de JPEG2000 sont nombreux : navigation internet, numérisation, photographie numérique, télédétection, communications mobiles, archives numériques, imagerie médicale... Nous présentons ici deux cas concrets d'utilisation.

Cinéma numérique. Sous l'impulsion de la DCI (Digital Cinema Initiative, LLC, [DCIL]), le standard MJPEG2000 a été adopté comme standard de compression du cinéma numérique. La DCI est une organisation créée en mars 2002 qui regroupe six majors du cinéma américain (Disney, Fox, Paramount, Sony Pictures Entertainment, Universal et Warner Bros. Studios). Son but principal est de définir un ensemble de spécifications assurant ainsi des performances uniformes et de qualité, ainsi que la fiabilité et le contrôle de la qualité de la diffusion du cinéma numérique. Ce standard a été choisi car il permet de favoriser la qualité de l'image par rapport à la bande passante qui est de moindre importance dans ce type d'application. De plus, chaque image étant codée indépendamment des autres, l'édition du flux compressé est facilitée. L'accès aléatoire est immédiat et toutes les opérations d'édition (suppression, découpage, ajout) sont aisées et sans pertes. Ce format est donc particulièrement adapté pour le montage vidéo. Un amendement a ainsi été ajouté au standard pour définir les profils à utiliser pour des applications de cinéma numérique ([ISO06]) et d'autres sont en voie de publication (ISO/CEI 15444-1/A2 : "Extended Profiles for Digital Cinema Applications", septembre 2009 et ISO/CEI 15444-1/A3 : "Guidelines for Digital Cinema Applications", juillet 2011).

Archives numériques. De nombreuses compagnies privées ou publiques recommandent MJPEG2000 pour les archives vidéo, en particulier pour son mode de compression sans pertes ([Gil06], Media Maters, llc). Cependant, la compression avec pertes reste très utile dans ce genre d'applications. Au vu d'études de comparaison de performances par rapport au standard H.264 en terme de courbes débit-distorsion telles que [Mar03], il est raisonnable de penser que MJPEG2000 sera préféré dans des applications maniant des vidéos de HD. En effet, Marpe et al. [Mar03] montrent que les performances relatives de H.264 et MJPEG2000 varient avec la résolution spatiale des vidéos. Pour du CIF, H.264 est le plus performant, pour du 720p les performances des deux standards sont comparables et pour du 1080p JPEG2000 tend à être supérieur.

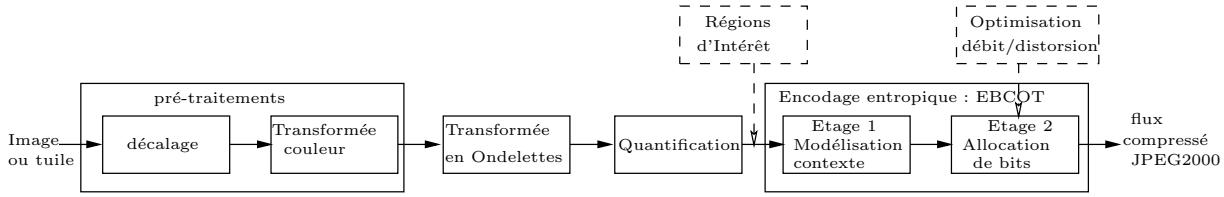


Fig. 2.1 – Schéma simplifié de l'encodeur JPEG2000.

Ainsi, même si JPEG2000 n'est pas destiné à être le standard universel de compression des vidéos HD, ses performances font qu'il est déjà utilisé pour coder une proportion importante des données vidéos. C'est pourquoi nous pensons que développer des solutions d'indexation automatique basées JPEG2000 est crucial pour gérer la masse de données existante et à venir.

2.2 Description

Après avoir introduit le standard JPEG2000 par ses objectifs et ses applications, nous présentons dans ce paragraphe sa structure. Le but ici n'est pas d'en faire une présentation complète mais d'indiquer le cadre dans lequel se placent nos travaux de recherche. Ainsi, même si deux schémas peuvent être utilisés pour coder une vidéo en JPEG2000 (un chemin réversible, sans pertes, et un chemin irréversible, avec pertes), nous avons choisi de nous focaliser sur la compression avec pertes. Pour une présentation plus complète, le lecteur pourra se référer à [Tau02].

Le schéma générique de codage JPEG2000 est présenté figure 2.1. L'image peut d'abord être découpée en sous-images appelées "tuiles" ("tiles" en anglais) sur lesquelles l'algorithme de codage est appliqué indépendamment. Cela permet de réduire la quantité de mémoire nécessaire à l'encodage. Cependant, une image entière peut tout à fait n'être constituée que d'une seule tuile. Chaque composante de l'image (de la tuile) est codée séparément. Des pré-traitements peuvent être appliqués (section 2.3) permettant de mettre les données sous une forme optimale pour la transformation suivante. Ces étapes facultatives étant passées, l'image est décorrélée par Transformée en Ondelettes Discrète (TOD) (section 2.4). La TOD constitue l'originalité du standard JPEG2000, en rupture technologique au sens de la base de représentation avec les familles d'encodeurs H.26X et MPEG qui utilisent la Transformée en Cosinus Discrète (TCD) sur des blocs de petite taille. Les coefficients ainsi obtenus sont ensuite quantifiés afin de diminuer la quantité d'information (section 2.5). Un codeur entropique prend alors en charge la mise en forme du train binaire, c'est-à-dire la représentation sous forme binaire de l'image compressée. L'algorithme de codage utilisé est l'EBCOT ("Embedded Block Coding with Optimized Truncation", [Tau00]) et constitue,

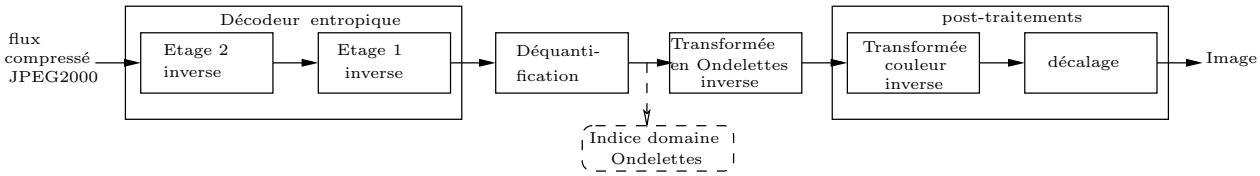


Fig. 2.2 – Schéma simplifié du décodage JPEG2000. Le lieu d'extraction de descripteurs proposé est indiqué en pointillé

avec la TOD, le coeur du standard (section 2.6).

JPEG2000 propose de nombreuses fonctionnalités. Parmi celles-ci, on peut évoquer la possibilité d'une gestion de ROI (représenté sur la figure 2.1 en pointillé). L'utilisateur peut choisir une région dans l'image qui sera codée avec plus de précision. La forme de cette région n'est pas conditionnée et, suivant la technique d'encodage de la ROI utilisée, il n'est même pas nécessaire de stocker et transmettre au décodeur des informations supplémentaires telles que le masque de la région.
JPEG2000 suit le principe de scalabilité (section 1.1). Ainsi, au moment de la compression, l'utilisateur décide de la résolution maximale et de la qualité de l'image maximale à utiliser. L'algorithme d'encodage utilise les informations de multirésolution intrinsèquement obtenues par la TOD et les différentes couches de qualité obtenues par une allocation de bits raisonnée au niveau du codeur EBCOT (en pointillé sur la figure 2.1) pour atteindre cet objectif. Toutes les qualités d'image ou toutes les tailles peuvent être alors décompressées du train binaire encodé résultant. Il est aussi possible d'avoir un accès aléatoire en ne décompressant qu'une certaine région de l'image ou une composante spécifique de l'image (par exemple la composante à niveaux de gris d'une image couleur). Il n'est pas nécessaire de décoder entièrement le flux avant d'effectuer l'extraction de la région d'intérêt ou de la composante désirée. En fait, les bits extraits et décodés sont généralement identiques à ceux que l'on aurait obtenus si seule l'image de résultat désirée avait été encodée. Cette propriété de JPEG2000 est très importante car elle permet de prévenir l'augmentation du bruit de compression au travers des différents cycles décompression/re-compression.
Enfin, les transformations géométriques basiques (rotation, symétrie axiale, translation, mise à l'échelle et découpage) peuvent facilement être appliquées à la représentation compressée de l'image. Cela permet d'éviter d'avoir à décompresser et re-compresser l'image pour faire ces traitements.

Le schéma de décompression du standard JPEG2000, présenté en figure 2.2, est l'inverse du schéma de codage. Dans le cadre du paradigme de l'Indexation Primaire [Man04], nous souhaitons indexer des vidéos déjà compressées en utilisant des informations obtenues par décompression partielle. Le lieu possible d'extraction d'indices que nous avons envisagé

dans nos travaux est indiqué en pointillé sur la figure 2.2. Nous avons décidé d'intervenir après la déquantification et de travailler dans le domaine des ondelettes. Les ondelettes sont intéressantes car elles représentent l'image à la fois en Basse Fréquence (BF), c'est-à-dire sous une forme proche du domaine pixel qui permet d'utiliser des approches images bien connues, et en Haute Fréquence (HF), c'est-à-dire sous une forme qui apporte des informations précieuses sur les contours et la texture des objets. Afin de définir la nature des données sur lesquelles nous allons bâtir notre méthode d'indexation, nous détaillons dans la suite les grandes étapes du processus d'encodage JPEG2000.

2.3 Pré-traitements

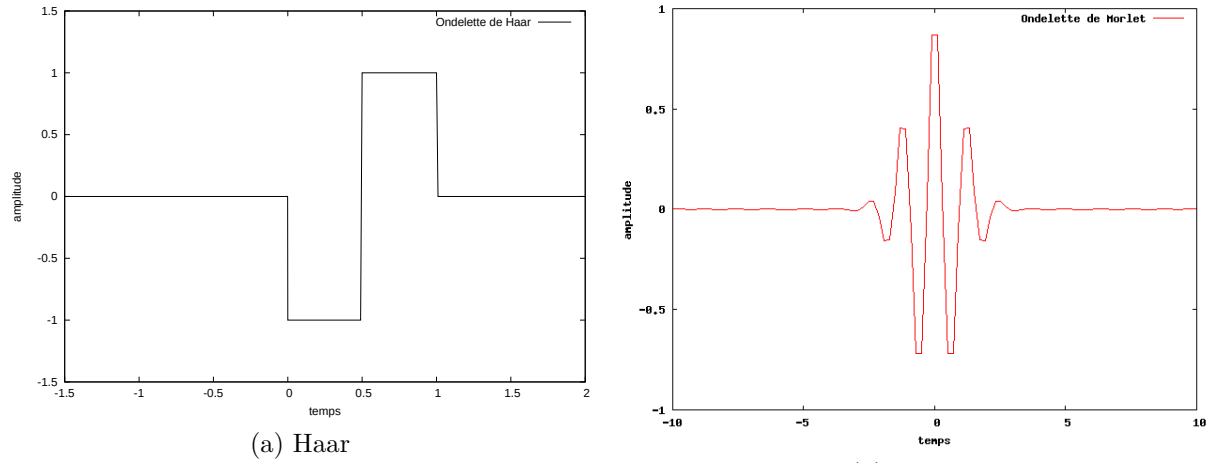
Les étapes de pré-traitement servent à préparer les données pour la TOD. Le standard JPEG2000 permet de compresser toutes sortes d'images. Cependant, dans les applications qui nous intéressent, les données d'entrée sont supposées être des images au format RVB.

Décalage
Les filtres Passe-Haut (PH) et Passe-Bas (PB) utilisés pour réaliser la TOD (section 2.4) ont été créés pour des données de dynamique centrée en zéro. Or les images d'entrée en RVB sont représentées par des valeurs entières positives codées sur B bits (typiquement, $B = 8$ pour chaque composante). Donc si $i[l, c]$ est la valeur du pixel au point de coordonnées (l, c), alors $0 \leqslant i[l, c] < 2^B$. La valeur décalée est obtenue par $\tilde{i}[l, c] = i[l, c] - 2^{B-1}$. Ainsi $\forall (l, c) \in \mathbb{N}^2, -2^{B-1} \leqslant \tilde{i}[l, c] < 2^{B-1}$; les données ont une dynamique centrée en zéro. Si les données sont déjà signées, il n'y a pas d'ajustement.

Transformée couleur.
Cette phase s'applique uniquement pour les images couleurs au format RVB. Elle permet de décorréler les composantes couleur en passant dans le système couleur YC_rC_b proposant la séparation de la luminance et de la chrominance. Cette transformation couleur irréversible (dans la mesure où les calculs en virgule flottante ne permettent pas de retrouver exactement les valeurs initiales par transformation inverse) est linéaire et s'écrit sous forme matricielle comme suit (2.1).

$$\begin{pmatrix} Y \\ C_r \\ C_b \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.16875 & -0.33126 & 0.5 \\ 0.5 & -0.41869 & -0.08131 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.1}$$

La transformation inverse est alors donnée par (2.2).



(a) Haar

(b) Morlet

Fig. 2.3 – Exemples d'ondelettes (a) Haar : $\psi(t) = -1$ si $t \in [\ 0\ ;\ 0.5\ [; 1$ si $t \in [0.5; 1[$ et 0 sinon, (b) Morlet $\psi(t) = e(-\frac{t^2}{2.0})cos(5t)$

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} 1.0 & 0.0 & 1.402 \\ 1.0 & -0.34413 & -0.71414 \\ 1.0 & 1.772 & 0.0 \end{pmatrix} \begin{pmatrix} Y \\ C_r \\ C_b \end{pmatrix} \tag{2.2}$$

2.4 La Transformée en Ondelettes Discrète

Par rapport aux familles d'encodeurs H.26X et MPEG, JPEG2000 représente une rupture technologique au sens de la base de représentation du signal image/vidéo. Il n'utilise pas de TCD ou son approximation entière, mais la TOD. Cette transformée permet d'assurer la scalabilité en résolution du standard. Cette section présente le principe de la transformée en ondelettes dans le domaine continu, puis détaille son implémentation dans le domaine discret. Au vu des définitions précédentes, l'intérêt que représente pour nous une telle transformation pour l'extraction d'un descripteur scalable est indiqué. La présente section n'est qu'une brève introduction au formidable outil mathématique que sont les ondelettes. Une présentation plus complète peut être trouvée dans les ouvrages de référence de Meyer [Mey90], Daubechies [Dau92] et Mallat [Mal98]. Citons aussi le livre de Misiti et al. [Mis03] qui présente de façon claire et pédagogique les ondelettes et leurs applications.

La Transformée en Ondelettes Continue

Dans cette partie, nous nous limitons à la présentation de la transformée en ondelettes 1D afin d'en faciliter l'explication sans surcharger les formules. Nous noterons alors de

façon traditionnelle la variable décrivant la dimension par t le temps. Une ondelette est une fonction ψ oscillante, localisée et suffisamment régulière (cf figure 2.3). Ces propriétés se traduisent par la condition d'admissibilité dans le domaine fréquentiel suivante :

$$\psi \in L^1 \cap L^2 \text{ et } \int_0^{+\infty} \frac{|TF[\psi](\omega)|^2}{|\omega|} d\omega = \int_{-\infty}^{0} \frac{|TF[\psi](\omega)|^2}{|\omega|} d\omega < +\infty \tag{2.3}$$

où L^1 désigne l'espace des fonctions intégrables sur $\mathbb{R}$, L^2 est l'espace des fonctions de carré intégrable sur $\mathbb{R}$, $TF[\psi]$ est la Transformée de Fourier de la fonction ψ et ω est la variable de fréquence. A partir de cette unique fonction ψ, il est possible de construire une base d'ondelettes qui sert à l'analyse de fonctions. Ainsi, les fonctions $\psi_{h,\tau}$ de la base sont obtenues par translation (de paramètre τ) et dilatation (de facteur h) de la fonction ψ appelée *ondelette mère*.

$$\psi_{h,\tau}(t) = \frac{1}{\sqrt{h}} \psi(\frac{t-\tau}{h}) \ , \ h \in \mathbb{R}^+, \ \tau \in \mathbb{R} \tag{2.4}$$

L'analyse d'une fonction f par l'ondelette ψ consiste alors à calculer les coefficients $w_f(h,\tau)$, résultant de la projection de f sur la fonction $\psi_{h,\tau}$.

$$w_f(h,\tau) = \int_{\mathbb{R}} f(t) \overline{\psi_{h,\tau}(t)} dt \tag{2.5}$$

où $\overline{\psi_{h,\tau}(t)}$ désigne le complexe conjuguée de $\psi_{h,\tau}(t)$. Ainsi, l'analyse par ondelettes donne une représentation des signaux localisée à la fois en temps et en fréquence. En effet, en faisant varier τ, l'intervalle où l'ondelette est non nulle "se déplace" et l'analyse se fait dans une fenêtre temporelle de la taille de l'intervalle précédent et centrée en τ. De même, en faisant varier h, la période temporelle et donc la fréquence de l'ondelette varie. Plus la fréquence de l'ondelette sera proche de celle du signal dans la fenêtre temporelle sélectionnée, plus le coefficient $w_f(h,\tau)$ sera élevé. Dans de nombreuses situations, on se limite au cas dyadique en posant :

$$h = 2^k \ , \ \tau = l2^k \ , \ (k,l) \in \mathbb{Z}^2 \tag{2.6}$$

Cette transformation est inversible sous les conditions d'admissibilité de l'équation (2.3). C'est le cas par exemple des bases d'ondelettes orthogonales et biorthogonales. Le signal est alors reconstruit par :

$$f(t) = \sum_{k \in \mathbb{Z}} \sum_{l \in \mathbb{Z}} w_f(2^k, l2^k) \psi_{2^k, l2^k}(t) \tag{2.7}$$

Dans les applications de compression des signaux, les ondelettes utilisées sont orthogonales ou biorthogonales. En effet, en utilisant une base d'ondelettes orthogonale normalisée, la transformée en ondelettes résultante préserve l'énergie. Les ondelettes biorthogonales ne

satisfont que quelques propriétés d'orthogonalité. Elles sont construites de telle façon que deux bases, la directe $B = \{e_i\}$ et la duale $\tilde{B} = \{\tilde{e}_i\}$ satisfassent la condition de dualité $(e_i, \tilde{e}_j) = \delta_{ij}$ et servent pour l'analyse et la synthèse respectivement. A ce titre, elles ne permettent pas la conservation d'énergie. Cependant certaines ondelettes biorthogonales sont quasiment orthogonales et n'introduisent qu'un faible changement dans l'énergie. Ainsi JPEG2000 utilise des bases d'ondelettes biorthogonales. Un tel choix est motivé par le calcul pratique de la TOD détaillé dans la suite.

La Transformée en Ondelettes Discrète

Effectuer l'analyse par ondelettes d'une fonction ou d'une image revient à en faire une analyse multirésolution. Partant de cette constatation, Mallat [Mal98] a montré que l'analyse par ondelettes pouvait être calculée par l'utilisation de bancs de filtres. L'objet de ce paragraphe est de présenter succinctement ce calcul des ondelettes par filtrage. Nous présentons ce processus dans le cas des images discrètes à deux dimensions.

Le banc de filtres utilisé ici est l'association de deux filtres à une dimension : un filtre passe-bas (L_a) et un filtre passe-haut (H_a). Pour obtenir les coefficients de la transformation en ondelettes, ces filtres sont appliqués séparément sur les lignes et les colonnes de l'image suivant l'ordre indiqué dans la figure 2.4. Rappelons que le filtrage d'un vecteur à une dimension (ici chaque ligne et chaque colonne de l'image) est le résultat de la convolution de ce vecteur avec la réponse impulsionnelle discrète et finie du filtre. Ainsi, par exemple, la convolution $\hat{I}_l$ d'une ligne I_l de l'image avec le filtre passe-haut H de longueur T est donnée par :

$$\hat{I}_l = I_l * H \text{ avec } \hat{I}_l(t) = \sum_{\tau=-T/2}^{T/2} I_l(t-\tau)H(\tau) \tag{2.8}$$

D'abord, les lignes de l'image sont traitées par le filtre passe-bas L_a et le filtre passe-haut H_a. Cela résulte en un nombre de coefficients deux fois plus importants que dans l'image originale. Afin de limiter cette expansion des coefficients, le résultat de chaque filtrage est sous-échantillonné d'un facteur 2 (indiqué par $\downarrow 2$ sur la figure 2.4). Ensuite, les colonnes des deux images résultantes sont traitées elles aussi par le filtre passe-bas L_a et le filtre passe-haut H_a puis sous-échantillonnées. Cette combinaison de filtrages résulte en quatre images (ou sous-bandes) notées LL^1, LH^1, HL^1 et HH^1 qui forment le premier niveau de la décomposition en ondelettes. Un niveau supplémentaire de décomposition peut être obtenu en appliquant le traitement précédent sur la sous-bande LL^1. Typiquement, le processus est répété Kfois, résultant en un ensemble d'images $W_t = \{LL^0, LL^k, LH^k, HL^k, HH^k, k = 1...K\}$ où $k$désigne le niveau de décomposition. W_t est la trame d'ondelettes à l'instant t correspondant à l'image I_t prise à l'instant t d'une vidéo. De même $W_t^k = \{LL^k, LH^k, HL^k, HH^k\}$désignera le niveau $k > 0$ de la décomposition en ondelettes. Ainsi, le "bas" de la pyramide, pour $k = 0$, est l'image à pleine résolution qui

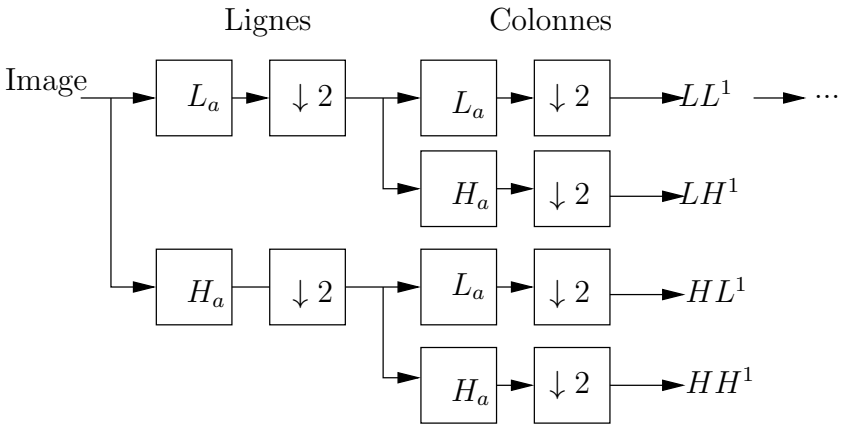
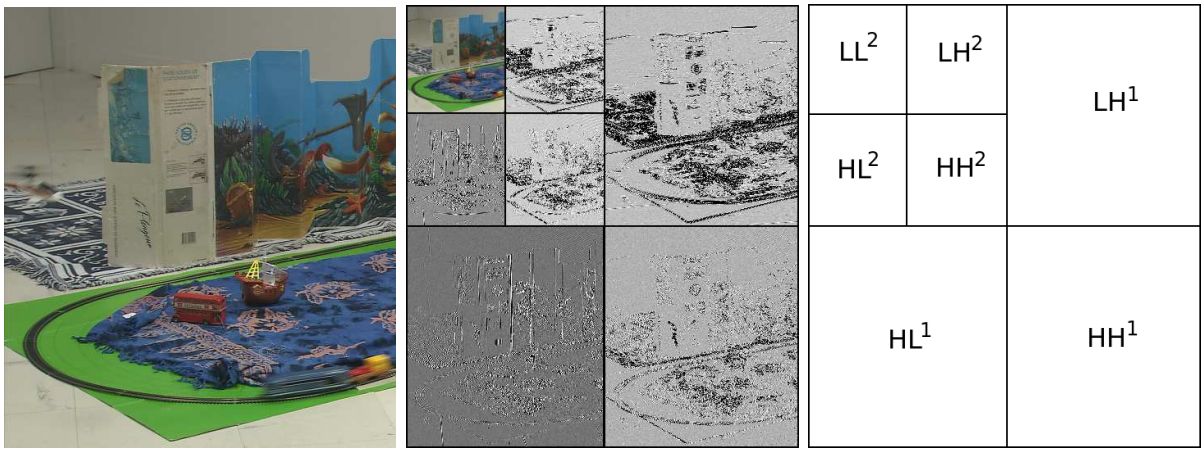


Fig. 2.4 – TOD : Analyse

est assimilée à la sous-bande LL^0. Dans ce cas $W_t^0 = \{LL^0\}$ car aucune décomposition en ondelettes n'a eu lieu. Le "haut" de la pyramide, pour $k = K$, correspond à la résolution d'image minimale. La trame d'ondelettes désigne collectivement les 4 sous-bandes de la décomposition en ondelettes. Quand il ne sera pas nécessaire de préciser le niveau de la sous-bande dont on parle, celle-ci sera désignée seulement par son type : LL, LH, HL ou HH. Enfin, le terme sous-bande de Basse Fréquence (BF) désigne l'ensemble $\{LL\}$ et celui de sous-bande de Haute Fréquence (HF) l'ensemble $\{LH, HL, HH\}$.
Un exemple de résultat de décomposition en ondelettes sur une image est donné dans la figure 2.5(b), l'image originale ou composante LL^0 associée étant montré figure 2.5(a). L'agencement des sous-bandes sur une tel affichage est précisé dans la figure 2.5(c). Les images des sous-bandes de HF ont été post-traitées indépendamment les unes des autres afin d'en rendre les détails visibles. Par construction, la sous-bande de type LL est une version réduite de l'image originale et les sous-bandes LH, HL et HH contiennent respectivement des informations sur les contours horizontaux, verticaux et diagonaux de l'image. On parle aussi d'analyse par ondelettes. Nous reviendrons plus en détail sur l'utilité d'une telle représentation pour notre approche dans la suite de ce paragraphe.

La transformation inverse (ou étape de synthèse) s'implémente elle aussi à l'aide d'un banc de filtres, composé d'un filtre passe-bas (L_s) et d'un filtre passe-haut (H_s). Le processus de reconstruction d'une image à partir de l'ensemble des sous-bandes de niveau $k = 1$ est illustré dans la figure 2.6.

Le schéma de synthèse est le strict schéma inverse de l'analyse. Les filtrages sont d'abord effectués sur les colonnes puis les lignes. Avant chaque étape, un sur-échantillonnage (indiqué par $\uparrow 2$ sur la figure 2.6) précède l'application du filtre afin de ramener toutes les sous-bandes à la résolution de l'image à reconstruire.



(a) (b) (c)

Fig. 2.5 – Illustration de la décomposition en ondelettes sur une image issue de la séquence “train”, LaBRI. (a) Image Originale (b) TOD sur 2 niveaux (c) Notations.

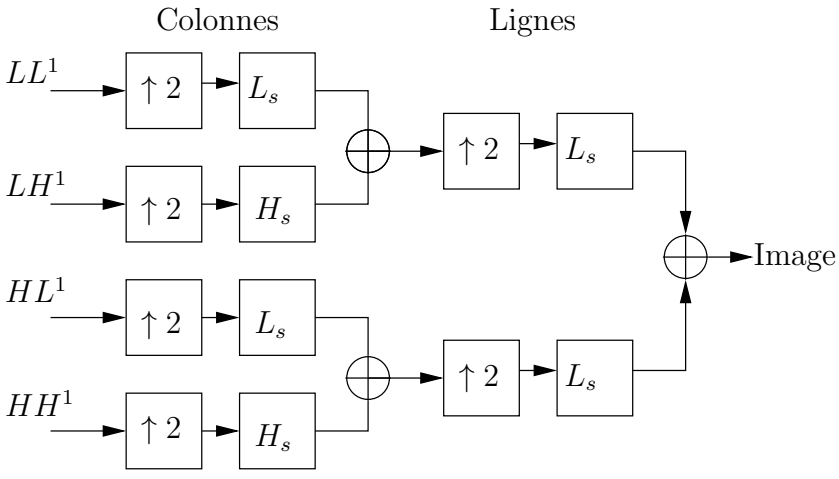


Fig. 2.6 – TOD : Synthèse

Filtres d'Analyse		
i	Filtre Passe-Bas L_a	Filtre Passe-Haut H_a
0	0.6029490182363579	1.115087052456994
± 1	0.2668641184428723	-0.591271763114247
± 2	-0.07822326652898785	-0.05754352622849957
± 3	-0.01686411844287495	0.09127176311424948
± 4	0.02674875741080976	
Filtres de Synthèse		
i	Filtre Passe-Bas L_s	Filtre Passe-Haut H_s
0	1.115087052456994	0.6029490182363579
± 1	-0.591271763114247	0.2668641184428723
± 2	-0.05754352622849957	-0.07822326652898785
± 3	0.09127176311424948	-0.01686411844287495
± 4		0.02674875741080976

Tab. 2.1 – Coefficients des filtres des ondelettes 9/7 de Daubechies (haut) filtres d'analyse (bas) filtres de synthèse

Utilisation de la TOD dans le standard JPEG2000

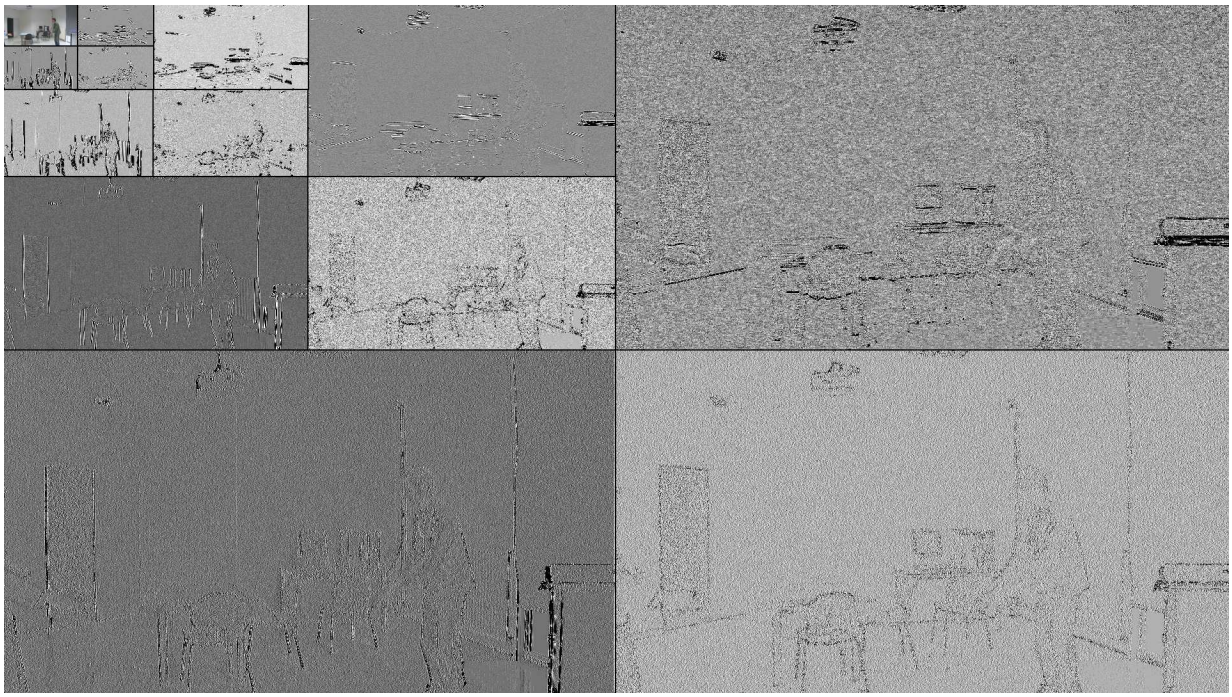
Les paragraphes précédents nous ont permis d'établir le cadre théorique des ondelettes. Nous allons indiquer ici comment ce cadre est utilisé dans le standard JPEG2000.
Dans le cas de la compression avec pertes, les ondelettes utilisées sont les ondelettes 9/7 de Daubechies [Ant92]. Ces ondelettes sont biorthogonales. Les raisons d'un tel choix sont liées au calcul de la TOD par bancs de filtres. Pour utiliser cet algorithme, il faut que la base d'ondelettes puisse s'écrire sous la forme de filtres à phase linéaire. Or une seule ondelette orthogonale possède cette propriété : l'ondelette de Haar (cf. figure 2.3(a)). Cette ondelette est peu intéressante pour la compression. C'est pourquoi des ondelettes biorthogonales, qui peuvent s'écrire sous la forme de filtres à phase linéaire, ont été définies. L'ondelette 9/7 de Daubechies est, par exemple, quasi-orthogonale. Les coefficients des filtres 9/7 pour l'analyse et la synthèse sont donnés dans les tableaux 2.1.

Les ondelettes sont donc calculées par filtrage. Nous avons expliqué comment calculer ce filtrage par convolution. Il existe néanmoins une autre manière d'effectuer ce calcul appelée "lifting". C'est cette deuxième technique qui est utilisée en pratique dans le compresseur JPEG2000 car elle est plus rapide en termes de temps de calcul. Nous avons décidé de présenter la solution par convolution car nous trouvons qu'elle permet de mieux comprendre la nature physique des sous-bandes BF et HF. Les valeurs des coefficients obtenues par les deux méthodes sont bien sûr identiques. Outre les ouvrages de référence cités en début de paragraphe, le lecteur peut trouver une description de la méthode de "lifting" dans l'article

de Sweldens [Swe95].
Ajoutons ici une remarque qui prendra toute son importance lorsque nous détaillerons nos méthodes d'extraction d'objet et d'indexation dans le domaine des ondelettes calculées par JPEG2000. Les sous-bandes LL^k, $k = 0...K-1$, peuvent être synthétisées à partir des sous-bandes de HF et de la sous-bande LL^K. Il n'est donc pas nécessaire de les encoder dans un flux compressé tel que JPEG2000. Ainsi, lorsque nous voudrons accéder aux informations contenues dans les sous-bandes LL^k, il nous faudra procéder préalablement à leur synthèse.

Propriétés des ondelettes
Avant de continuer la description du standard JPEG2000 dans les paragraphes suivants, nous présentons ici quelques propriétés des ondelettes qui sont utilisées par nos méthodes d'extraction d'objets et d'indexation.
Comme nous l'avons évoqué, chaque sous-bande d'un niveau de décomposition de la TOD contient des informations spécifiques. La sous-bande de type LL est une version réduite de l'image originale. A ce titre, elle possède des caractéristiques similaires à une image et les algorithmes d'extraction d'objets dans le domaine pixel existants peuvent lui être appliqués à de petits ajustements près. Les sous-bandes de HF contiennent les informations de contour, chaque sous-bande ayant une direction privilégiée : LH regroupe les informations sur les contours horizontaux, HL celles sur les contours verticaux et HH décrit les contours diagonaux de l'image. Nous utilisons ici le terme contour au sens large. Il désigne aussi bien les contours des objets que les directions privilégiées des textures.
L'information contenue dans les sous-bandes est modifiée suivant le niveau de décomposition considéré. A Haute Résolution (HR), c'est-à-dire pour k proche de 1, les sous-bandes de HF décrivent en plus le bruit dans l'image. Notons que ce bruit est plus présent dans la sous-bande HH que dans les sous-bandes LH et HL (cf. figure 2.7). En effet, l'utilisation du filtrage passe-bas sur les lignes (resp les colonnes) atténue le bruit dans la sous-bande HL (resp LH) alors qu'aucun filtrage passe-bas n'est appliqué pour l'obtention de la sous-bande HH. A Basse Résolution (BR), c'est-à-dire pour k proche de K (typiquement $K = 4$ ou 5), l'image à décomposer est une sous-bande LL qui est issue de plusieurs filtrages passe-bas. Le bruit présent dans l'image en est d'autant réduit. Les détails de textures ont eux aussi été gommés. L'information que contient les sous-bandes de HF est alors majoritairement l'information de contours des objets.
La transformée en ondelettes permet à la fois la localisation en espace et en fréquence. Ainsi, un lieu dans une sous-bande de la décomposition en ondelettes a un correspondant spatial dans chacune des autres sous-bandes de la décomposition (cf. figure 2.8). Tous ces lieux de sous-bandes sont à l'origine localisés au même endroit dans l'image non transformée. Du fait des changements de résolution, un lieu de taille 1 pixel correspond à un lieu de taille 4 pixels au niveau de résolution immédiatement supérieure. Cette propriété de localisation est un des fondements de la propriété de scalabilité en résolution des ondelettes

(a)



(b)

Fig. 2.7 – (a) Image extraite de la séquence “lancer_trousse” et (b) sa TOD sur $K = 4$ niveaux de décomposition.

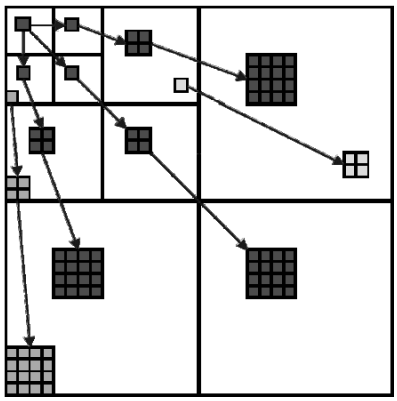


Fig. 2.8 – Principe de localisation des ondelettes. Un pixel donné est projeté sur quatre pixels au niveau de résolution immédiatement supérieur.

et par conséquent du standard JPEG2000. Nous l'exploiterons lors de l'extraction d'objets et de l'indexation, elle nous permet en effet de mettre en correspondance les contours des objets extraits sur toutes les sous-bandes.

2.5 Quantification

L'étape suivante du processus d'encodage JPEG2000 est la quantification. Jusqu'à présent, les étapes de pré-traitement et la TOD ont transformé les données mais n'ont pas réduit la quantité de données à encoder. La réduction des données est l'objet de la quantification. Le but est d'associer à chaque coefficient un entier appartenant à un ensemble fini. Deux valeurs de coefficients très proches l'une de l'autre peuvent alors être associées au même entier. Cette étape conduit à une perte d'information.
La quantification utilisée dans JPEG2000 est une quantification scalaire uniforme avec zone morte. La règle associée est donnée par :

$$q = sign(w)\lfloor\frac{|w|}{\Delta}\rfloor \tag{2.9}$$

avec w un coefficient d'ondelette à quantifier, $sign(w)$ le signe de w, Δ le pas de quantification choisi et q l'indice de quantification résultant. L'opération $\lfloor x \rfloor$ arrondi l'argument x à la valeur entière inférieure la plus proche. Le qualificatif de "zone morte" indique que l'intervalle des valeurs quantifiées par la valeur 0 est de taille 2Δ, c'est-à-dire deux fois supérieure à la taille des autres intervalles de valeurs quantifiées par un entier non nul. Une telle approche permet de favoriser l'apparition de zéros. Cette propriété est désirable pour la compression. Coder des valeurs nulles est en effet très peu couteux en bits ; plus il y a de zéros, plus le taux de compression est grand. On voit apparaître ici l'intérêt d'avoir appliqué la TOD avant cette étape. En effet, les sous-bandes de HF d'une décomposition

en ondelettes ont beaucoup de coefficients proches de zéros dans les zones plates de l'image d'origine.
Le choix du pas de quantification Δ est crucial. Dans le schéma de compression JPEG2000, un pas de quantification Δ_{SB} différent est choisi pour chaque sous-bande de la décomposition en ondelettes. Dans la partie I du standard, il est calculé par :

$$\Delta_{SB} = 2^{\epsilon_{SB}}(1 + \frac{\mu_{SB}}{2^{11}}) \tag{2.10}$$

où ϵ_{SB} et μ_{SB} sont des entiers positifs tels que :

$$0 \leqslant \epsilon_{SB} < 2^5 \text{ et } 0 \leqslant \mu_{SB} < 2^{11} \tag{2.11}$$

Les valeurs de ϵ_{SB} et μ_{SB} sont choisies en fonction de la dynamique des coefficients de la sous-bande. Le choix du pas de quantification peut être fait a posteriori lors de la compression par l'algorithme EBCOT suivant l'importance accordée à la sous-bande associée. Nous ne détaillerons pas plus le choix du pas de quantification qui ne nous intéresse pas directement pour construire nos méthodes d'extraction d'objets et d'indexation qui se placent après la déquantification et s'adaptent aux données fournies. Notons que ce pas de quantification peut être défini par l'utilisateur pour tenir compte des phénomènes de masquage du système visuel humain comme, par exemple, la fonction de sensibilité au contraste [Sto05].
La valeur de reconstruction $\hat{w}$ associée à l'indice de quantification q est assignée par l'opération de déquantification par

$$\hat{w} = \begin{cases} 0 & \text{si } q = 0 \\ (q + \delta_q)\Delta & \text{sinon} \end{cases} \tag{2.12}$$

où δ_q est un paramètre de reconstruction choisi par le décodeur, $\delta_q = \frac{1}{2}$ correspond à la reconstruction au point median de l'intervalle.

2.6 Codage Entropique

Cette étape de la compression détermine la quantité de bits à utiliser pour encoder chaque coefficient. Le but est évidemment d'utiliser le moins de bits possible, afin que la taille du fichier de sortie soit la moins grande possible. L'algorithme d'encodage utilisé dans le standard JPEG2000 est l'EBCOT ("Embedded Block Coding with Optimised Truncation").
Le codeur EBCOT se décompose en deux étages, l'étage 1 est chargé de la modélisation du contexte et du codage entropique alors que l'étage 2 génère le flux de sortie et assure l'allocation des bits de sortie. Dans *l'étage 1* de l'encodeur, chaque sous-bande à encoder est partitionnée en blocs relativement petits (par exemple 64x64 ou 32x32 échantillons) appelés "code-blocks". Chaque "code-block" est codé indépendamment des autres. Ce codage

indépendant assure une meilleure robustesse aux erreurs. Les indices de quantification ne sont pas codés au niveau du symbole mais au niveau du plan de bits. Pour cela, un codeur arithmétique adaptatif au context est utilisé. Il agit en trois passes : propagation de la significance, affinage de l'amplitude et nettoyage. Dans *l'étage 2*, la décision d'inclure ou non un plan de bits est prise. A cette étape, l'ordre d'assemblage de ces plans de bits est décidé afin de former le message (ou train binaire) final ayant le débit visé. Pour ce faire, les plans de bits sont arrangés en couches de qualité. Une fois que l'image a été compressée, une opération de post-traitement de contrôle du débit est effectuée. Elle vise à altérer le train binaire associé à chaque "code-block" de façon à ce que le taux d'échantillonage désiré du train binaire final soit atteint.
L'utilisation des "code-blocks" permet une organisation du train binaire flexible. Ces "code-blocks" sont regroupés en paquets qui sont ensuites multiplexés et ordonnés pour former le train binaire. L'ordre dans lequel les paquets sont ordonnés est appelé progression. Les paquets sont rangés du plus significatif au moins significatif suivant la propriété mise en avant (par exemple qualité ou résolution). C'est cet arrangement qui assure la scalabilité effective du flux.
Nous ne détaillerons pas plus le codeur EBCOT. La présentation faite est volontairement courte pour ne pas surcharger ce manuscrit. De plus amples détails sur le codeur peuvent être trouvés dans l'ouvrage de Taubman et Marcellin [Tau02] déjà mentionné et dans l'article de Taubman [Tau00].

2.7 Conclusion

Le standard de compression JPEG2000 fait appel à deux outils originaux qui permettent la scalabilité du flux : le codeur EBCOT et la TOD. Le codeur EBCOT organise le train binaire d'une façon souple, en décomposant le flux binaire en petites entités décodables séparément. Le codage en plan de bits intégré permet de définir plusieurs couches de qualité qui assurent la scalabilité en qualité. La transformée en ondelettes permet de transformer l'image dans sa globalité et de réduire le nombre de coefficients significatifs à encoder. Le fait que cette transformée soit multirésolution assure la scalabilité en résolution du flux. D'autres scalabilités, que nous n'avons pas détaillées, sont aussi disponibles (par exemple, la scalabilité en localisation spatiale assurée d'une part par la TOD qui possède la propriété de localisation et d'autre part par le codeur EBCOT qui, avec la structure de blocs, permet de décoder certaines régions de l'image spécifiquement).
Dans notre approche, nous avons décidé de nous focaliser sur la TOD et la scalabilité en résolution. Une part importante du chapitre a ainsi été consacrée à la TOD. Les autres étapes de l'encodage JPEG2000 ont été évoquées mais non détaillées. Nous avons ainsi mis l'accent sur la nature des données que nous allons utiliser dans nos méthodes. Rappelons que

la TOD fournit une représentation multirésolution du signal. Chaque niveau est composé de sous-bandes qui fournissent des informations de BF et HF sur l'image. Les coefficients de ces sous-bandes sont liés entre eux par le principe de localisation des ondelettes. Les images traitées par la transformée en ondelettes sont codées au format YUV. De plus les valeurs des coefficients des images sont centrées en 0 (contrairement à l'habitude en image où les valeurs des pixels sont représentées dans l'intervalle $[0, 255]$). Dans la pratique, nous avons utilisé le logiciel Kakadu [Tau] pour assurer la compression par JPEG2000 et extraire les coefficients d'ondelettes après déquantification. Les auteurs de ce logiciel, pour des questions de performances, ajoute une étape de pré-traitement (cf section 2.3) qui consiste à normaliser les valeurs des images avant la TOD. Cette normalisation est faite par division par un facteur constant de 2^B (cf section 2.3, décalage) et n'aura donc pas incidence pour l'application de nos algorithmes sur des flux compressés avec un autre logiciel (tels que ceux cités dans l'annexe A). Seul l'ordre de grandeur des valeurs changent et certains seuils définis expérimentalement devront être multiplié par 2^B pour être adaptables à des données non normalisées.

Chapitre 3

Etat de l'art en indexation vidéo et extraction d'objets

L'indexation automatique des vidéos par le contenu est un domaine de recherche très actif. Bien que de nombreuses techniques aient été proposées, le problème reste non entièrement résolu. La difficulté vient du "fossé sémantique" [Sme00] qui existe entre les informations dites de "bas-niveau" que l'on peut extraire des images de la vidéo et l'interprétation qu'un cerveau humain est capable de faire de ces mêmes images. L'indexation par le contenu se compose de la définition et l'extraction automatique d'une ou plusieurs caractéristiques de la vidéo et de la définition d'une mesure de similarité sur ces signes distinctifs. Ces caractéristiques, aussi appelées descripteurs, sont soit spatio-temporelles soit simplement spatiales si la vidéo est considérée comme un simple ensemble d'images. Dans les deux cas, elles peuvent être définies à différentes échelles spatiales au niveau de la trame : i) sur l'image entière (indexation globale) ou sur un sous-ensemble de l'image constitué soit ii) d'un ensemble de points caractéristiques (indexation locale) soit iii) de l'union de composantes connexes de l'image (indexation basée objet). Dans ce dernier cas (iii), une étape préliminaire d'extraction automatique des composantes connexes est nécessaire : il s'agit généralement d'extraire les objets de la scène. Ce problème est lui-même le sujet de nombreuses recherches. En effet, la segmentation est un problème mal posé. Ainsi, trouver les composantes connexes puis décider comment les grouper en objets par traitements automatiques sont des tâches complexes.

Dans la suite de ce chapitre, nous passons en revue quelques unes des techniques majeures existantes à ce jour d'indexation puis d'extraction d'objets. En nous appuyant sur cet état de l'art, la méthode d'indexation basée objet que nous proposons est présentée.

3.1 Indexation par le contenu des vidéos

Une séquence vidéo est une succession d'images, appelées trames, dont le contenu change progressivement et, la plupart du temps, est "lisse" au sens de la variation temporelle. C'est pourquoi l'indexation vidéo doit tenir compte de cette continuité et de l'ensemble des images présentes dans une séquence. Néanmoins, l'aspect spatial reste très significatif. De plus, beaucoup de solutions d'indexation des vidéos par le contenu proposent de considérer chaque séquence comme une collection d'images fixes - le plus souvent ramenée à un sous-ensemble d'images clés - et utilisent alors des techniques d'indexation d'images purement spatiales. C'est pourquoi nous proposons de classer les différents types d'indexation en trois groupes, suivant le domaine spatial utilisé dans chaque trame pour calculer le descripteur, que ces trames soient considérées à titre individuel ou comme faisant partie d'un ensemble :

- **Indexation globale** : le descripteur est calculé sur la trame entière
- **Indexation locale** : le descripteur est calculé sur un ensemble de points caractéristiques de la trame
- **Indexation basée objet** : le descripteur est calculé sur les objets d'intérêt de la trame.

De même qu'il existe différents domaines spatiaux de définition des descripteurs, il existe différents domaines de représentation liés aux domaines de représentation de l'image. Nous nous limitons dans la suite au domaine pixel (la représentation couleur n'étant pas limitée : RBG, YUV, HSV...) et au domaine transformé compressé (c'est-à-dire les transformations utilisées dans les standards de compression : TCD et TOD).
La recherche en indexation vidéo ayant donné lieu à de nombreux travaux, il n'est pas possible d'en rendre compte ici en intégralité. Nous nous limiterons à quelques exemples. Ainsi, nous n'évoquerons pas les techniques d'indexation d'images-clés purement spatiales qui n'ont pas eté étendues à un groupe d'images. En revanche, certaines techniques d'abord pensées pour les images puis étendues à la vidéo seront présentées. Une description plus complète du domaine peut être trouvée dans les articles [Man99b, Vel02, Sme00, Lew06, Ren09] et le livre [Han04]. L'indexation des vidéos a fait l'objet d'une normalisation au travers du standard MPEG7 [MPE01] dans la partie 3 [ISO02]. Une description détaillée de ces descripteurs peut être trouvée dans [**?**].

3.1.1 Indexation globale

Le principe de l'indexation globale est, à partir de caractéristiques extraites de l'image, de proposer une "signature" globale d'une image-clé de la séquence vidéo à indexer. Une telle signature peut être étendue et calculée sur l'ensemble des images de la séquence.

De nombreuses techniques de construction de descripteurs globaux ont été développées à ce jour. La suite de ce paragraphe présente les grands groupes formant ce type d'indexation.

Histogrammes couleurs. Une des premières signatures des séquences proposée dans la littérature est **l'histogramme couleur** des images ou des images-clés d'une vidéo. L'idée est de caractériser les images par la PDF des couleurs dont l'histogramme est une approximation. En pratique, ce dernier est défini dans un espace couleur donné en quantifiant les couleurs de cet espace et en relevant la fréquence d'apparition de chaque couleur quantifiée dans l'image. Ainsi, pour une image I, l'histogramme est défini par $H(I) = \{(C^m, card(C^m)), m = 1..M\}$ où M est le nombre de couleurs de quantification et $C^m = (C_1^m, C_2^m, C_3^m)$ désigne un vecteur couleur quantifié dans l'espace de représentation considéré (par exemple RGB, YUV, HSV,...). Cette signature est à l'origine de nombreux travaux. Ainsi, Gargi et al [Gar95] l'utilisent pour indexer les discontinuités dans les documents vidéo montés, ces discontinuités étant typiquement les frontières des plans de montage. Un tel changement de plans est détecté lorsque les histogrammes de deux images consécutives sont différents ; la mesure de similarité utilisée étant l'intersection d'histogrammes de Swain et Ballard [Swa91] (cf. section 6.1). Ce descripteur peut être étendu à un segment vidéo en utilisant les vecteurs couleurs de l'ensemble des images de la séquence pour calculer un seul histogramme. C'est ce que font Dumont et Merialdo [Dum09] pour retrouver les multiples prises d'un même plan dans des "rushes"[1] vidéo. Dans leur travaux, la vidéo est découpée en blocs de une seconde. L'histogramme couleur de l'ensemble des images de chacun de ces blocs est calculé dans le domaine HSV. L'originalité de l'approche réside dans le choix de la mesure de similarité par alignement des séquences. Une telle mesure indique le nombre de transformations élémentaires nécessaires pour transformer un bloc en un autre. L'algorithme d'alignement des séquences vidéo que les auteurs proposent est inspiré de l'algorithme classique en bioinformatique de Smith-Waterman [Smi81]. Une présentation plus détaillée de l'indexation par histogramme sera donnée dans le chapitre 6.
L'histogramme couleur est un descripteur très fréquemment utilisé. Il présente l'avantage d'être robuste aux transformations géométriques, aux changements d'échelle et aux changements de prises de vues modérés. Il est en revanche sensible aux changements d'illumination ; ce problème pouvant être résolu en recalant les histogrammes sur la même plage de définition avant de calculer la similarité.

Le principal problème du descripteur précédent est que deux trames de contenus différents peuvent avoir le même histogramme. Cela est dû au fait que la distribution spatiale des couleurs dans le plan-image n'est pas prise en compte. Plusieurs types d'approches permettent de tenir compte à la fois du contenu couleur et de la distribution spatiale associée.

[1] Ensemble des prises de vues après développement et avant montage des séquences

Nous allons les présenter dans la suite.

Variantes d'histogramme. L'idée des histogrammes a ainsi été reprise et normalisée dans le standard de description de contenu MPEG7 [MPE01] par le descripteur appelé **couleur dominante** ("dominant color"). Il est défini par $F = \{(c_i, p_i, v_i), s\}$, $i = 1, 2, ...N$ où N est le nombre de couleurs dominantes, c_i le vecteur de couleur, p_i la fraction de pixel dans l'image correspondant à cette couleur. Le paramètre v_i représente la variance des valeurs des couleurs des pixels dans une région autour de la couleur représentative correspondante. La cohérence spatiale s est un nombre unique qui représente l'homogénité spatiale globale des couleurs dominantes dans l'image. Le nombre de couleurs N peut varier d'une image à une autre et, dans la plupart des cas, $N = 8$ est suffisant pour décrire l'image. Le **corrélogramme** [Hua97] est une matrice à trois dimensions dont les éléments $\gamma(i, j, k)$ représentent la probabilité de trouver deux pixels dans l'image ayant les couleurs C_i et C_j placée à une distance de k l'une de l'autre. Cette technique a été étendue au domaine de la transformée en ondelettes de Gabor dans [Mog05]. Une autre approche consiste en l'utilisation des **vecteurs de cohérence de couleur** [Pas96]. Les pixels d'une couleur C_m donnée sont séparés en deux classes afin d'affiner l'histogramme. Les pixels cohérents, c'est-à-dire appartenant à une composante connexe de taille raisonnable (au moins 1% de l'image), sont distingués des pixels incohérents.

Outre les variantes d'histogrammes, il existe d'autres méthodes tenant compte à la fois du contenu couleur et de la distribution spatiale associée. Il s'agit par exemple de la mesure ordinale ou de la création de mosaïques 1D.

Mesure Ordinale. Afin de structurer l'information de couleur en régions, Bhat et Nayar [Bha98] ont proposé **la mesure ordinale** pour calculer les correspondances entre images. L'image-clé de la vidéo est partitionnée en N blocs ; ces blocs sont ensuite classés suivant leur niveau de gris moyen. La signature $S(t)$ (3.1) utilise le rang r_i de chaque bloc i.

$$S(t) = (r_1, r_2, ..., r_N) \tag{3.1}$$

La distance $D(t)$ (3.2) est définie pour calculer la similarité entre deux vidéos (une référence R et une candidate C) à l'instant t (T est la taille du segment considéré) :

$$D(t) = \frac{1}{T} \sum_{i=t-\frac{T}{2}}^{t=+\frac{T}{2}} |R(i) - C(i)| \tag{3.2}$$

Différentes études utilisent cette mesure ordinale [Hua04, Kim05, Ham01] et il a été prouvé qu'elle est robuste aux changements de résolution, décalages d'illumination et formats d'affichage. La mesure ordinale a été étendue au temporel par L. Chen et F. Stentiford [Che06]

en utilisant le rang des blocs au cours du temps. Si chaque trame est divisée en N blocs et si λ^n est la mesure ordinale du bloc n dans une fenêtre temporelle de longueur M, la dissimilarité D (3.3) entre une vidéo requête V_q et une vidéo de référence V_r au temps t est :

$$D(V_q, V_r) = \frac{1}{N} \sum_{n=1}^{N} d^p(\lambda_q^n, \lambda_r^n) \tag{3.3}$$

où

$$d^p(\lambda_q^n, \lambda_r^n) = \frac{1}{C_M} \sum_{i=1}^{M} |\lambda_q^n - \lambda_r^{n+ip}| \tag{3.4}$$

p est le décalage temporel testé et C_M un facteur de normalisation. Le meilleur décalage temporel p est sélectionné. Selon [Law07], la mesure ordinale présente le défaut d'être peu robuste au regard de l'insertion de logos, du décalage ou du rognage, qui sont des transformations très fréquentes dans les post-productions TV.

Mosaïque 1D. Une autre façon de prendre en compte l'information spatiale et temporelle dans la signature vidéo a été proposée dans [Cou99, Ben03]. Il s'agit de construire des **mosaïques spatio-temporelles compensées en mouvement** dans le domaine de la transformée de Radon discret. Pour chaque image d'une séquence vidéo, sa projection par transformée de Radon est calculée par :

$$R_\phi[f](u) = \int\int_D f(x,y)\delta(u - xsin(\phi) - ycos(\phi))dxdy \tag{3.5}$$

où $f(x, y)$ est la fonction, définie dans $\mathbb{R}^2$, des intensités des pixels de l'image, δ est la fonction de Dirac et ϕ est l'angle de projection. Ensuite, un modèle de mouvement à deux paramètres (zoom, translation) est calculé entre les projections de la même direction définie par l'angle ϕ. Finalement, toutes les images de la séquence sont compensées dans le repère de la projection de l'image de référence à l'instant t_0 et la mosaïque est calculée. Cette mosaïque représente une signature spatio-temporelle d'un plan-séquence vidéo. Les dictionnaires visuels à base de signatures de mosaïques ont été proposées dans [Ben01] par groupement des plans similaires et dans [Del04] pour la navigation dans des documents représentés par des sous-graphes.

Jusqu'à présent, aucune approche n'a pris en compte la cohérence temporelle qui existe entre les trames d'une vidéo. Nous allons maintenant présenter les descripteurs définis à partir du mouvement dans la séquence. L'information de mouvement est prise en compte en exploitant par exemple le mouvement d'objets visibles ou les mouvements de la caméra.

Activité temporelle globale. L'information globale de mouvement peut être représentée sous la forme d'une mesure de l'activité temporelle globale. Ainsi dans [Law07], les auteurs

définissent une activité temporelle globale $a(t)$ qui dépend de l'intensité I de chaque pixel (N étant le nombre de pixels dans chaque image) (3.6) :

$$a(t) = \sum_{i=1}^{N} K(i)(I(i,t) - I(i,t-1))^2 \tag{3.6}$$

où $K(i)$ est une fonction de poids qui permet d'accorder plus d'importance aux pixels du centre. Le descripteur est obtenu par analyse spectrale par TFD sur la séquence et conduit à un vecteur de dimension 16 fondé sur la phase de l'activité. Hampapur et al [Ham01] quant à eux proposent de quantifier les vecteurs de mouvement en Q classes. La signature de mouvement est constituée des histogrammes des vecteurs de mouvement sur ces Q classes calculés en chaque trame. Les clips requêtes et ceux de la base pouvant ne pas être de la même taille, la signature de test est convoluée le long du clip de référence et en chaque point la mesure de similarité utilisée est le coefficient de corrélation normalisé. L'instant de temps où la corrélation est maximum est alors la meilleure mise en correspondance. Parmi les travaux utilisant le mouvement, signalons le système VideoQ [Cha98], qui décrit diverses façons d'extraire des mouvements d'objets dans les vidéos et de formuler et traiter des requêtes portant sur ces mouvements.
Analyse du mouvement apparent. Une autre approche permettant de décrire les vidéos par le mouvement est de décomposer le flot optique sur des fonctions de bases ; les coefficients de la décomposition servant de descripteur. Ainsi, Augereau et al [Aug05] proposent de décomposer le flux optique couleur sur une base polynomiale. Cette description a été utilisée pour le groupement des séquences similaires dans la tâche de résumé vidéo [Que08]. De la même manière, Bruno et Pellerin [Bru02] proposent d'utiliser la décomposition en ondelettes.

Les méthodes précédentes sont les courants dominants qui existent actuellement en indexation globale. De nombreux travaux de recherche, reprenant ces principes de bases, sont disponibles dans la littérature. Il serait trop long de les détailler ici. D'autres techniques ont été développées, qui se basent sur les descripteurs de forme, la caractérisation des textures (matrices de co-occurrence [Har79] par exemple), les fonctions de hachage [Oos02, Cos06].

Enfin, terminons cette section par la présentation de méthodes globales développées dans le domaine des ondelettes. Wang et al [Wan98] proposent d'utiliser les coefficients d'ondelettes comme descripteurs des images. Chaque image est d'abord sous-échantillonnée à la taille 128x128 puis convertie du domaine couleur RGB à un domaine couleur repré-

sentant l'intensité et les contrastes perçus. La conversion se fait par :

$$\begin{cases} C_1 = (R+G+B)/3 \\ C_2 = (R+(255-B))/2 \\ C_3 = (R+2*(255-G)+B)/4 \end{cases} \tag{3.7}$$

Chaque composante couleur de l'image est décomposée sur $K = 5$ niveaux de décomposition en utilisant les ondelettes 8 de Daubechies. Le descripteur est alors constitué de l'ensemble des coefficients d'ondelettes W^5 et W^4 et de la variance des coefficients d'ondelettes LL^4 pour les trois composantes couleurs. L'idée est qu'une image ayant une variance faible, c'est-à-dire homogène en couleur, a peu de chances d'avoir la même signification sémantique qu'une image avec de grandes variations, c'est-à-dire une variance élevée. La recherche dans la base de donnée se fait de façon scalable. D'abord, les variances sont comparées. Cela conduit à un premier tri des réponses possibles. Ensuite, pour les images qui ont passé le premier test avec succès, les coefficients W^5 sont comparés. Le résultat est ensuite affiné sur les images restantes en comparant les coefficients W^4. La mesure de distance utilisée pour les comparaisons est la distance euclidienne.

Un des premiers travaux utilisant la transformée en ondelettes pour faire de l'indexation vidéo est proposé par Wen et al [Wen99]. Ils proposent de calculer la transformée en ondelettes de Haar 3D (filtrage sur les lignes puis les colonnes puis le long de l'axe temporel) sur des blocs de taille 16x16x16 de la vidéo. Le calcul du descripteur se fait sur ces données. Notons que ce découpage en blocs ne nous semble pas satisfaisant. En effet, la particularité de la transformation en ondelettes est qu'elle peut être appliquée sur l'image entière tout en fournissant une localisation spatiale. Cette approche ne travaille pas dans le domaine transformé ondelettes et n'est pas intéressante pour nous.

3.1.2 Indexation locale

Bien que performante, l'approche globale ne peut répondre de façon absolue au problème d'indexation. Elle souffre d'un manque de robustesse aux changements locaux d'intensité et de couleur par exemple. L'idée est alors de n'utiliser que les points de la vidéo qui sont hautement invariants. C'est le principe de l'indexation locale. Une signature de la vidéo est construite à partir des caractéristiques du voisinage local de points d'intérêt spatiaux ou spatio-temporels. Deux étapes sont donc nécessaires : l'extraction des points d'intérêt et la définition de la signature sur ces points.

L'utilisation des points d'intérêt pour la mise en correspondance d'images remonte aux années 1980 avec les travaux de Moravec [Mor81]. L'indexation locale est vraiment

devenue populaire avec l'apparition des **points SIFT** (Scale Invariant Feature Transform) [Low99, Low04]. Dans un premier temps, les points candidats sont déterminés. Un ensemble de filtres DoG (Différences de Gaussiennes) multirésolution est appliqué sur l'image. Les maxima locaux, en espace et en résolution, sont conservés. Une telle sélection permet de garantir l'invariance des points au changement d'échelle des images. Parmi les points trouvés, seuls ceux de fort contraste (et donc moins sensibles au bruit) et bien localisés aux extrémités des contours sont gardés. Enfin, une orientation est assignée à chacun des points sélectionnés. Cela permet de rendre le descripteur robuste à la rotation. Pour chaque pixel dans le voisinage du point d'intérêt, l'amplitude et la direction du gradient sont calculées. Un histogramme d'orientation est alors créé avec 36 classes, chaque classe couvrant un angle de 10°. Les pics dans l'histogramme correspondent à l'orientation dominante et sont assignés au point d'intérêt. Une fois les points détectés, le descripteur associé à chaque point peut être construit. Il est calculé comme étant un ensemble d'histogrammes d'orientations. Les orientations du gradient d'intensité sont tournées relativement à l'orientation du descripteur trouvée précédemment. Les histogrammes sont calculés sur des blocs voisins de taille 4x4 et quantifiés en 8 classes. 16 blocs de voisinage sont considérés, ce qui conduit à un descripteur de taille 128. Des travaux postérieurs ont proposé de sélectionner les points d'intérêts sur les régions dites MSER ("Maximally Stable Extremal Regions") [Mat02] et de les décrire de la même façon que les points SIFT [Liu08, For07]. Les MSER sont des régions qui sont plus sombres ou plus claires que leur voisinage et qui sont stables par rapport au seuillage de la fonction d'intensité.
La distance entre deux points SIFT est calculée par la mesure euclidienne. La mesure de similarité entre images et/ou vidéos se fait comme suit. Les points des deux images/vidéos à comparer sont appariés en utilisant la mesure de distance entre points. Si cette mesure est inférieure à un seuil donné, les points sont considérés comme identiques. La similarité entre deux images/vidéos est donnée par un algorithme standard de vote qui compte le nombre de descripteurs qui ont été appariés, sous l'hypothèse d'unicité de l'appariement, c'est-à-dire qu'un descripteur ne peut être mis en correspondance qu'une seule fois.
Dans le cas de la vidéo, une méthode a été proposée par Shi et Tomasi [Shi94] pour le suivi des points caractéristiques portant le nom de ScavFT (Scale Variant Feature Transform). Ces points caractéristiques sont extraits en se basant sur les valeurs propres de la matrice des gradients (3.8)

$$Z = \begin{pmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{pmatrix} \tag{3.8}$$

Les pixels où les valeurs propres de la matrice Z sont fortes et de même ordre de grandeur représentent des points caractéristiques avec un bon contraste qui correspondent aux coins et textures fortes faciles à suivre dans le temps. Le suivi de points d'intérêt est une tâche difficile. Des travaux d'amélioration ont été proposés tels que [Gou06].

La prise en compte de la cohérence spatiale dans la détection de points d'intérêt est le sujet de nombreuses recherches. Dans [Jol07], les caractéristiques sont extraites seulement sur les images-clés. Les points d'intérêt sont déterminés à l'aide d'une version améliorée du détecteur de Harris [Har88]. A chaque point est associé une description différentielle de la région locale alentours. La caractéristique locale différentielle est un vecteur $\vec{S}$ de dimension 20, $\vec{S} \in [0, 255]^{20}$, et est définie par

$$\vec{S} = (\frac{\vec{s_1}}{\|\vec{s_1}\|}, \frac{\vec{s_2}}{\|\vec{s_2}\|}, \frac{\vec{s_3}}{\|\vec{s_3}\|}, \frac{\vec{s_4}}{\|\vec{s_4}\|}) \tag{3.9}$$

$\vec{s_i}$ correspond à un vecteur de dimension 5 calculé à la position i qui est une position située dans un voisinage spatio-temporel du point d'intérêt. Chaque $\vec{s_i}$ est la décomposition différentielle du niveau de gris du signal d'intensité 2D $\vec{I}(x, y)$ au 2nd ordre :

$$\vec{s_i} = (\frac{\partial \vec{I}}{\partial x}, \frac{\partial \vec{I}}{\partial y}, \frac{\partial^2 \vec{I}}{\partial x \partial y}, \frac{\partial^2 \vec{I}}{\partial x^2}, \frac{\partial^2 \vec{I}}{\partial y^2}) \tag{3.10}$$

Le système de détection de copies **ViCopT** [Law06] (Video Copy Tracking) utilise une description similaire à la méthode précédente. Les points d'intérêt de Harris sont extraits sur toutes les trames de la vidéo. La différence est que la décomposition différentielle du signal en niveaux de gris est calculée pour 4 positions spatiales, c'est-à-dire dans la même trame. Ces points d'intérêt sont associés de trame en trame pour construire des trajectoires avec un algorithme similaire au suivi de points Kanade-Lucas-Tomasi [Tom91]. Pour chaque trajectoire, la description du signal conservée est la moyenne de chaque composante de la description locale. En utilisant les propriétés de la trajectoire ainsi construite, une étiquette de comportement peut être assignée à la description locale correspondante. Pour les besoins de la Détection de Copies par le Contenu, deux étiquettes particulières sont choisies : fond (points persistants et sans mouvement au fil des trames) et Mouvement (points persistants qui bougent). La signature finale de chaque trajectoire est composée d'un vecteur de dimension 20 similaire à (3.9), des propriétés de la trajectoire et de l'étiquette de comportement.

Laptev et Lindenberg [Lap03] ont proposé la technique des STIP (Space Time Interest Points) pour détecter les évènements spatio-temporels. Les points d'intérêt spatio-temporels correspondent à des points où les valeurs de l'image présentent des variations locales significatives dans les deux domaines espace et temps. L'application première de ce détecteur concerne la classification d'activités humaines et la détection de mouvements périodiques. Les points d'intérêt spatio-temporels sont décrits par un jet local au 3ème ordre spatio-temporel qui conduit à un vecteur j de dimension 34 :

$$j = (L_x, L_y, L_t, L_{xx}, ..., L_{ttt}) \tag{3.11}$$

où L_{x^m,y^n,t^k} sont les dérivées normalisées Gaussiennes Spatio-temporelles à l'échelle de détection spatiale σ et temporelle τ.

$$L_{x^m,y^n,t^k} = \sigma^{m+n}\tau^k(\partial_{x^m,y^n,t^k}g) * f \tag{3.12}$$

La distance L_2 est utilisée comme métrique de comparaison des signatures locales.

Citons enfin le descripteur SURF ("Speeded Up Robust Features") [Bay06] qui décrit une distribution des réponses aux ondelettes de Haar dans le voisinage du point d'intérêt.

Dans le domaine des ondelettes, qui est notre centre d'intérêt, le choix des points caractéristiques a été proposé dans [Seb01]. Le principe de la méthode consiste en l'extraction des points caractéristiques "où il se passe quelque chose à tous les niveaux de la décomposition en ondelettes". Ainsi il s'agit de considérer tous les coefficients dans des sous-bandes de Haute Fréquence (HF) du niveau le plus élevé de la décomposition. Ensuite, en utilisant le principe de localisation des ondelettes, tous les coefficients correspondants dans des sous-bandes du niveau inférieur de décomposition sont récupérés. A chaque niveau, le coefficient d'amplitude maximale est sélectionné. La procédure est répétée récursivement jusqu'à atteindre la pleine résolution. De cette façon, chaque coefficient de niveau de la résolution minimale est "tracé". Pour chaque pixel dans l'image d'origine on additionne les coefficients déterminant ainsi une valeur de "saillance" du pixel. La matrice des valeurs de saillance est ensuite seuillée et les pixels les plus "saillants" sont sélectionnés. Les résultats expérimentaux montrent que les ondelettes 4 de Daubechies permettent d'obtenir une meilleure carte de saillance correspondant aux contours significatifs que les ondelettes de Haar. Par ailleurs, la détection de ces points caractéristiques est plus robuste que celle des coins de Harris. Une fois les points caractéristiques détectés, comme dans le cas de SIFT, les descripteurs locaux dans le voisinage de ces points peuvent être calculés : moments couleur, moments des ondelettes, etc.

Une analyse assez complète de ces différents descripteurs et points caractéristiques est présentée dans [Mik04]. Ces approches semblent avoir leurs limites. Tout d'abord, la persistance des points caractéristiques n'est pas garantie dans une vidéo, car, à la différence de l'image statique représentant la même scène, les phénomènes d'occultation masquent des points caractéristiques identifiés dans les images précédentes. Ainsi des points identifiés sur le "fond d'une scène" risquent d'être masqués par l'objet en mouvement. Par ailleurs, si la recherche dans la vidéo se fait par l'identification des points caractéristiques dans une zone d'intérêt, les difficultés d'identification de ces derniers par rapport à la prise de vue "plan éloigné" ou "rapproché" peuvent être observées [Bag07].

3.1.3 Indexation basée objets

Les deux types d'indexation précédents travaillent à un bas niveau sémantique. Le problème de l'indexation globale est qu'elle travaille sur tous les pixels sans exception et est par là-même redondante. L'indexation locale tente de résoudre ce problème en ne calculant le descripteur que sur des points d'intérêt. Bien que performants, ces descripteurs se révèlent peu stables et très peu porteurs d'information sémantique. La recherche s'est alors intéressée à l'utilisation des objets d'intérêt. Une étude menée par Sav et al. [Sav06] montre que lorsque les utilisateurs en ont la possibilité, ils utilisent la description basée objet. Ce genre d'étude montre bien l'importance des objets et l'aide qu'ils apportent pour combler le "fossé sémantique".

Trajectoire de l'objet. Hsu et Teng [Hsu02] proposent un indice fondé sur la modélisation de la trajectoire de l'objet en mouvement. L'objet en mouvement est défini dans la première trame de la séquence par l'utilisateur qui détermine un rectangle englobant. Ce rectangle est ensuite transmis automatiquement aux autres trames de la séquence par compensation de mouvement. La trajectoire est alors un ensemble de points $\{(x_i, y_i), i = 0, ..., n-1\}$, où (x_i, y_i) est le centre de l'objet en mouvement dans la $i$ème trame. Le descripteur proposé utilise trois courbes polynomiales pour représenter séparément la trajectoire : le mouvement vertical, horizontal et la traine de mouvement. Le descripteur se compose :

- instant de début t_0 et durée n
- position horizontale minimale x_{min} et position verticale minimale y_{min}, déplacements maximaux Δx et Δy
- vitesses moyennes le long des directions horizontales et verticales $v_x = \frac{x(t_{n-1})-x(t_0)}{n}$ et $v_y = \frac{y(t_{n-1})-y(t_0)}{n}$
- Coefficients polynomiaux $a_0, ..., a_k$ de $x(t)$: $x(t) = a_0 + a_1(t-t_0) + ... + a_k(t-t_0)^k$
- Coefficients polynomiaux $b_0, ..., b_l$ de $y(t)$: $y(t) = b_0 + b_1(t-t_0) + ... + b_l(t-t_0)^l$
- axe principal : x-axe (si $\Delta x \geqslant \Delta)y$ ou y-axe (sinon)
- Coefficients polynomiaux $c_0, ..., c_m$ de $y(x)$ si x et l'axe principal, $x(y)$ sinon. $y(x) = c_0 + c_1(x - x_{min}) + ... + c_m(x - x_{min})^m$

Puisque un polynome d'ordre faible (1 ou 2) peut ne pas représenter de façon efficace la trajectoire de mouvement sur tout le clip, la trajectoire de mouvement est divisée en sous-trajectoires et chacune de ces sous-trajectoires est décrite par le descripteur proposé. La trajectoire complète est constituée de la concaténation des sous-trajectoires.

DISCOV. Dans l'algorithme DISCOV ("Discovering Objects in Video Sequences") [Liu08], Liu et Chen proposent de combiner la détection de points d'intérêt à la découverte d'un seul objet d'intérêt dans la vidéo. Dans un premier temps, les descripteurs SIFT des points d'intérêt déterminés par les régions MSER [Mat02] sont calculés sur chaque trame indépendamment et quantifiés vectoriellement sur la vidéo. Dans un deuxième temps, la cohérence temporelle de la vidéo est utilisée pour construire les modèles d'apparence et de mouvement de l'objet d'intérêt en suivant une approche probabiliste. La probabilité pour un point SIFT détecté d'appartenir à l'objet étant donné une trame est ainsi déterminée. Cette probabilité a été utilisée dans les tâches d'indexation - telles que la détection de changement de plans et la création de résumé vidéo - basées objet.

Reconnaissance d'objet dans les images. Notons que le problème d'extraction d'objets ne se résume pas à l'extraction d'objets en mouvement dans la vidéo. Des techniques de découverte d'objets dans des images fixes sont aussi étudiées. Ainsi, dans les articles de Lui et Izquierdo [Lui03] et Carson et al [**?**], les primitives -ou régions- des images sont modélisées par des mixtures de gaussiennes dont les paramètres sont estimés par l'algorithme d' "Expectation Maximization" (EM). Pour être plus précis, Lui et Izquierdo associent à chaque pixel de l'image un vecteur de caractéristiques formé des trois composantes couleur écrites dans le système $L * a * b$ et de trois descripteurs de textures : l'anisotropie, le contraste de texture normalisé et l'orientation. Les vecteurs de caractéristiques ainsi formés sont groupés en classe. Pour cela, l'espace des vecteurs de caractéristiques de l'image est modélisé à l'aide d'une mixture de gaussiennes. Le nombre de gaussiennes à utiliser est choisi automatiquement par l'algorithme de MDL. Seulement 2 à 5 gaussiennes sont nécessaires pour décrire toute l'image. Sur chacune des régions ainsi définie, un descripteur hiérarchique est calculé à l'aide d'un algorithme de "K-means". La mesure de similarité utilisée est la "Earth Mover distance".
Ce type d'algorithme ne sera pas repris dans l'état de l'art en extraction d'objets (section 3.2).

Signalons le travail original de Chevalier et al [Che07] qui proposent une mesure de similarité pour l'indexation basée objet dans les graphes.

3.1.4 Evaluation de l'indexation

Dans le cadre de la recherche dans des BD, la qualité d'un indice se définit par sa capacité à ne renvoyer que les vidéos de la BD qui répondent à la requête d'indexation envisagée. Cette requête peut être par exemple : recherche de copies, recherche de vidéos de même catégorie visuelle, recherche de vidéos de même catégorie sémantique, etc. La qualité d'un indice dépend donc de la tâche d'indexation. Les méthodes d'évaluation actuelles se font

par confrontation avec une vérité terrain, c'est-à-dire la réponse donnée par un opérateur humain à la même requête. Dans le cadre de l'évaluation, la vérité terrain définie parfaitement la requête. Il n'est donc pas nécessaire ici de présenter avec précision les tâches d'indexation ; cette présentation sera faite dans la suite de ce manuscrit (chapitre 7).
En tenant compte des réponses de l'algorithme par rapport à la vérité terrain, quatre valeurs sont définies :

- *Vrais Positifs* VP : Nombre de réponses qui sont considérées comme étant réponse à la requête par l'algorithme et par la vérité terrain.
- *Faux Positifs* FP : Nombre de réponses qui sont considérées comme étant réponse à la requête par l'algorithme mais pas par la vérité terrain.
- *Vrais Négatifs* VN : Nombre de réponses qui sont considérées comme n'étant pas réponse à la requête par l'algorithme et par la vérité terrain.
- *Faux Négatifs* FN : Nombre de réponses qui sont considérées comme n'étant pas réponse à la requête par l'algorithme mais qui sont réponses pour la vérité terrain.

A partir de ces observations, les mesures de Rappel (R) et Précision (P) sont définies dans l'équation (3.13).

$$R = \frac{VP}{VP + FN} \qquad P = \frac{VP}{VP + FP} \tag{3.13}$$

Ces valeurs sont comprises dans l'intervalle $[0, 1]$; plus la valeur est proche de 1, meilleur est l'indice. Le rappel permet d'évaluer la capacité à retrouver les vidéos similaires, la précision indique le taux de fausses alarmes. Ces deux mesures sont considérées conjointement pour définir la qualité de l'indice. En fonction de la tâche de recherche, l'une ou l'autre mesure peut être favorisée. Néanmoins, assez fréquemment, on cherche à équilibrer les performances du système. La mesure permettant d'évaluer cette performance est la F-mesure normalisée :

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}} \tag{3.14}$$

Un autre critère est la Précision Moyenne ($Pmoy$). Cette mesure permet de favoriser les systèmes qui renvoient les vidéos les plus significatives le plus tôt dans la liste des réponses. C'est la moyenne des précisions calculées après chaque troncature de la liste, troncature obtenue à chaque fois qu'une détection significative a été trouvée.

$$Pmoy = \frac{1}{N_{max}} \sum_{r=1}^{N_{detecte}} (P(r) * \delta(r)) \tag{3.15}$$

où r est le rang, N_{max} le nombre de détection pertinentes, $N_{detecte}$ le nombre de détections, $\delta(r)$ une fonction binaire de pertinence d'un rang donné et $P(r)$ la précsion à un rang de

troncature donné. En fonction de la complexité de la requête la mesure P_{moy} peut varier.

Afin de caractériser les performances en moyenne d'un système par rapport à un grand ensemble de requêtes avec des contenus différents tout en préservant l'ordre des réponses, les courbes des Rappel-Précision Interpolées sont utilisées. L'axe X dans une telle courbe représente les niveaux standards des rappels : $R_S = \{0, 0.1, ..., 1\}$. L'axe Y correspond à la valeur moyenne (sur toutes les vidéos ou images requêtes) de toutes les valeurs maximales des précisions obtenues aux niveaux de rappel standard de l'ensemble ordonné R_S. Cette mesure est notamment utilisée dans la campagne d'évaluation TRECVid [TRE] et sera utilisée dans notre étude.

3.2 Extraction Spatio-temporelle d'Objets

L'extraction d'objets est une étape de base pour de nombreuses applications incluant les conférences vidéo, le biomédical, la surveillance, l'indexation et la recherche par le contenu, la représentation multimédia (MPEG4). Nous présentons dans la suite quelques techniques d'extraction spatio-temporelle d'objets. Les objets à extraire sont des formes homogènes en texture et/ou en couleur vis-à-vis d'un critère donné et qui sont animées d'un mouvement propre différent du mouvement global de la scène.

Segmentation par modèles de Markov. Une revue détaillée de la modélisation markovienne pour le traitement d'images peut être trouvée dans [Dub89]. Une présentation plus complète de la théorie sera donnée dans le paragraphe 5.3.2.2.
L'extraction d'objets est vue comme un problème d'étiquetage à N classes, une classe représentant le "fond" et les $N-1$ classes restantes représentant les $N-1$ objets en mouvement propre présents dans la scène. En considérant un cadre probabiliste, une image (ou la succession d'images) de la vidéo est vue comme la réalisation y de la variable aléatoire Y des observations. A ces observations est associée une carte de segmentation, elle-même vue comme la réalisation x de la variable aléatoire X des étiquettes. Le problème d'extraction d'objets consiste alors à trouver la carte d'étiquettes $\hat{x}$ (3.16) la plus probable étant donnée l'observation y, c'est à dire la séquence vidéo. Cette estimation est appelée estimation au sens du MAP (Maximum a Posteriori).

$$\hat{x} = \underset{x}{\operatorname{argmax}}\ p(X = x|Y = y) \tag{3.16}$$

La probabilité a posteriori $p(X = x|Y = y)$ est difficilement modélisable. Le problème (3.16) se reformule grâce à la théorie de Bayes qui relie les probabilités a priori et a poste-

riori. Le problème se reformule par :

$$\hat{x} = \underset{x}{\operatorname{argmax}}\ p(Y = y|X = x)p(X = x) \tag{3.17}$$

En supposant la markoviannité du champ de étiquettes X, la probabilité $p(X = x)$ peut être formulée en fonction d'une somme de potentiels. Toute la difficulté réside dans la définition de ces fonctions de potentiel qui sont définies de façon plus qualitative que quantitave. Un des premiers travaux appliquant ces idées est décrit dans l'article de Black [Bla92]. Trois potentiels sont définis,E_M, E_I et E_L qui expriment les hypothèses faites respectivement sur le champ de mouvement, la structure des valeurs d'intensité et l'organisation des discontinuités. Les potentiels tiennent compte entre autres des mesures de l'image précédente. Tsaig et Averbruch [Tsa02] proposent d'utiliser les champs de Markov pour réguler le résultat d'une première segmentation. Le problème est formulé comme un problème d'étiquetage de graphe basé sur l'information de mouvement. Le champ des étiquettes est modélisé par un Champs de Markov Aléatoire (MRF). Une partition initiale de chaque trame est obtenue par l'algorithme de Ligne de Partage des Eaux (LPE) (cf section 5.2.2). Le mouvement de chaque région est estimé par un schéma de validation du mouvement hiérarchique. Le suivi temporel des objets, par mémoire dynamique, est incorporé au processus de segmentation. L'étiquetage est finalement obtenu en combinant les informations de mouvement, les informations spatiales et les informations de suivi dans un MRF. Zeng et al [Zen05] proposent de faire de la détection d'objets en mouvement comme un processus de Markov d'étiquetage des macro-blocs du flux H264. La définition des potentiels est faite pour intégrer des notions de suivi d'objets. Enfin, Brouard et al [Bro08] proposent une combinaison de quatre potentiels définis pour des sites qui sont des tubes spatio-temporels.

Segmentation par contours actifs. La segmentation par contours actifs permet le suivi des objets en mouvement. Le contour actif est une courbe que l'on fait évoluer pour réaliser la segmentation. Soit Γ la courbe. Elle suit la loi d'évolution temporelle :

$$\frac{\partial \Gamma}{\partial t} = F\vec{N} \tag{3.18}$$

où $\vec{N}$ est la normale à la courbure orientée vers l'intérieur de la courbe et F traduit les "forces" qui s'appliquent sur le contour et lui permettent d'évoluer. La force F est dérivée d'une énergie. Cette énergie doit être minimum pour une courbe lisse aux bords de l'objet à détecter. Cette énergie peut être décomposée en une énergie interne (traduisant l'élasticité et/ou la rigidité de la courbe) et une énergie externe (traduisant l'attache aux données). Une équation d'évolution est déduite de 3.18 permettant de déplacer le contour actif. Le principal problème des contours actifs est la précision limitée de la détection des contours due aux incertitudes sur l'estimation du mouvement et du fond.

Extraction d'objets par modélisation du fond. Dans les applications de vidéo surveillance, une technique d'extraction d'objets très utilisée est la soustraction du fond comme par exemple dans [Col98]. La distribution des intensités du fond est estimée de façon récursive à partir des vraies données de l'image. Soit $I_t(x, y)$ l'intensité du pixel à la position (x, y) dans la trame I_t. La valeur d'intensité du fond $B_{t+dt}(x, y)$ à la même position à l'instant $t + dt$ est donnée par :

$$B_{t+dt}(x,y) = \begin{cases} aB_t(x,y) + (1-a)I_t(x,y) \text{ si le pixel } (x,y) \text{ ne bouge pas} \\ B_t(x,y) \text{ si le pixel } (x,y) \text{ bouge} \end{cases} \tag{3.19}$$

où $B_t(x, y)$ est l'estimée précédente de la valeur d'intensité du fond à la même position de pixel. Le paramètre de mise-à-jour a est un nombre positif proche de 1. $B_0(x, y)$ est initialisé comme étant égal à la première trame $I_0(x, y)$. Les instants t et $t + dt$ sont pris immédiatement successifs. Un pixel positionné en (x, y) est supposé en mouvement si les valeurs d'intensité à cette position entre les images I_t et I_{t-dt} satisfont l'inégalité suivante :

$$|I_t(x,y) - I_{t-dt}(x,y)| > T_t(x,y) \tag{3.20}$$

$T_t(x, y)$ est un seuil décrivant un changement significatif d'intensité à la position (x, y). Ce seuil est mis-à-jour récursivement pour chaque pixel comme suit :

$$T_{t+1} = \begin{cases} aT_t(x,y) + (1-a)(c|I_t(x,y) - B_t(x,y)| \text{ si le pixel } (x,y) \text{ ne bouge pas} \\ T_t(x,y) \text{ si le pixel } (x,y) \text{ bouge} \end{cases} \tag{3.21}$$

où c est un nombre réel supérieur à 1. Plus ce paramètre est grand, plus le seuil est élevé et moins le schéma de détection est sensible. On suppose que le régions qui diffèrent du fond sont les régions en mouvement. C'est-à-dire qu'elles sont constituées des pixels vérifiant :

$$|I_t(x,y) - B_t(x,y)| > T_t(x,y) \tag{3.22}$$

Töreyin et al [Ugu05] proposent d'utiliser cette approche dans le domaine compressé des ondelettes 9/7 de Daubechies. Le fond est calculé séparément sur chaque sous-bande comme indiqué précédemment. Les pixels en mouvement sont définis individuellement dans chaque sous-bande. Le résultat est projeté par bloc sur la pleine résolution. Les pixels en mouvement dans l'image courante sont définis comme étant l'union de tous les pixels en mouvement dans chaque sous-bande.

Coopération segmentation en mouvement et segmentation spatiale. Le problème d'extraction spatio-temporelle d'objets en mouvement peut être vu simplement comme la fusion d'une segmentation temporelle et d'une segmentation spatiale. Ce type de méthode a d'abord été utilisé dans les schémas de codage par objet. Dans un tel scenario, le but est de

définir des régions de mouvement homogène qui constituent les objets. Dans [Wu96, Sal97], le principe est de fusionner les régions d'une segmentation couleur par un critère d'homogénéité sur le mouvement. Le mouvement de chaque région est supposé suivre soit un modèle de translation [Wu96] soit un modèle affine à 6 paramètres [Wu96, Sal97] et est estimé par descente de gradient. Le critère de fusion est basé sur l'Erreur de Prédiction par le modèle de mouvement, cette erreur est mesurée par l'Erreur Quadratique Moyenne (MSE). Afin de conserver la cohérence le long de l'axe du temps, un suivi temporel des objets est assuré en projetant, grâce au mouvement estimé, le résultat de la segmentation à l'instant $t-1$ sur l'image à l'instant t. Un ajustement de la carte de segmentation est alors effectué. Wu et al [Wu96] proposent d'utiliser une segmentation spatiale fondée sur la détection de contours et la fermeture de ces contours. Cette segmentation est ensuite utilisée pour fournir un modèle polygonal des contours des régions. Après projection, l'ajustement se fait alors par la technique des "polygones ajustables" [Del94], un polygone ajustable peut être vu comme un ensemble de contour actifs élémentaires. Une étape de gestion des occlusions est ensuite réalisée. Dans [Sal97], la segmentation spatiale est effectuée à l'aide de l'algorithme de la LPE (cf. section 5.2.2). Après projection temporelle, les régions obtenues servent de marqueurs à l'algorithme de LPE. Pour chaque trame, un "arbre de partitions" est crée. Il est formé à partir d'une représentation hiérarchique de la segmentation précédemment obtenue. Les niveaux inférieurs sont obtenus par re-segmentation des régions, les niveaux supérieurs par fusion des régions suivant le mouvement. Cet arbre sert à déterminer la meilleure partition à utiliser pour l'encodage.
Manerba et al [Man04] proposent de fusionner les segmentations couleur et mouvement par vote majoritaire. La segmentation en mouvement est effectuée par estimation du mouvement global avec rejet des valeurs non conformes à partir des vecteurs du flux MPEG2. Une segmentation couleur fine, par LPE modifiée, permet d'obtenir les contours des objets avec précisions. La fusion se fait par vote majoritaire : les régions de la segmentation couleur couvertes par un pourcentage minimum de la segmentation en mouvement sont conservées. Dans [Man08], les mêmes auteurs proposent d'assurer le suivi des objets à l'aide de tubes spatio-temporels.
Dans ce type d'approche, une étape de suivi doit être effectuée, alors que dans les méthodes précédentes, elle peut être intégrée à l'algorithme d'extraction d'objets. Pour résoudre ce problème, Saban et Manjunath [ElS03] proposent d'utiliser un algorithme de LPE dans le domaine 2D+t compensé en mouvement.

Segmentation hiérarchique. Ces méthodes cherchent à déduire des informations sémantiques de niveau moyen à partir des segmentations de bas niveau. A partir d'une segmentation détaillée, il s'agit de produire une hiérarchie des segmentations emboîtées. Chaque niveau dans cette hiérarchie étant caractérisé par une certaine valeur (absolue ou relative) d'un critère d'homogénéité. La fusion des régions peut être effectuée soit deux par deux selon

le parcours d'un arbre binaire [Sal00] ou encore lors de fusion de plusieurs régions voisines suivant le parcours du graphe de voisinage [BP98]. Remarquons qu'aux différents niveaux de la hiérarchie emboîtée les critères d'homogénéité peuvent être de nature diverse. Dans le cas de la segmentation vidéo, les premiers niveaux peuvent être construits à base de critères d'homogénéité spatiale, ensuite, pour les autres niveaux, l'homogénéité du mouvement peut être mise en jeu. En tant que niveau initial de segmentation, il est possible de considérer par exemple soit une segmentation morphologique fine soit que chaque pixel représente une région-germe (comme dans le cas des pyramides irrégulières [Ber95]). Nous allons voir par la suite que notre méthode d'extraction des objets des séquences vidéo peut être vue comme une hiérarchie à deux niveaux, le premier niveau est constitué par la segmentation morphologique couleur, le deuxième par l'objet d'intérêt constitué des régions dont les pixels ne suivent pas le mouvement de la caméra.

3.2.1 Evaluation du résultat de la segmentation

L'extraction d'objets est une segmentation d'images en deux classes : l'objet et le fond. L'évaluation du résultat d'une méthode de segmentation est une tâche très difficile. Elle dépend fortement de l'application dans laquelle la segmentation doit être utilisée. Un passage en revue des différentes techniques d'évaluation de la segmentation a été fait par Zhang [Zha96] et Philipp-Foliguet et Guignes [PF06]. L'évaluation peut être faite soit de façon quantitative, soit de façon qualitative.

Les critères d'évaluation quantitative peuvent être séparés en deux classes, selon que l'on possède ou non une "vérité-terrain" qui constitue une *segmentation de référence*. Celle-ci est directement accessible dans le cas d'images de synthèse, mais elle doit être construite "à la main" par un expert du domaine de l'application dans le cas des images réelles. Si l'on veut comparer de manière objective les méthodes, il est plus simple d'utiliser des images de synthèse, pour lesquelles la "vérité" est parfaitement connue, à savoir la segmentation qui a servi à synthétiser l'image. L'inconvénient d'une telle démarche est que ces images ne représentent pas toutes les situations possibles d'une prise de vue réelle. Bien que l'évaluation sur des images réelles soit certainement plus réaliste, elle pose d'autres difficultés, la principale étant qu'il n'existe généralement pas de solution unique à la division d'une image en régions "pertinentes". La "pertinence" d'une région est en effet une notion éminemment dépendante de l'application. Néanmoins, deux segmentations humaines d'une même image tendent à être cohérentes dans le sens où elles sont des raffinements mutuels l'une de l'autre (par exemple, une personne peut être vue comme l'ensemble formé de la tête, des bras, du torse et des jambes...).
Sans segmentation de référence, la qualité de la segmentation se mesure par des mesures

établies suivant l'intuition humaine de ce qu'une "bonne" segmentation doit être, pour obtenir une image agréable à l'oeil par exemple. Plusieurs méthodes existent qui cherchent à favoriser l'uniformité intra-région, le contraste inter-région ou le degré de lissage des contours des régions. Une revue de ces méthodes est faite par [Zha96]. La plupart des chercheurs préfèrent se reposer sur un jugement qualitatif humain pour l'évaluation. En effet, la qualité d'une extraction d'objets dépend fortement de l'application visée et est donc subjective.
Les méthodes d'évaluation de la segmentation sont encore aujourd'hui un sujet de recherche actif, citons par exemple les travaux de Gelasca et Ebrahimi [DG09].

3.3 L'approche proposée

Les sections précédentes nous ont permis de passer en revue les diverses méthodes d'indexation et d'extraction d'objets qui existent dans la littérature. Ces problèmes ont été et sont toujours largement étudiés. Nous avons ainsi été amené à faire des choix dans les méthodes présentées et à rester concis. L'objet de cette section est de présenter les choix que nous avons faits pour résoudre notre problématique d'indexation basée objet des vidéos.
Tout d'abord, nous avons choisi de développer une indexation basée objet. Nous avons vu que d'autres types d'indexation, plus largement étudiés, existent. Ce sont les indexations globale et locale. L'indexation globale souffre d'un problème d'invariance liée à sa définition sur des zones de chaque trame de la vidéo pouvant être largement invariantes. L'indexation locale tente de résoudre ce problème en ne calculant un descripteur que sur les points les plus invariants de la vidéo. Le problème de ces points est qu'ils ne sont pas assurés d'être stables le long des trames d'une vidéo d'une part et que peu de points sont effectivement appariés (environ 30% d'après [Mik04]). L'avantage d'utiliser des objets est qu'ils permettent d'apporter un niveau d'interprétation sémantique de moyen niveau par rapport à des points d'intérêt qui sont de bas niveau. C'est ce dernier argument, que nous estimons très fort, qui nous a amené à choisir une indexation basée objet. Une fois ce choix fait, il nous reste deux sous-problèmes à envisager. Quelle méthode utiliser pour extraire les objets et quel type d'indice construire sur les objets extraits ?
La méthode d'extraction d'objet choisie est tout naturellement hiérarchique. En effet, sous le "paradigme de l'indexation primaire" [Man04], nous avons été amené à considérer le domaine des ondelettes, c'est-à-dire le domaine compressé JPEG2000. Rappelons que ce type de représentation est multirésolution. Nous avons décidé d'adopter une approche utilisant le moins d'information possible. C'est-à-dire que, par analogie avec la recherche par image-clé, nous proposons de travailler sur des couples d'images ; le couple d'images étant la plus petite unité permettant d'obtenir une information de mouvement. D'autre part, nous n'avons fait aucune hypothèse quant au contenu des vidéos étudiées. Les applications

visées par notre approche sont la recherche dans des bases de données, en particulier celles du cinéma numérique. Vu les contenus extrêmement variés suivant le genre des films, il n'est pas aisé de trouver d'*a priori* fort suivi par tous les types de vidéo. Notre approche se doit d'être la plus générale possible. Au vu de la littérature, la combinaison d'une segmentation en mouvement et d'une segmentation couleur, sans donner le meilleur résultat pour chaque vidéo individuellement, donne un résultat tout à fait correct sur une grande diversité de données. Pour les mêmes raisons, c'est une approche de segmentation morphologique qui est envisagée pour la couleur. Les méthodes d'extraction multirésolution des objets par projection et ajustement successifs ont largement été utilisées dans la littérature et avec succès. Notre approche se propose d'appliquer un tel schéma dans le domaine des ondelettes. Contrairement aux travaux existants dans le domaine, nous proposons d'utiliser directement les valeurs des coefficients HF et n'intégrons pas notre ajustement dans le processus de transformation inverse [Jun03, Kim03].

Intéressons-nous maintenant à la solution de segmentation en mouvement. Dans les approches précédentes travaillant dans le domaine compressé, ce type de segmentation est réalisé à partir des vecteurs de mouvement des macro-blocs encodés dans le flux compressé. Dans un standard tel que MJPEG2000, aucune information de mouvement n'est disponible. Il faut donc envisager une stratégie d'estimation de mouvement. D'une part, cette estimation sera hiérarchique du fait de la nature des ondelettes. D'autre part, l'approche hiérarchique se fera de la Basse Résolution (BR) vers la Haute Résolution (HR) pour rester cohérent avec la structure scalable du flux et les modes de transmission associés. Enfin, comme nous travaillons dans une optique flux compressé, nous avons décidé d'utiliser les mêmes techniques d'estimation de mouvement que dans les standards de compression actuels. Ainsi, nous proposons une approche par Mise en Correspondance de Blocs (MCB). Il est bien connu cependant que le processus de décimation dans les pyramides multirésolution rend ces représentations non invariantes par translation. Le processus de Mise en Correspondances de Blocs en est plus imprécis. Afin de corriger les erreurs ainsi introduites, nous proposons de régulariser les vecteurs obtenus à l'aide d'une estimation robuste du mouvement global. Ce processus d'estimation robuste permet en plus d'aboutir simplement à une segmentation en mouvement des trames.

Une fois les objets extraits, il nous reste à choisir une stratégie de définition du descripteur. Nous avons déjà évoqué le fait que le défaut des descripteurs locaux était que seul un faible pourcentage d'entre eux est effectivement mis en correspondance. Il faut donc qu'un nombre certains de descripteurs aient pu être extraits. Dans le cas d'une indexation par objet, l'aire effectivement utilisée pour la définition des descripteurs est relativement petite. Il en résulte que très peu de descripteurs locaux pourront en être extraits. L'indexation par descripteurs locaux nous semble alors inefficace. Nous nous proposons alors d'utiliser des descripteurs globaux. Parmi ces descripteurs, nous avons choisi d'utiliser l'histogramme. Ce type de descripteur a largement été étudié, il est simple et robuste. Son désavantage

Paramètre	valeur
Taille de l'image	1920x1080psf
Système Couleur	YCrCb
Nombre de tuiles	1
Type de compression	avec pertes
Base d'ondelettes	9/7 de Daubechies
Nombre de niveaux de décomposition en ondelettes	5
Nombre de couches de qualité	2

Tab. 3.1 – Caractéristiques des vidéos et Paramètres de l'encodeur JPEG2000 utilisés dans cette thèse

est que, comme il ne tient pas compte de la répartition spatiale des coefficients, il peut ne pas distinguer deux contenus spatiaux très différents. Cependant, comme nous travaillons uniquement sur les objets en mouvement, l'extraction et la prise en compte de l'information spatiale sont déjà faites. Nous pensons qu'alors le descripteur histogramme est suffisamment discriminant. Nous améliorerons de plus la qualité du descripteur en définissant les histogrammes des coefficients de HF des ondelettes. Par là-même, nous ajoutons à notre descripteur une information de contour et de texture.
Avant de passer à la description des méthodes, les caractéristiques des vidéos compressées par JPEG2000 sont présentées dans le tableau 3.1. Ces caractéristiques sont celles préconisées par la norme pour le cinéma numérique [ISO06].

3.4 Conclusion

Le présent chapitre nous a permis d'indiquer quelques travaux dans les domaines de l'indexation des vidéos et de l'extraction d'objets. Ces domaines sont étudiés intensivement depuis plus d'une décennie. Cependant, il reste encore de nombreux problèmes à résoudre avant de considérer ces recherches comme achevées. L'indexation et l'extraction d'objets souffrent en effet toutes deux du fossé sémantique. Ce n'empêche que des systèmes opérationnels de recherche par le contenu de vidéos ont été développés [Pin, Han01].
Nous proposons de nouvelles méthodes pour l'indexation par le contenu des vidéos dans le domaine compressé JPEG2000. Ces méthodes sont présentées dans les chapitres suivants.

Deuxième partie

Segmentation spatio-temporelle dans le domaine des ondelettes

Chapitre 4

Estimation de mouvement dans le domaine des ondelettes : une base pour l'extraction spatio-temporelle d'objets

Le préalable à une extraction spatio-temporelle d'objets dans le domaine compressé est d'être capable d'estimer le mouvement dans la vidéo. Pour les standards H.26x et MPEG, une information de mouvement est immédiatement disponible sous la forme de vecteurs de déplacement directement codés dans le flux. Dans le cas du standard JPEG2000, aucune donnée de ce type n'existe (section 2.1). Il est donc nécessaire de l'estimer dans le domaine de la TOD, c'est-à-dire le domaine compressé. L'algorithme que nous proposons est d'abord décrit comme une méthode indépendante ; nous indiquerons dans les chapitres suivants comment l'intégrer aux méthodes d'extraction d'objets en mouvement.
Dans la suite du chapitre, les techniques classiques d'estimation de mouvement dans le domaine pixel sont indiquées et les difficultés de leur adaptation au domaine d'ondelettes sont décrites. Nous présentons ensuite notre méthode d'estimation de mouvement et en illustrons l'intérêt sur quelques séquences d'exemple.

4.1 Etat de l'art : estimation du mouvement et ondelettes

4.1.1 Estimation du mouvement dans le domaine pixel

Le mouvement apparent, ou flot optique, est le mouvement observé dans une séquence d'images et résultant des mouvements des objets dans une scène 3D et du mouvement de la caméra. La problématique consistant à estimer ce mouvement à partir des informations d'intensité et de couleur des trames successives de la vidéo a largement été étudiée au cours des dernières décennies. Nous nous limitons dans ce paragraphe au rappel des principales techniques existantes et à la présentation succincte de quelques techniques représentatives. Une présentation plus complète peut être trouvée dans l'article de Stiller et al [Sti99] et dans les ouvrages de Tekalp [Tek95] (chapitre 3), et Barlaud, Labit et al. [Bar02] (Chapitre 5).
L'estimation de mouvement repose sur l' **hypothèse de conservation de l'intensité** : l'intensité d'un pixel est constante le long de sa trajectoire dans le temps. Ainsi, si $I(x, y, t)$ est l'intensité du pixel à la position (x, y) à l'instant t et (dx, dy) le vecteur de déplacement de ce pixel pendant le temps $dt > 0$, la Displaced Frame Difference (DFD) est définie par (4.1) :

$$DFD(x, y, t) = I(x + dx, y + dy, t + dt) - I(x, y, t) \tag{4.1}$$

L'hypothèse de conservation de l'intensité se traduit alors par (4.2) :

$$DFD(x, y, t) = 0 \tag{4.2}$$

Par développement en série de Taylor au premier ordre au point (x, y, t) de $I(x + dx, y + dy, t + dt)$, l'équation (4.2) se réécrit

$$\nabla\vec{I}.\vec{w} = -grad_t I \tag{4.3}$$

où $\nabla\vec{I} = (grad_x I, grad_y I)^T$ est le vecteur des gradients d'intensité spatiaux, $grad_t I$, le gradient d'intensité temporel et $\vec{w} = (u, v)$ le vecteur de vitesse de déplacement du point considéré. Cette équation est appelée **Equation de Contrainte du Mouvement Apparant (ECMA)** (4.3). L'estimation de mouvement est un problème mal posé. Ainsi, seule la composante parallèle au gradient spatial d'intensité peut-être déterminée. En effet, posons $\vec{w} = \vec{w}_{\parallel} + \vec{w}_{\perp}$ où $\vec{w}_{\parallel}$ (resp.$\vec{w}_{\perp}$) désigne la composante parallèle (resp. perpendiculaire) au contour local soit perpendiculaire (resp. parallèle) au gradient d'intensité. Alors (4.3) se réécrit $\nabla\vec{I}.\vec{w}_{\parallel} + \nabla\vec{I}.\vec{w}_{\perp} = -grad_t I$. Par définition, $\nabla\vec{I}.\vec{w}_{\parallel} = 0$ et donc seule $\vec{w}_{\perp}$ peut être estimée par cette méthode. Ce phénomène bien connu est appelé le phénomène d'ouverture ("aperture") et est lié à l'observation de la scène au travers d'une fenêtre spatiale réduite.

En réalité, l'égalité (4.2) n'est pas respectée car des variations d'intensité existent, liées au propriétés physiques intrinsèques de l'éclairage de la scène et du capteur de la caméra. Le problème d'estimation du mouvement est alors défini comme un problème de minimisation de la DFD (4.4) : le déplacement d'un pixel est celui qui minimise la différence absolue d'intensité.

$$(dx, dy) = \underset{(dx,dy)}{argmin} |DFD(x, y, t)| \tag{4.4}$$

Dans la suite, nous allons présenter trois grandes catégories de résolution du problème : les méthodes de minimisation locale, de minimisation globale et de modélisation du champ de vecteurs. Afin de rester concis, seul un exemple représentatif de méthode sera développé par catégorie.

La méthode de résolution de *minimisation locale* la plus répandue dans le domaine des codeurs vidéos tels que H.26x [ITRH93, ITRH05] et MPEGx [ISO93, ISO00, ISO04b] est la Mise en Correspondance de Blocs (MCB). L'idée est de découper l'image en blocs et de minimiser une fonctionnelle de la DFD individuellement pour chacun de ces blocs afin d'en déterminer le déplacement. Le principe détaillé de cette technique est présenté dans la section 4.2.2.

La *minimisation globale* se fait en minimisant une énergie, fonctionnelle de l'équation (4.3). Il est alors aussi possible de rajouter des contraintes de lissage du champ des vecteurs. C'est le cas de la méthode de Horn et Schunk [Hor81] qui rajoute une contrainte de lissage du champ en imposant que les gradients locaux des vecteurs de déplacement soient faibles. Ainsi, le problème de minimisation s'écrit comme suit (4.5).

$$\mathcal{W} = \underset{\mathcal{W}}{argmin} \sum_{(x,y),w\in\mathcal{W}} (E_{ECMA} + \alpha^2 E_w^2) \tag{4.5}$$

$\mathcal{W} = \{w = (u, v)\}$ est le champs de déplacement, $E_{ECMA} = grad_x I.u + grad_y I.v + grad_t I$ est l'énergie associée à l'ECMA et $E_w^2 = ||\vec{\nabla}u||^2 + ||\vec{\nabla}v||^2$ est la contrainte de lissage pondérée par le paramètre α.

Enfin, il est possible de rajouter une contrainte en définissant un *modèle pour le champs de vecteurs*. Le modèle le plus utilisé est affine à 6 paramètres décrit par (4.6) qui lie l'amplitude du déplacement à la position.

$$\begin{cases} dx = a_1 + a_2 x + a_3 y \\ dy = a_4 + a_5 x + a_6 y \end{cases} \tag{4.6}$$

Le problème ne consiste plus à trouver pour chaque point ou région le vecteur de mouvement mais d'estimer les paramètres du mouvement $\theta = (a_1, a_2, a_3, a_4, a_5, a_6)^T$qui minimisent globalement la DFD. La formalisation d'un tel problème conduit à un système d'équations sur-déterminé et les techniques d'optimisation adéquates doivent être utilisées. Une solution très satisfaisante est d'utiliser un estimateur robuste afin d'identifier et de rejeter des

mesures aberrantes dues par exemple aux objets en mouvement. C'est ce qu'ont proposé Odobez et Bouthemy [Odo95] et qui a donné lieu au logiciel de référence français en terme d'estimation de mouvement : Motion2D [Mot].
Pour clore cette rapide présentation de l'estimation de mouvement dans le domaine pixel, il faut évoquer les techniques multirésolution [Odo95]. Un tel schéma peut s'appliquer à toutes les approches évoquées précédemment. Les images sont d'abord décomposées en pyramides gaussiennes. L'estimation du mouvement se fait par la technique choisie entre les images de Basse Résolution des pyramides. Le résultat trouvé est projeté au niveau de résolution immédiatement supérieur en s'adaptant au facteur d'échelle. La technique d'estimation de mouvement est alors appliquée à nouveau sur cette nouvelle résolution, les valeurs trouvées par projection servant d'initialisation. La procédure est répétée jusqu'à la pleine résolution. Avec une telle approche, il est possible de capter les mouvements de forte amplitude à la basse résolution. A contrario, les mouvements de plus faible amplitude constituent des éléments de détail et sont récupérés au fur et à mesure que l'on augmente la résolution dans le processus ce qui rend l'estimation plus précise que dans le cas d'une estimation mono-niveau, en évitant ainsi de converger vers un minimum local.

4.1.2 Estimation du mouvement dans le domaine des ondelettes

Le problème d'estimation du mouvement dans le domaine des ondelettes discrètes est non trivial et a été largement abordé dans la littérature à l'occasion de la construction de codeurs en ondelettes compensées en mouvement dites 2D+t [Ohm04]. En effet, le processus de décimation de la TOD la rend non-invariante par translation. De ce fait, obtenir les vecteurs de mouvement par MCB dans le domaine des ondelettes n'est pas efficace. Le signal de basse fréquence est habituellement lisse et la différence des coefficients entre le signal original et le signal translaté est faible. Cependant, il existe une très grande différence entre les coefficients des bandes de HF du signal original et de son translaté. Un tel phénomène apparaît souvent au bord des objets. La différence dans les signaux dans les sous-bandes de haute fréquence dépend de la valeur de la translation ainsi que du filtre utilisé pour la TOD. Les erreurs de prédiction rendent ainsi difficile l'estimation des vecteurs de mouvement dans le domaine des ondelettes par un algorithme de MCB classique. Plusieurs méthodes d'estimation de mouvement dans le domaine des ondelettes ont été développées [Kim01, Liu07, Mae] . Parmi celles-ci, l'estimation de mouvement directe bande à bande n'est pas efficace du fait de la propriété de non-invariance par translation de la TOD [Mae]. Une autre approche est d'effectuer l'estimation de mouvement seulement sur la sous-bande de BF et de compenser en mouvement les sous-bandes de HF avec ces vecteurs. Pour limiter les effets de la non-invariance par translation, une autre technique consiste à décomposer l'image de référence non pas avec la TOD mais avec la Transformée en Ondelettes Complète, qui est une TOD sans décimation. La décimation

n'intervenant plus dans l'image de référence, la mise en correspondance de blocs devient plus efficace [Liu07], mais la taille de l'image de référence devient très grande. Dans la méthode de translation de la sous-bande de BF ([Kim01]) la MCB est faite entre l'image courante et un ensemble d'images de référence constitué des transformées en ondelettes de versions translatées de l'image de référence.
Les méthodes précédemment décrites n'utilisent pas la TOD classique mais proposent des améliorations (du point de vue de l'estimation de mouvement) de cette transformation. Comme le but de notre approche est de travailler à partir des données du flux compressé, la TOD nous est imposée par l'encodeur MJPEG2000. C'est pourquoi nous ne détaillerons pas d'avantage l'état de l'art de l'estimation de mouvement dans le domaine des ondelettes. Comme estimer le mouvement sur les sous-bandes de HF est un problème compliqué et que l'information supplémentaire apportée par une telle estimation n'est pas importante, nous avons décidé de n'utiliser que la sous-bande de BF dans notre stratégie d'estimation du mouvement.

4.1.3 Approche retenue

Dans notre approche, les seules données disponibles sont celles du flux compressé. Du fait de la scalabilité, seule une partie du flux et non le flux entier peut être disponible. Cela signifie que dans un premier temps, seuls les coefficients d'ondelettes de basse résolution sont disponibles, c'est-à-dire ceux de la couche de base. Les couches de détails peuvent ensuite être mises à disposition et permettre d'améliorer la qualité des informations disponibles. Cette structuration des données est typiquement celle des stratégies hiérarchiques d'estimation de mouvement. La différence est que ce n'est pas une pyramide gaussienne mais une pyramide d'ondelettes qui est utilisée. Il s'agit donc de choisir une méthode d'estimation de mouvement à appliquer sur chaque niveau de la pyramide. Conformément à nos conclusions sur l'état de l'art dans le domaine des ondelettes, seule la composante de BF est utilisée et les stratégies du domaine pixel peuvent être appliquées.
Rappelons que notre travail suit le Paradigme de l'Indexation Primaire. Une telle hypothèse nous incite à adopter une solution orientée codeur afin de pouvoir appliquer notre démarche, moyennant des adaptations mineures, à des vidéos compressées à l'aide d'encodeurs 2D+t qui sont actuellement le sujet de nombreuses recherches. Une information de mouvement est potentiellement contenue dans ce flux et se présente alors sous la forme d'un ensemble de vecteurs de mouvement. Ces vecteurs sont choisis de manière à obtenir un PSNR (Peak Signal to Noise Ratio) de reconstruction minimum, autrement dit de manière à minimiser la redondance temporelle. La méthode choisie est donc la méthode de MCB utilisée dans la plupart des standards de compression vidéo.
Cependant, la technique de MCB étant pensée pour le codage et non pour l'analyse de mouvement, elle se prête peu à l'extraction des objets en mouvement qui est notre but.

C'est pourquoi nous proposons de la combiner avec une estimation robuste du mouvement global. Le mouvement de l'arrière-plan de la scène est ainsi régulé par approximation des vecteurs de mouvement obtenus par MCB par le modèle trouvé. De plus, cela apporte une indication sur les lieux où le mouvement a besoin d'être estimé avec plus de précision, typiquement sur les objets d'avant-plan.
Notre approche est détaillée dans la suite de ce chapitre.

4.2 Méthode hiérarchique proposée

4.2.1 Vue d'ensemble de la méthode d'estimation de mouvement

Le schéma global de l'estimation de mouvement dans le domaine des ondelettes est présenté figure 4.1. Du fait de la compression par JPEG2000, chaque image de la séquence est décomposée en ondelettes sur K niveaux (section 2.4). D'abord, une MCB est effectuée sur les sous-bandes LL du Kième niveau de la décomposition (section 4.2.2). Ensuite, le mouvement global, assimilé à celui de la caméra, est estimé en utilisant un estimateur robuste (Estimation du Mouvement Global (EMG), section 4.2.3) [Dur01]. Ce dernier permet non seulement de déterminer un modèle de mouvement mais aussi d'attribuer un poids à chaque vecteur suivant son adéquation à ce modèle. Les deux informations obtenues sont fusionnées ; les vecteurs de poids faible indiquant les blocs qui ne suivent pas le mouvement global. Le signal est ensuite synthétisé niveau par niveau (c'est-à-dire que la sous-bande LL du niveau considéré est calculée). Les vecteurs de mouvement au niveau courant sont initialisés (bloc initialisation par projection, indiqué par des pointillés variables sur la figure 4.1) en utilisant les vecteurs de mouvement prédits au niveau précédent. Pour les blocs de poids faible (section 4.2.4), c'est le vecteur issu de la MCB initiale qui sert de prédiction ; pour les blocs conformes au modèle, c'est le modèle global qui détermine la prédiction. Cette différentiation est appliquée à la propagation des tailles des blocs à travers la pyramide. Les blocs ayant un vecteur de poids faible préservent leur taille, les autres, conformes au modèle global, voient leur taille s'agrandir proportionnellement au facteur d'échelle (section 4.2.5). Les vecteurs de mouvement sont ensuite affinés par une MCB au niveau courant de la pyramide d'ondelettes et le mouvement global est ré-estimé. La procédure est réitérée jusqu'à la pleine résolution ($k = 0$).

Il faut souligner que la différenciation entre mouvement global (c'est-à-dire le mouvement du fond) et mouvement local, sur laquelle s'appuie notre démarche, sert à l'extraction des objets en mouvement propre. Malgré notre soucis de respecter le Paradigme de l'Indexation Primaire, il n'est pas question ici d'avoir une démarche visant uniquement à améliorer la qualité de reconstruction des images de la vidéo par compensation de mouvement mais seulement de respecter le cadre de travail imposé par les techniques de codage.
Les diverses étapes de l'estimation de mouvement sont détaillées dans les paragraphes

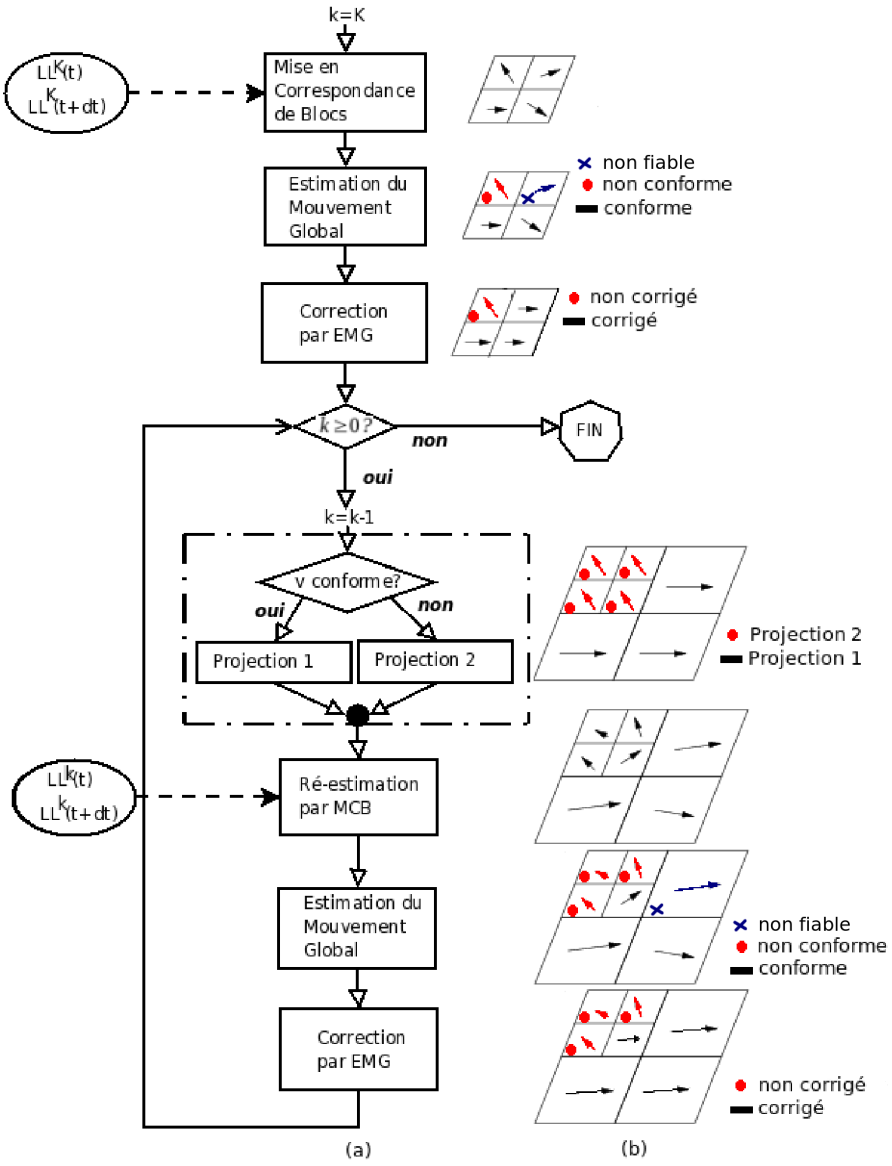


Fig. 4.1 – Schéma général de l'estimation de mouvement dans le domaine des ondelettes

suivants.

4.2.2 Estimation du mouvement à Basse Résolution : Mise en Correspondance de Blocs

La MCB est utilisée à deux endroits de notre démarche. A BR (bloc fonctionnel Mise en Correspondance de Blocs de la figure 4.1), elle permet d'initialiser le processus d'estimation de mouvement. Aux niveaux intermédiaires (bloc fonctionnel ré-estimation par MCB de la

figure 4.1), elle sert à ajuster les vecteurs obtenus par projection des résultats du niveau précédent. Dans la suite de ce paragraphe, la stratégie de MCB choisie parmi toutes les variantes existantes est d'abord présentée dans le cas de la BR. Son application à la ré-estimation des vecteurs sera présentée en fin de paragraphe.
Soient LL^K_{t-dt} et LL^K_t les sous-bandes LL de niveau K de la décomposition en ondelettes de deux images I_{t-dt} et I_t d'une même séquence vidéo prises à deux instants de temps $t-dt$ et t différents, avec $dt > 0$ (figure 4.2). Dans la suite, les sous-bandes LL sont assimilées à des images (section 2.4). L'algorithme de MCB permet d'estimer le mouvement de l'image LL^K_t dite image courante par rapport à l'image LL^K_{t-dt} appelée image de référence. L'image courante est divisée en N blocs de taille nxn constituant une partition de l'image. Dans le cas où la hauteur et/ou la largeur de l'image ne sont pas des multiples de n, les blocs des bords "droit" et "bas" de l'image auront une taille n^i_1xn^i_2 inférieure, c'est-à-dire $n^i_1 \leqslant n$ et $n^i_2 \leqslant n$ avec $n^i_1 < n$ si $n^i_2 = n$ et $n^i_2 < n$ si $n^i_1 = n$. Ces blocs sont notés B_i, $i \in \{0, ..., N-1\}$. Si dt est suffisamment petit, le mouvement de chaque bloc B_i entre $t-dt$ et t peut être approximé par une translation. Il s'agit alors de déterminer la position de B_i à l'instant $t-dt$; la position d'un bloc étant la position du coin supérieur gauche de ce bloc. En supposant qu'il y a conservation de l'intensité lumineuse au cours du temps (autrement dit les différences entre les images LL^K_{t-dt} et LL^K_t sont uniquement dues au mouvement), il doit exister pour chaque bloc B_i un bloc B^r_i identique dans l'image de référence. La position de B^r_i est donc la position de B_i à l'instant $t-dt$. En réalité, de petites variations de luminance apparaissent, liées, d'une part, aux propriétés physiques intrinsèques de la lumière environnante et du capteur de la caméra et, d'autre part, à la non-invariance par translation de la TOD. B^r_i est alors défini comme le bloc le plus ressemblant en terme de luminance. Cette ressemblance est quantifiée en pratique par le critère de la Moyenne des Différences des valeurs Absolues (MAD, "Mean of Absolute Differences", équation 4.7).

$$MAD_{B_i}(dx, dy) = \frac{1}{Card(\mathcal{B})} \sum_{(x,y)\in\mathcal{B}} |LLY^K_t(x, y) \; - \; LLY^K_{t-dt}(x + dx, y + dy)| \qquad (4.7)$$

avec $\mathcal{B} = \mathcal{B}_{B_i,d_x,d_y,LL^K_{t-dt}} = \{(x, y) \in B_i, \setminus (x+dx, y+dy) \in LL^K_{t-dt}\}$, l'ensemble des couples (x, y) de coordonnées des points du bloc B_i dans le référentiel image qui, déplacé de (dx, dy) sont encore dans le domaine de définition spatial de LL^K_{t-dt} ; $Card(\mathcal{B})$ est le cardinal de $\mathcal{B}$.$LLY^K_t(x, y)$ (respectivement $LLY^K_{t-dt}(x + dx, y + dy)$) désigne la valeur de luminance de la sous-bande LL^K de l'image I_t (resp. I_{t-dt}) au point (x, y) (resp. $(x + dx, y + dy)$).
Le déplacement (dx_i, dy_i) de B_i est alors défini par l'équation 4.8.

$$(dx_i, dy_i) = \underset{(dx,dy)\in\mathcal{D}_{B_i}}{argmin} \; MAD_{B_i}(dx, dy) \qquad (4.8)$$

avec $\mathcal{D}_{B_i} = \{(dx, dy), \exists(x, y) \in B_i \setminus (x + dx, y + dy) \in LL^K_{t-dt}\}$ l'ensemble de tous les déplacements possibles du bloc B_i pour lesquels au moins une position $(x + dx, y + dy)$ est dans le domaine de définition spatial de LL^K_{t-dt}.

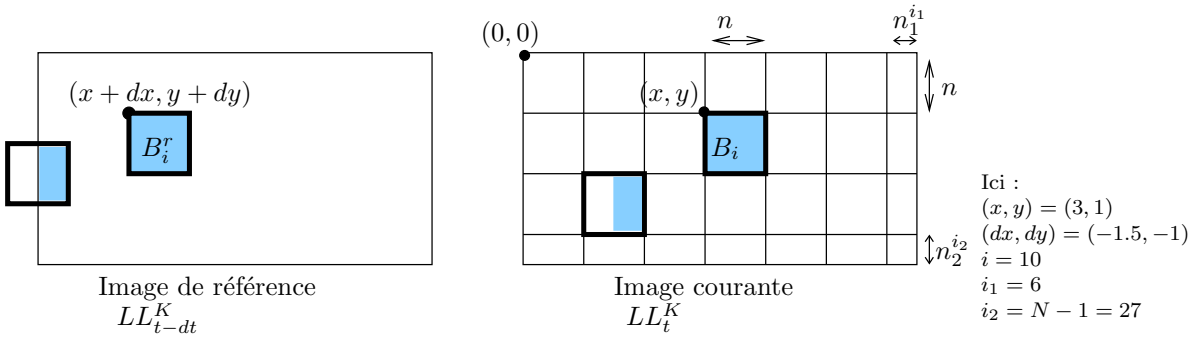


Fig. 4.2 – Principe de la MCB. Les zones colorées indiquent les parties du bloc qui servent au calcul du MAD (4.7)

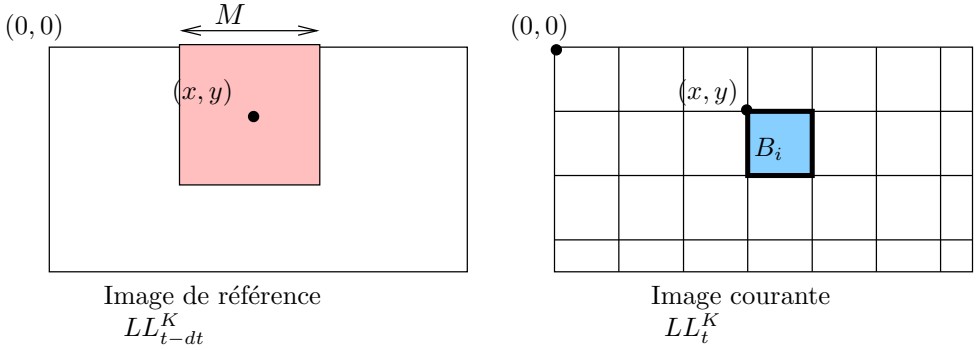


Fig. 4.3 – Principe de la RE. Les zones colorées indiquent un bloc dans l'image courante et la zone de recherche correspondante dans l'image de référence

En pratique, afin de réduire les temps de calcul, la recherche du vecteur de déplacement optimal (dx_i, dy_i) de B_i n'est pas effectuée de manière exhaustive sur l'espace $\mathcal{D}_{B_i}$ mais de manière approchée sur un espace $\mathcal{D}'_{B_i} \subset \mathcal{D}_{B_i}$; seuls certains blocs de l'image de référence sont testés. Plusieurs méthodes de réduction de l'espace de recherche existent. Nous nous limitons dans cette étude à la Recherche Exhaustive RE dans laquelle l'espace de recherche $\mathcal{D}'_{B_i}$ est une fenêtre de taille fixe M centrée sur la position initiale du bloc B_i (figure 4.3).

Enfin, aucune indication quant à la quantification des vecteurs de déplacements n'a encore été donnée. Naturellement, seules les valeurs des intensités des pixels aux coordonnées entières des repères de l'image courante et de l'image de référence sont connues. Cela conduirait à ne tester que des valeurs de déplacements de précision de 1 pixel. Cependant, une telle approximation paraît très réductrice et peu réaliste, la probabilité pour que le mouvement apparent soit effectivement une bijection pixel à pixel est très faible. Il est possible de définir des pas de recherche plus fins en interpolant les valeurs d'intensité à partir des coordonnées entières. Nous nous limiterons dans cette étude au pas de recherche spatial p_s de 1 et 1/2 pixels (4.9) ; les valeurs d'intensité aux coordonnées non-entières étant

obtenues par interpolation bilinéaire.

$$(dx, dy) = (p_s A, p_s B) \text{ avec } (A, B) \in \mathbb{N}^2 \tag{4.9}$$

Ré-estimation par MCB

Dans ce contexte, une information supplémentaire par rapport à la BR est ajoutée sous la forme d'un champ de vecteurs initial. La démarche de MCB reste identique. La nouvelle information est prise en compte dans l'étape de RE : la fenêtre de recherche est dans ce cas centrée sur la position indiquée par le vecteur initial et plus sur la position du bloc dans l'image courante.

Paramétrage Numérique de l'algorithme

L'algorithme de MCB que nous nous proposons d'utiliser sur la séquence de sous-bandes LL de BR dépend de 4 paramètres : la taille des blocs (n), le pas temporel (dt, espacement temporel entre deux images successives), la taille de la fenêtre de recherche (M) et le pas de recherche spatial (p_s).

Taille des Blocs (n) : La finalité de notre travail est de parvenir à extraire des objets d'intérêt en mouvement propre. Nous considérons qu'un objet est significatif s'il est d'une taille d'au moins 15% de la taille de l'image. Les images de HD utilisées ont une taille de 1920x1080 pixels (section 7.2). Les sous-bandes LL de Basse Résolution utilisées pour l'estimation de mouvement sont prises au niveau de décomposition $K = 4$ et ont donc une résolution $2^K = 16$ fois plus petite soit 120x68. La taille minimale à détecter est alors prise sur la hauteur et est de $\lfloor 68 * 15/100 \rfloor = 10$. Afin de décrire au mieux le lieu des objets en mouvement, nous souhaitons les définir avec plusieurs blocs par opposition à un bloc isolé qui serait dû à une mauvaise estimation du mouvement. Par ailleurs, dans une estimation hiérarchique, les tailles des blocs sont pris égales à des multiples de 2 pour être cohérent avec le sous-échantillonnage dyadique. Nous choisissons $n = 4$.

Pas temporel (dt) : Afin d'observer un mouvement en haut de la pyramide, il faut que le mouvement à la pleine résolution soit significatif. Ainsi, un déplacement de 1 pixel à BR correspond à un mouvement de $2^K = 16$ pixels à pleine résolution. Pour avoir de si grands déplacements, il faut choisir un pas temporel dt suffisant. Notre choix de dt est guidé par le temps de fixation du système visuel humain qui est d'environ 200ms [LM06]. Sur ce temps, le mouvement est suffisamment cohérent pour pouvoir apparier les images. Pour les vidéos utilisées qui sont échantillonnées à 25 images/s cela correspond à $dt = 5$. Notons que dans le contexte de la scalabilité, la résolution temporelle du flux transmis peut être plus faible. Dans la suite de ce travail et de nos expérimentations, nous n'avons pas imposé de contraintes sur la scalabilité temporelle. Dans le cas contraire, il est toujours possible de choisir un pas proche de celui défini ici.

Taille de la fenêtre de recherche (M) : Pour tirer pleinement partie de l'utilisation d'une RE nous avons défini empiriquement une taille de fenêtre de recherche suffisant grande

(a) lancer_trousse

(b) pieton_exterieur

(c) pedestrian_area

Fig. 4.4 – Résultats de la MCB à BR pour trois séquences issues du corpus ICOS-HD. L'image de référence est à gauche, l'image courante au milieu et les vecteurs de mouvement estimés sont représentés sur l'image de droite. (Pour la lisibilité de l'affichage, $K = 3$ et la taille des vecteurs est doublée).

pour capter une large gamme de mouvements. Il s'agit de $M = 5$ pour la BR (à HR, cela revient à détecter des mouvements d'amplitude maximale 16 pixels entre deux images éloignées de $dt = 1$).
Pas de recherche spatial (p_s) : Le pas de recherche spatial est fixé au demi-pixel près pour rester en accord avec ce qui ce fait dans les standards de compression vidéo. On a $p_s = 0.5$.

Dans le cadre de la phase de ré-estimation, les paramètres dt et p_s restent identiques à ceux fixés à BR. La taille des blocs n est fixé par la projection qui est détaillée dans la section 4.2.5. Il reste à déterminer la taille de la fenêtre de recherche M. Nous la fixons à $M = 2$ car l'ajustement, si l'initialisation est correcte, est limité par le facteur de sous-échantillonnage de la pyramide d'ondelettes, ici 2. Les techniques proposées dans la suite de ce chapitre assurent que l'hypothèse d'une initialisation correcte est vérifiée.

La figure 4.4 présente le résultat de l'estimation de mouvement par BM sur un couple

de sous-bandes LL de BR avec les paramètrages indiqués pour 3 séquences de notre corpus. La séquence "pedestrian_area" (figure 4.4c) présente un faible mouvement de translation de la caméra et des mouvements de personnes forts. Les vecteurs du fond de la scène sont visuellement réguliers et traduisent bien le mouvement global, les objets ont quant à eux des vecteurs très différents permettant la distinction objet/fond. La séquence "lancer_trousse" (figure 4.4a) est une séquence en caméra fixe. Là-encore, la plupart des vecteurs de fond sont visuellement corrects mais quelques vecteurs, dans les zones plates, sont mal estimés. Enfin, dans la séquence "pieton_exterieur" (figure 4.4b), les vecteurs de mouvement dans le fond sont bien estimés seulement vers le milieu de l'image et l'objet est "noyé" au milieu de vecteurs erronés.

4.2.3 Estimation du mouvement global

Les vecteurs de mouvement obtenus précédemment à BR sont fortement bruités (cf figure 4.4). Pour corriger les vecteurs erronés, nous proposons d'utiliser le mouvement global de la caméra (cf figure 4.1, blocs fonctionnels "Estimation du mouvement global" et "Correction par EMG"). Nous détaillons dans cette section la méthode d'EMG (son utilisation pour corriger les vecteurs de mouvement sera expliquée plus loin). Elle est adaptée des travaux de Durick et al. [Dur01], eux mêmes inspirés de ceux de Odobez et Bouthemy [Odo95], et a notamment été utilisée avec succès par Kraëmer et al. [Kra05] comme base pour l'identification des mouvements de caméra en utilisant les vecteurs de déplacement des macro-blocs du flux MPEG2.

Le mouvement global définit le mouvement du contenu principal de la scène et est principalement dû aux mouvements de la caméra ou aux changements de focus. Nous proposons de le caractériser par un modèle affine à 6 paramètres, $\theta = (a_1, a_2, ..., a_6)^T$ (cf (4.6)) avec l'origine des coordonnées des points au centre de l'image. En considérant les N couples constitués des mesures (les vecteurs de mouvement) et des observations (les positions des blocs) et après changement de repère, on obtient un système linéaire d'équations pouvant être résumé sous la forme matricielle (4.10) suivante :

$$Z = H\theta \tag{4.10}$$

où

$$Z = (dx_1, \cdots, dx_N, dy_1, \cdots, dy_N)^T \tag{4.11a}$$

$$H = \begin{pmatrix} 1 & x_1 & y_1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & y_N & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 & x_N & y_N \end{pmatrix} \tag{4.11b}$$

Z est le vecteur de mesures représentant les vecteurs de mouvement des blocs (ces mesures sont bruitées) et H est la matrice d'observation contenant les centres des blocs, l'origine du repère étant placée au centre de l'image. Nous proposons d'utiliser la méthode des moindres carrés pondérés pour résoudre ce système sur-déterminé d'inconnue θ, l'estimée $\hat{\theta}$ étant alors calculée par :

$$\hat{\theta} = (H^T W H)^{-1} H^T W Z \tag{4.12}$$

où

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 & 0 \\ 0 & w_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & w_{2N-1} & 0 \\ 0 & 0 & \cdots & 0 & w_{2N} \end{pmatrix} \tag{4.13}$$

La matrice W est une matrice de poids w_i, avec $w_i \in [0, 1]$. Ces poids permettent d'accorder plus ou moins d'importance à chaque mesure v_i, suivant qu'elle est proche du modèle (poids fort, mesure faiblement bruitée) ou loin du modèle (poids faible, mesure fortement bruitée). Les poids sont fonctions des résidus r_i définis par (4.14) :

$$r_i = v_i - v_i(\hat{\theta}) \tag{4.14}$$

où v_i est la mesure du $i$ème vecteur de déplacement ($i \in [1, ..., N]$) et $v_i(\hat{\theta})$ son estimation calculée par (4.6) en utilisant le modèle estimé $\hat{\theta}$. La fonctionnelle utilisée est, comme suggéré par [Odo95], l'estimateur de Tukey ρ (4.15) de dérivée ψ (4.16).

$$\rho(r_i, \lambda_r) = \begin{cases} \frac{r_i^6}{6} - \frac{2\lambda_r^2 r^4}{4} + \frac{\lambda_r^4 r_i^2}{2} & \text{si } |r_i| < \lambda_r \\ \frac{\lambda_r^6}{6} & \text{sinon} \end{cases} \tag{4.15}$$

$$\psi(r_i, \lambda_r) = \begin{cases} r_i(r_i^2 - \lambda_r^2)^2 & \text{si } |r_i| < \lambda_r \\ 0 & \text{sinon} \end{cases} \tag{4.16}$$

avec λ_r un seuil à fixer par expérimentations. Dans la méthode des moindres carrés pondérés, le poids associé à la $i$ème mesure est défini par $\psi(r_i)/r_i$. Après normalisation, la formule d'obtention du poids w_i est donnée par (4.17) :

$$w_i = \frac{\psi(r_i, \lambda_r)}{\lambda_r^4 r_i} \tag{4.17}$$

La définition des poids est donc conjointe à l'estimation des paramètres. Comme dans [Odo95, Dur01], nous utilisons un schéma multirésolution. Deux pyramides dyadiques, une des mesures et une des observations, de K_{EMG} niveaux et respectant la structure spatiale des images, sont construites par moyennage des valeurs des vecteurs de mouvement et des centres des blocs. La stratégie de résolution va du haut de la pyramide (là où les données sont le plus réduites) vers le bas. En haut de la pyramide, les poids sont pris égaux à 1. Le modèle $\hat{\theta}^{k_{EMG}}$ est estimé. Pour chaque niveau suivant, les poids sont initialisés grâce à la valeur du modèle $\hat{\theta}^{k_{EMG}-1}$ trouvé au niveau supérieur et $\hat{\theta}^{k_{EMG}}$ est estimé au niveau courant par (4.12).

Pour améliorer la qualité des résultats, nous avons pris en considération le fait que les éléments a_i de θ n'ont pas tous le même ordre de grandeur. Ainsi, a_1 et a_4 sont des paramètres de translation et sont mesurés en unités de distance (ici, le nombre de pixels). A contrario, a_2, a_3, a_4 et a_5 sont des grandeurs sans dimensions et généralement inférieures à 1. Cette constatation nous amène à estimer un modèle à 2 paramètres en haut de la pyramide et à n'introduire le modèle à 6 paramètres que dans les niveaux inférieurs de la pyramide. En effet, comme dans toute représentation multirésolution, le haut de la pyramide informe sur le mouvement dominant de la scène, les niveaux inférieurs successifs contenant plus de détails. L'amélioration apportée par la considération séparée des paramètres de translation des autres paramètres est illustrée dans la figure 4.5. Nous avons reproduit le champ de vecteur de mouvement global sur l'image sans tenir compte du mouvement propre de l'objet afin de faciliter la visualisation. Dans cette scène (présentée dans la figure 4.4), le mouvement global de la caméra est un mouvement de translation horizontale pure comme l'on obtient avec l'amélioration proposée (figure 4.5, image de droite). En déterminant simultanément tous les paramètres (figure 4.5, image de gauche), des erreurs d'estimation apparaissent sur les bords de l'image mais pas au centre. En effet, considérons une erreur de 10^{-2} dans l'estimation du paramètre a_3. Cette erreur est multipliée par la valeur de y dans le modèle. Pour des valeurs de y faibles, vers le centre de l'image, cette erreur est négligeable. Pour des valeurs de y fortes, telles que $y = 100$, l'erreur engendrée sur dx est de l'ordre de 1 pixel, ce qui n'est pas négligeable à la résolution K de la pyramide d'ondelettes considérée.

Une fois cette méthode appliquée, le mouvement global est déterminé et un poids est

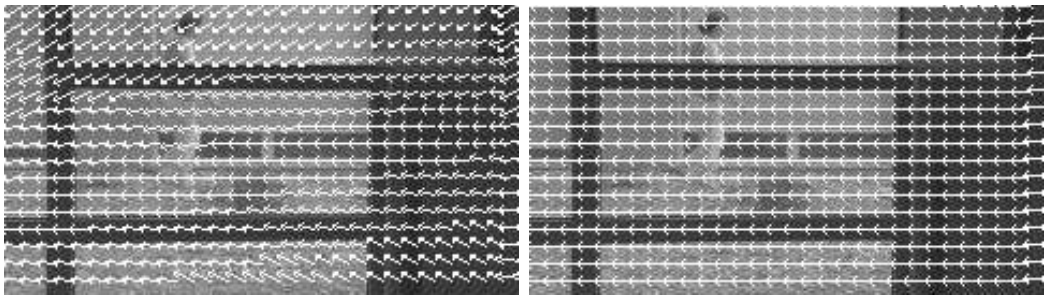


Fig. 4.5 – EMG par multirésolution. Le modèle utilisé en haut de la pyramide est (à gauche) à 6 paramètres et (à droite) à 2 paramètres.

attribué à chaque mesure de vecteur de mouvement. Plus ce poids est proche de 1 et plus la mesure est "conforme" au modèle global. A l'inverse, lorsque le poids est proche de 0, on parle de mesure "non conforme" au modèle global.

Les vecteurs de mouvement estimés par la MCB peuvent être très bruités par rapport au modèle de mouvement global (cf figure 4.4b). Si ces vecteurs erronés sont en nombre dominant par rapport aux vecteurs bien estimés sur le fond, le mouvement global ne peut pas être déterminé correctement. Nous proposons d'appliquer la méthode d'EMG précédente sur un nombre réduit $N_1 \leqslant N$ de mesures fiables. Le rejet des vecteurs de mouvement non fiables se fait en deux étapes.
Etape 1 : La première phase consiste en un rejet explicite des premières ligne et colonne du bord de l'image. En effet, le mouvement de la caméra entraîne l'apparition de zones "entrantes" et "sortantes" qui n'existent pas dans l'image précédente ; les vecteurs associés ont alors une très forte probabilité d'être faux.
Etape 2 : La seconde phase est le rejet des blocs correspondant à une région plate où la MCB a une très forte probabilité de ne pas trouver le vrai vecteur de mouvement. Nous proposons de déterminer si une région est plate en vérifiant si son activité HF est faible. Nous mesurons cette activité par les valeurs d'écart-type des coefficients de HF sur le bloc. Pour un bloc dans l'image courante, le vecteur d'écart-type $\sigma_i = (\sigma_{i,LH}, \sigma_{i,HL}, \sigma_{i,HH})^T$des coefficients HF du bloc au niveau k est calculé et comparé à un seuil donné 4.18.

$$\sigma_i < T_\sigma^k <=> \begin{cases} \sigma_{i,LH} < T_{LH}^k \\ \sigma_{i,HL} < T_{HL}^k \\ \sigma_{i,HH} < T_{HH}^k \end{cases} \tag{4.18}$$

Le seuil $T_\sigma^k = (T_{LH}^k, T_{HL}^k, T_{HH}^k)^T$ est adaptatif au niveau k de la pyramide pour tenir compte de l'influence du bruit : à Haute Résolution, la plupart du bruit se retrouve dans les sous-bandes de HF, alors qu'à Basse Résolution, le signal a été filtré et le bruit réduit. Ces seuils ont été défini de façon empirique à partir du corpus de test ICOS-HD (cf section 7.2). Les

type sous-bande / niveau	LH (T^k_{LH})	HL (T^k_{HL})	HH (T^k_{HH})
4	$3.162.10^{-6}$	$7.94.10^{-7}$	$7.94.10^{-7}$
3	$3.981.10^{-6}$	$1.585.10^{-6}$	$1.585.10^{-6}$
2	10^{-6}	10^{-6}	$1.585.10^{-6}$
1	$1.995.10^{-6}$	10^{-6}	$3.16.10^{-7}$

Tab. 4.1 – Seuils empiriques de rejet d'activité HF

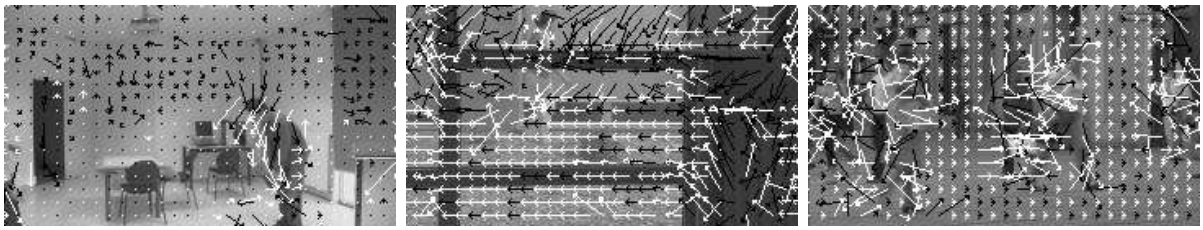


Fig. 4.6 – Vecteurs exclus par l'étape 2 du rejet. Ces vecteurs sont représentés en noir. (Pour la lisibilité de l'affichage, $K = 3$ et la taille des vecteurs est doublée).

seuils obtenus sont résumés dans le tableau 4.1. Des précisions sur la méthode empirique utilisée pour déterminer ces seuils peuvent être trouvées dans l'annexe B. Rappelons que nous travaillons sur les coefficients d'ondelettes extraits du flux compressé juste après la déquantification. Ces coefficients ont donc été calculés à partir d'images à valeurs normalisées dans $[-\frac{1}{2}, \frac{1}{2}[$, ce qui explique l'ordre de grandeur des variances calculées.

La figure 4.6 représente les vecteurs exclus par cette étape de rejet pour les trois couples d'images de la figure 4.4. Dans les séquences "lancer_trousse" et "pieton_exterieur", les vecteurs visuellement mal estimés par la MCB sur les zones plates ont bien été rejetés. En contrepartie, certains vecteurs bien estimés dans des zones peu texturées des séquences "pieton_exterieur" et "pedestrian_area" sont considérés comme non fiables. Il reste cependant suffisamment de vecteurs dans le fond pour permettre une estimation du mouvement global correcte. Enfin, des vecteurs correspondants aux objets en mouvement ont été exclus. Comme ceux-ci ne correspondent pas au mouvement global, leur rejet est acceptable.

Paramétrage Numérique de l'algorithme

L'algorithme d'EMG proposé dépend de deux paramètres : le nombre de niveaux K_{EMG} utilisé dans la décomposition en multirésolution et λ_r le seuil sur les résidus de la fonction de Tukey (4.16).

Nombre de niveaux K_{EMG} : le choix du nombre de niveaux de multirésolution est fait de façon empirique. Sur les données de tests utilisées, nous avons trouvé que $K_{EMG} = 3$ est

suffisant.
Constante de Tukey λ_r : Du fait de l'approche multirésolution utilisée, il nous semble pertinent d'adapter ce seuil au niveau considéré. A pleine résolution, nous supposons que la MCB fournit un nombre suffisant de vecteurs qui suivent le mouvement global. Du fait de la précision de recherche utilisée dans la MCB, ces vecteurs ont une erreur associée de $\pm p_s$. Le seuil est fixé en fonction de cette erreur à $\lambda_r^0 = 2p_s$. Aux niveaux supérieurs de la pyramide, ce seuil doit être relâché. Nous proposons une loi expérimentale $\lambda_r^{k_{EMG}} = 2^{k_{EMG}} * \lambda_r^0$. On rappelle que $\lambda_r^0 = 2p_s = 1.0$.

4.2.4 Fonction caractéristique des valeurs non conformes. Application à l' extraction des masques des objets en mouvement.

Le processus d'EMG a permis non seulement de résumer le mouvement dominant par 6 paramètres mais aussi de séparer les N mesures de vecteurs de déplacement obtenues par la MCB en 3 catégories symbolisées sur la figure 4.1 en face du bloc fonctionnel "Estimation du mouvement global" :

1. les mesures rejetées a priori car considérées comme non fiables (étape 2 du rejet, signalées par une croix),
2. les mesures non conformes au modèle global (de poids $w_i \leqslant T_w$ faible, signalées par un point),
3. les mesures conformes au modèle global (de poids $w_i > T_w$ fort, sans signe particulier associé).

Les mesures des points 1 et 2 peuvent être classées en deux catégories : soit un nouveau mouvement a été induit par un objet en mouvement propre dans la scène et l'on parle de valeur "non conforme" au modèle global, soit le vecteur est erroné du fait d'une mauvaise estimation par la MCB. Dans ce deuxième cas, nous pensons qu'il est plus intéressant de corriger le vecteur erroné par son estimation par le modèle de mouvement global que de le conserver (cf figure 4.1, bloc fonctionnel "correction par EMG"). Pour déterminer si un vecteur a besoin d'être corrigé, nous utilisons la mesure de qualité de reconstruction du bloc après compensation de mouvement définie dans la suite de cette section. De plus, nous introduisons dans ce paragraphe la fonction caractéristique de "valeurs non conformes".

Rappelons que la majorité des cas où la MCB abouti à un vecteur erroné est due au fait que le bloc B_i correspondant appartient à une zone plate de la scène. En effet, dans une telle situation, la différence entre les MAD (4.7) pour différents vecteurs, pointant tous

dans la même zone plate, est faible et de l'ordre de grandeur du bruit. Ainsi, dans un tel cas, la valeur de MAD obtenue pour le vecteur modélisé $v_i(\theta)$ sera très proche de celle trouvée avec le vecteur estimé v_i. A contrario, pour un bloc appartenant à un objet en mouvement propre, il est raisonnable d'évaluer que cette différence sera forte. Le critère de qualité de reconstruction du bloc après compensation de mouvement est donc défini par la différence entre les mesures de MAD pour le vecteur modélisé $v(\theta)$ et le vecteur estimé v : $MAD_{B_i}(v(\theta)) - MAD_{B_i}(v)$. Le seuillage par T^k_{MAD} de ce critère permet de déterminer quels vecteurs ont besoin d'être corrigés. La fonction caractérisque de valeurs "non conformes" est donc définie par

$$f_o(v) = \begin{cases} 1 & \text{si } (w_i < T_w \vee \sigma_{i,HF} < T^k_\sigma) \wedge (MAD(v(\theta)) - MAD(v) > T^k_{MAD}) \\ 0 & \text{sinon} \end{cases} \tag{4.19}$$

Le critère $(w_i < T_w \vee \sigma_{i,HF} < T^k_\sigma)$ permet de ne tester la qualité de reconstruction que sur des vecteurs soit "non conformes" soit non fiables. Le seuil T^k_{MAD} est adaptatif au niveau de résolution k de la pyramide d'ondelettes pour tenir compte de l'influence du bruit. Ces seuils ont été défini de façon empirique (cf. annexe B) à partir du corpus de test ICOS-HD (cf section 7.2). Le seuil commun obtenu à tous les niveaux est 0.001. Des précisions sur la méthode empirique utilisée pour déterminer ces seuils peuvent être trouvées dans l'annexe B.

La fonction caractéristique ainsi définie indique tous les blocs de la MCB ayant un mouvement "non conforme" au modèle global. Du fait des corrections que nous avons définies, ces mouvements sont dus à la présence d'objets en mouvement propres et non à des erreurs d'estimation. La fonction caractéristique indique donc le lieu des objets en mouvement. Nous notons M^k_t le masque binaire de mouvement où un pixel à une valeur de 1 si il appartient à un bloc tel que $f_o(v) = 1$ et une valeur de 0 dans les autres cas.

Paramétrage Numérique de l'algorithme
La définition de la fonction f_o nécessite la définition de deux paramètres : le seuil de conformité des vecteurs T_w et le seuil de qualité de reconstruction T^k_{MAD}. Le *Seuil de conformité* T_w est pris à $\lambda_r = 0.4$ afin de tenir compte de la valeur du seuil de rejet des valeurs aberrantes $\lambda_r = 2 * p_s$, p_s étant l'erreur tolérée sur le vecteur de mouvement. Le seuil de qualité de reconstruction a été défini empiriquement à $T_{MAD} = 0.001$.

Nous présentons ici une comparaison des masques obtenus sans la correction et avec la correction proposée sur la séquence "Pieton Exterieur, Clip2" ©LaBRI (Figure 4.7). Les blocs correspondants à des vecteurs non fiables ou considérés comme "non conformes" par l'estimateur de mouvement global sont marqués en noir. Les blocs ayant servi de support

(a) (b)

Fig. 4.7 – Blocs dont le vecteur de mouvement est “non conforme” au modèle de mouvement global (a) avant et (b) après correction par critère de qualité basé MAD

à l'estimation du mouvement global sont indiqué en blanc sur la figure 4.7 (a). L'image de droite (4.7 (b)) montre le masque de mouvement obtenu après application de notre critère de qualité de reconstruction. Les blocs décrivant l'objet en mouvement sont (presque) les seuls à être étiquetés “non conformes” après l'application de notre traitement.

4.2.5 Estimation multirésolution

L'objet de ce paragraphe est l'opération de projection encadré par un trait pointillé dans le schéma général de la méthode (figure 4.1). La figure 4.8 reprend la partie de la figure 4.1 correspondant à la projection.

Des étapes précédentes, nous avons distingué deux types de vecteur : les vecteurs “non conformes” au modèle de mouvement global (marqués d'un point sur la figure 4.8) et les vecteurs “conformes”. Nous proposons de traiter la projection de ces deux types de vecteurs de façon différente. Ainsi, les vecteurs “non conformes” correspondent à des blocs susceptibles d'appartenir à un objet en mouvement propre. Certains de ces blocs contiennent alors les frontières entre objet et fond. Il faut donc conserver un découpage fin aux niveaux de résolution supérieure pour bien capter le mouvement. Aussi nous projetons chaque bloc correspondant du niveau k en $p \times p$ blocs de niveau $k-1$, où $p = 2$ est le facteur de sous-échantillonage de la pyramide. Les vecteurs de mouvement initiaux de ces blocs de niveau $k-1$ sont initialisés chacun avec la projection du vecteur de mouvement trouvé au niveau k. Dans le cas où le vecteur de mouvement est “conforme” au modèle global, le mouvement peut être approximé par le mouvement global et n'a besoin d'être affiné que modérément. Chaque bloc correspondant au niveau k est projeté sur un seul bloc au niveau $k-1$ de taille pn où n est la taille du bloc au niveau k. Le vecteur de mouvement au niveau $k-1$ est initialisé par la projection du vecteur déduit du modèle global $v(\theta)$. L'approximation par le modèle aide à régulariser les vecteurs de mouvement de l'arrière-plan. Cela compense aussi les artefacts introduits par la non-invariance par translation de la TOD.

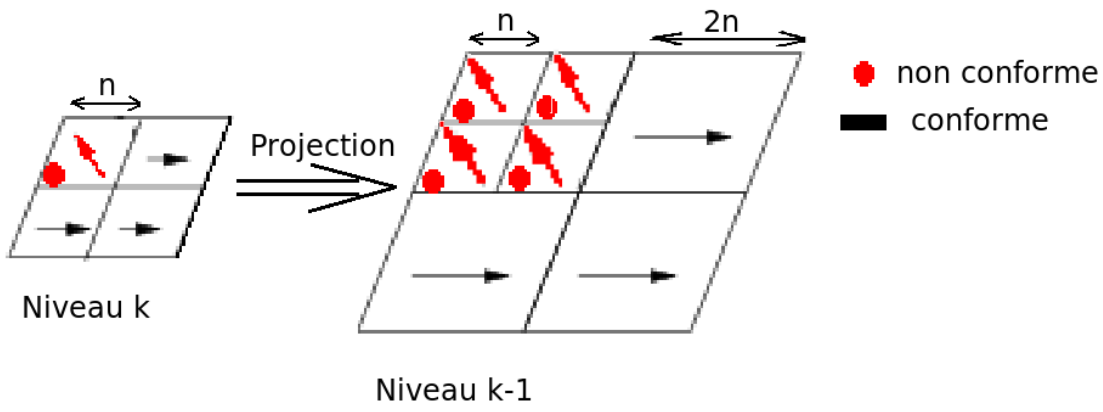


Fig. 4.8 – Détail du principe de projection des vecteurs de mouvement

4.3 Résultats et évaluation de la méthode

Le but de cette section est de présenter quelques résultats supplémentaires obtenus par notre méthode d'estimation de mouvement sur les pyramides d'ondelettes pour permettre au lecteur de juger de ses performances. L'évaluation d'une telle méthode par des outils conventionnels est peu adaptée. Dans un premier temps, nous allons évaluer les performances de la méthode en terme de PSNR de reconstruction. Les valeurs de références seront données par la comparaison avec une Mise en Correspondance de Blocs hiérarchique traditionnelle sur une pyramide gaussienne. Puis, toujours en terme de PSNR, nous chercherons à savoir si l'approximation par le modèle de mouvement global améliore ou non les résultats par rapport à une MCB.

Les résultats sont calculés pour la séquence "pieton_exterieur" du LaBRI. Nous avons choisi cette séquence car elle rassemble plusieurs caractéristiques correspondant aux conditions les plus courantes de vidéo : mouvement de caméra de translation, objet complexe articulé en mouvement propre, zones plates, zones texturées et zones fortement texturées. Afin de montrer l'importance du calcul du mouvement global, nous avons comparé les PSNR de reconstruction obtenus respectivement avec les vecteurs de MCB, avec le mouvement global et sans correction de mouvement (cf. figures 4.9 et 4.10).

Bien sûr, la correction par le modèle global ne permet pas d'améliorer le PSNR de reconstruction par rapport à la MCB. Mais on voit bien que ce soit au niveau 4 de la pyramide (figure 4.9), c'est-à-dire la Basse Résolution, aussi bien qu'au niveau 0 (figure 4.10) que la perte en PSNR est limitée et que l'estimation par le mouvement global reste meilleure que sans compensation de mouvement. Il arrive en de très rare occasions que cette perte soit très forte et conduise à des PSNRs très faibles. Cela peut se produire si le mouvement global est mal estimé. Cela ne prouve pas cependant que notre correction

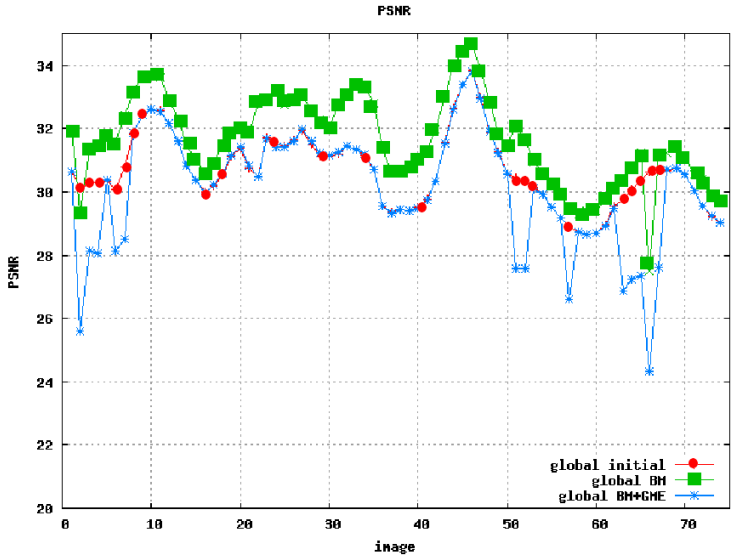


Fig. 4.9 – Comparaison des PSNR de reconstruction de la séquence “Pieton_exterieur” obtenus sans compensation de mouvement, avec une MCB hiérarchique classique et avec notre approche, au niveau 4 de la pyramide

n’est pas adéquate ou que l’estimateur de mouvement global n’est pas performant. Il s’agit seulement des erreurs normales due à la réalité qu’une méthode automatique ne peut pas réussir à tous les coups. Comme notre méthode permet cette distinction, nous comparons les PSNRs de reconstruction uniquement sur le fond de la scène et sur l’objet. Sur l’objet, du fait de l’initialisation par les résultats de la MCB de notre méthode, les résultats sont identiques. Sur le fond, on constate encore que ce soit à Basse Résolution (figure 4.11) ou à Haute Résolution (figure 4.12) que la perte de PSNR est négligeable.

Afin de compléter cette étude, nous avons comparé les PSNR de reconstruction obtenus avec ceux que l’on peut obtenir par une MCB classique sur une pyramide gaussienne. Si le résultat du PSNR au niveau de plus faible résolution (figure 4.13) n’est pas très bon avec notre méthode on voit qu’au fil des niveaux, la correction par le mouvement global porte ses fruits et notre méthode supplante la méthode classique sur pyramide gaussienne (figure 4.14).

4.4 Conclusion

Dans ce chapitre, nous avons proposé une méthode d’estimation de mouvement dans le domaine des ondelettes obtenu par compression JPEG2000. La méthode doit être scalable

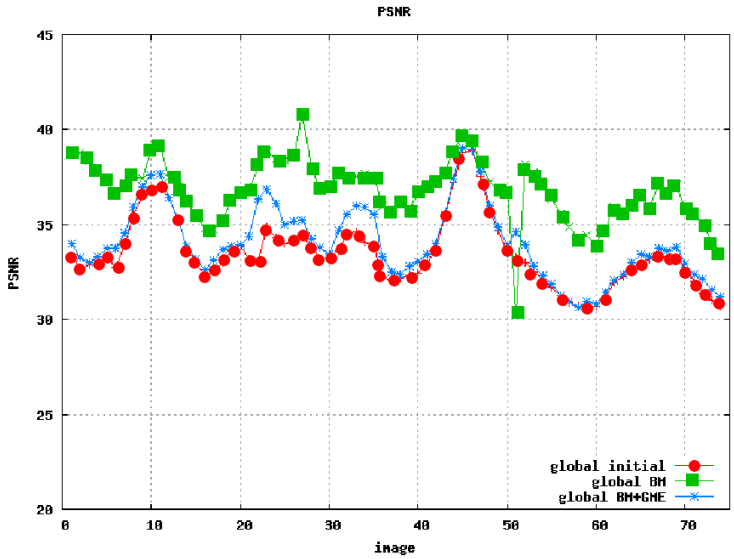


Fig. 4.10 – Comparaison des PSNR de reconstruction de la séquence "Pieton_exterieur" obtenus sans compensation de mouvement, avec une MCB hiérarchique classique et avec notre approche, au niveau 0 de la pyramide

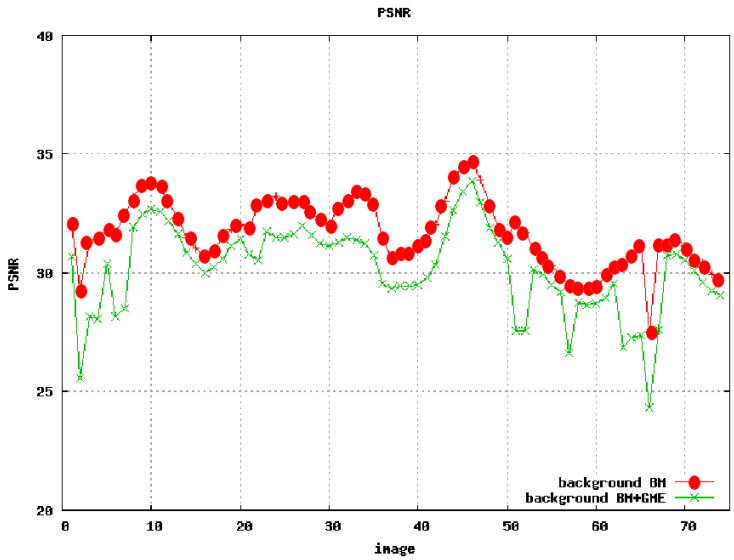


Fig. 4.11 – Comparaison, uniquement sur les pixels décrivant le fond, des PSNR de reconstruction de la séquence "Pieton_exterieur" obtenus avec une MCB hiérarchique classique et avec notre approche, au niveau 4 de la pyramide

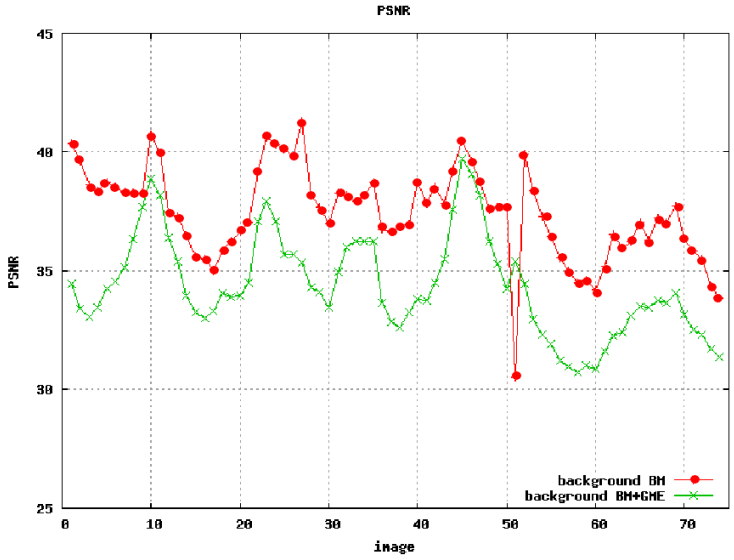


Fig. 4.12 – Comparaison, uniquement sur les pixels décrivant le fond, des PSNR de reconstruction de la séquence “Pieton_exterieur” obtenus avec une MCB hiérarchique classique et avec notre approche, au niveau 1 de la pyramide

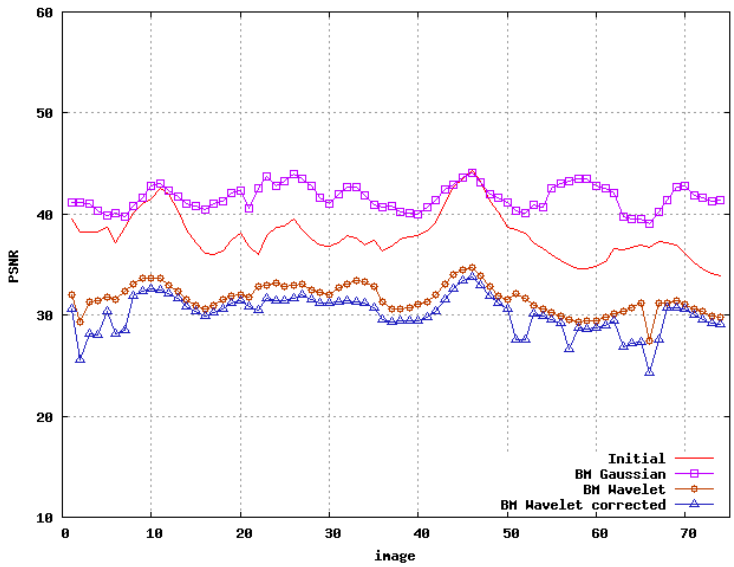


Fig. 4.13 – Comparaison des PSNR de reconstruction de la séquence “Pieton_exterieur” obtenus par notre méthode et une méthode de MCB hiérarchique classique sur pyramide gaussienne, au niveau 4 de la pyramide

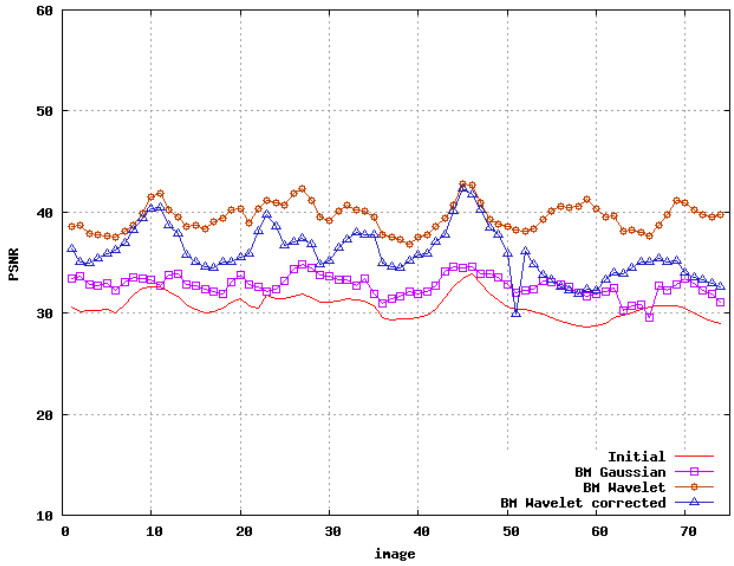


Fig. 4.14 – Comparaison des PSNR de reconstruction de la séquence "Pieton_exterieur" obtenus par notre méthode et une méthode de MCB hiérarchique classique sur pyramide gaussienne, au niveau 1 de la pyramide

dans le sens où elle doit suivre la progression de transmission du flux JPEG2000. Autrement dit, elle doit évaluer le mouvement en utilisant l'information de Basse Résolution en premier lieu, puis utiliser les informations de moyenne et Haute Résolution au fur et à mesure de leur transmission.
La méthode de MCB hiérarchique est une technique classique du traitement des vidéos. L'originalité de notre approche est de proposer une correction par le modèle de mouvement global. Nous pouvons ainsi traiter les régions de l'image correspondant à des objets en mouvement avec précision. Sur le fond, une telle précision n'est pas nécessaire, une approche moins fine est utilisée pour alléger les temps de calcul. De plus, nous proposons d'approximer les vecteurs de mouvement du fond par le modèle de mouvement global. De cette façon, nous régularisons tout au long de la pyramide l'estimation du mouvement dans le fond de la scène. De plus, les vecteurs "aberrants" mal estimés par la MCB sont corrigés et ne sont donc pas conservés dans les niveaux de résolution inférieure.
Enfin, nous avons proposer d'utiliser le rejet des valeurs de vecteurs de mouvement non fiables avant de calculer le mouvement global. Cela permet de rendre plus robuste encore l'estimation du modèle global par la méthode des moindres carrés pondérés classique.

Chapitre 5

Extraction spatiale d'objets guidée par l'information temporelle de deux images

Une des approches pour faire de l'analyse et de l'indexation vidéo est de faire l'analyse et l'indexation des images-clés de cette vidéo. Nous proposons, par analogie, une méthode d'extraction spatio-temporelle d'objets en mouvement sur un couple d'images ; le couple d'images étant la plus petite unité permettant d'obtenir une information de mouvement.

5.1 Vue d'ensemble de la méthode proposée

Le principe de la méthode est décrit dans la figure 5.1. La méthode de segmentation spatio-temporelle scalable que nous proposons consiste en plusieurs étapes. D'abord, une segmentation spatio-temporelle dans le domaine des ondelettes est accomplie au niveau de résolution le plus bas ($k = K$). Les vecteurs de mouvement grossiers V_t^ksont déterminés entre les trames d'ondelettes W_{t-dt}^k et W_t^k et le masque de mouvement M_t^k, dans lequel les objets d'avant-plan en mouvement sont contenus, est estimé. Une segmentation morphologique couleur est appliquée sur la trame d'ondelettes W_t^k, $k = K$. La carte de segmentation couleur S_t^k et le masque de mouvement M_t^k sont fusionnés, ce qui correspond à l'extraction de l'ensemble des objets d'avant-plan $O_t^k = \{O_{t,i}^k\}$ extraits du couple de trames d'ondelettes (W_t^k, W_{t-dt}^k). De cette façon, l'ensemble des objets d'avant-plan a été extrait à Basse Résolution. Pour les niveaux de résolution supérieure $k = K-1, K-2, ..., 0$, le processus d'extraction commence par la projection, sur $W_t^{(k-1)}$, du masque d'objet O_t^k, de la carte de segmentation S_t^k et des vecteurs de mouvement V_t^k , ce qui donne respectivement $\hat{O}_t^{k-1}$,$\hat{S}_t^{k-1}$ et $\hat{V}_t^k$ (figure 5.1, étapes projection du mouvement et projection couleur).

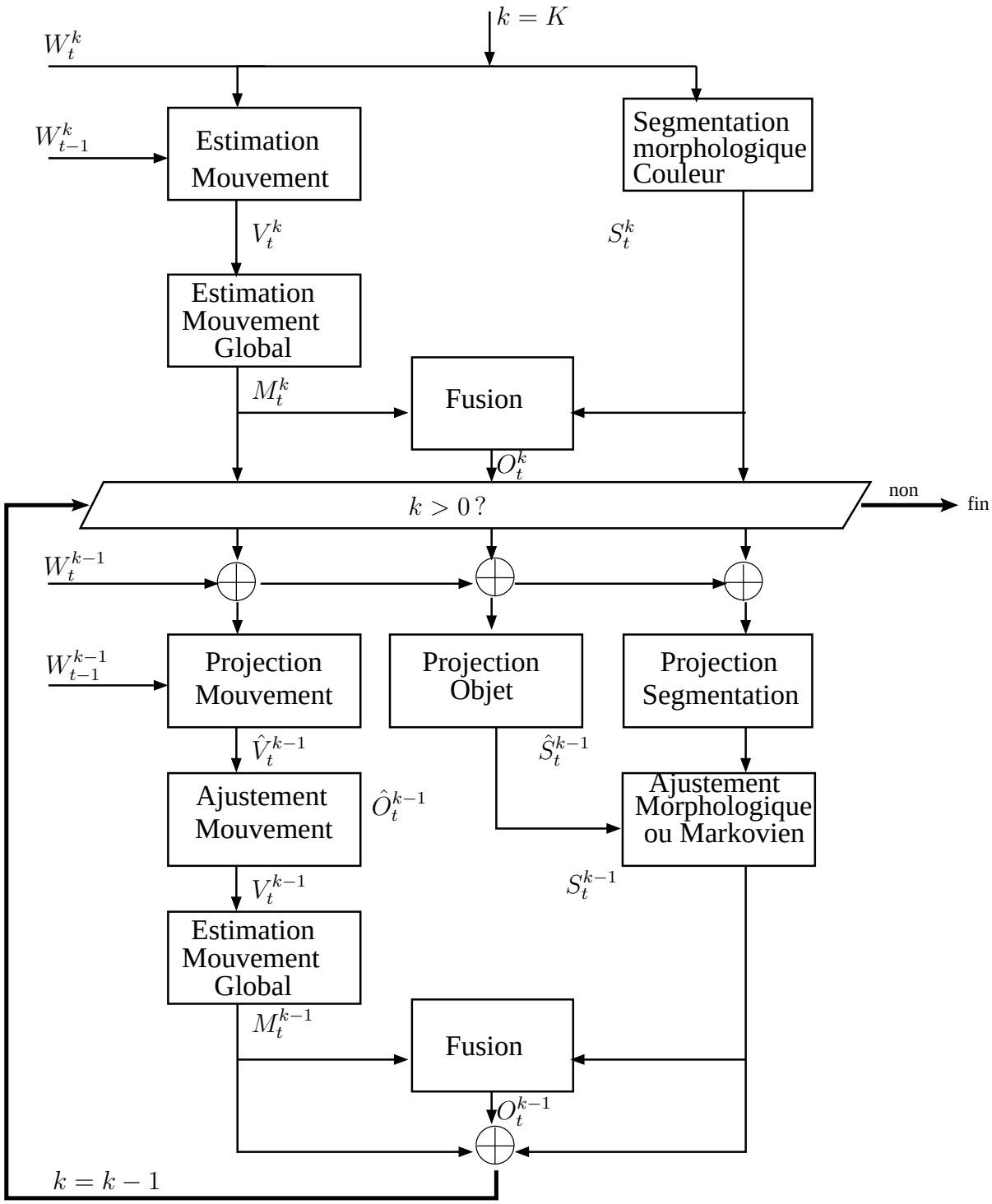


Fig. 5.1 – Schema général d'extraction spatio-temporelle scalable des objets

Fig. 5.2 – Trame d'ondelettes LL à Basse Résolution et masque de mouvement extrait, séquence "lancer trousse, clip2", LaBRI

L'extraction des objets d'avant-plan O_t^k avec $k \in [K-1, 0]$ consiste en l'ajustement de la carte de segmentation projetée $\hat{S}_t^{k-1}$ limitée à l'aire de l'objet $\hat{O}_t^{k-1}$, soit $\hat{S}_t^{k-1} \cap \hat{O}_t^{k-1}$. Deux stratégies d'ajustement de $\hat{S}_t^{k-1} \cap \hat{O}_t^{k-1}$ ont été définies : par approche morphologique et par modélisation markovienne. Dans le même temps, les vecteurs de mouvement sont eux aussi ajustés et le mouvement global est recalculé permettant d'obtenir un nouveau masque de mouvement au niveau $k-1$: M_t^{k-1}.

5.2 Extraction d'objets à Basse Résolution

L'objet de cette section est l'étude de l'extraction spatio-temporelle d'objets à Basse Résolution qui constitue l'étape d'initialisation de la méthode hiérarchique que nous proposons (partie supérieure du schéma de la figure 5.1). Cette extraction est l'adaptation à notre contexte de la méthode proposée par Manerba [Man04].

5.2.1 Segmentation en mouvement

Pour effectuer la segmentation en mouvement, nous utilisons directement les résultats de l'estimation de mouvement dans le domaine des ondelettes que nous avons proposée dans le chapitre 4. Cette méthode définit, entre autres, la fonction caractéristique de valeurs non conformes au modèle global $f_o(v)$ (section 4.2.4). Un vecteur estimé est alors non conforme lorsqu'il est induit par le mouvement propre d'un objet indépendant dans la scène. Rappelons en effet que les vecteurs mal estimés par la MCB, qui constituent une source de valeurs non conformes au modèle global, ont été identifiés grâce au critère (4.19) et corrigés. La représentation par une image de la fonction caractéristique de valeurs non conformes constitue le masque des objets en mouvement (figure 5.2).

La segmentation ainsi obtenue donne une information de localisation des objets en

mouvement dans la scène. Cependant, les masques obtenus sont bruités et peu précis car obtenus à la résolution des blocs utilisés pour l'estimation de mouvement. Ces masques sont par la suite ajustés par combinaison avec le résultat d'une segmentation morphologique couleur à la résolution du pixel de la sous-bande LL^K.

5.2.2 Segmentation morphologique couleur intra-trame

L'algorithme présenté ici est repris des travaux de Manerba et al [Man05]. Il suit le principe de l'algorithme morphologique de Ligne de Partage des Eaux (LPE) avec marqueurs. Les étapes d'un tel processus sont :

- pré-traitement de l'image afin de réduire le bruit
- détermination des marqueurs indicateurs de régions par l'utilisation du gradient morphologique
- croissance des régions pour former la partition de l'image

Une modification est apportée dans la fonction de similarité utilisée dans la croissance de régions afin de tenir compte des caractéristiques du système visuel humain. Les différentes étapes de cette segmentation sont détaillées dans la suite de cette section et illustrées dans la figure 5.3.

5.2.2.1 Pré-traitement

L'étape de pré-traitement (cf. figure 5.3 (b)) a pour but de simplifier l'image avant le calcul du gradient morphologique en atténuant les petites variations d'intensité qui la rende trop bruitée. Cette étape est réalisée à l'aide de filtres par reconstruction qui sont des opérateurs connexes ("connected operators", [Sal95]). Ainsi, un filtre d'ouverture par reconstruction partielle puis un filtre de fermeture par reconstruction partielle sont appliqués sur chaque composante Y, U et V de l'image.
Le filtre d'ouverture est défini dans l'équation (5.1) où LL_t^K est la trame d'ondelettes sur laquelle est appliquée le filtre, $\delta_m()$ est l'opération de dilatation unitaire par un élément structurant M (5.2) appliquée m fois et $\epsilon_n()$ est l'opération d'érosion unitaire par l'élément structurant M (5.3) appliquée n fois. Nous noterons $\tilde{LL_t}^K$ la trame d'ondelettes filtrée.

$$\tilde{LL_t}^K = \gamma_{m,n}(LL_t^K) = \delta_m(\epsilon_n(LL_t^K)) \tag{5.1}$$

$$\delta(f(x)) = \underset{k \in M}{max}\ f(x+k) \tag{5.2}$$

$$\epsilon(f(x)) = \underset{k \in M}{min}\ f(x+k) \tag{5.3}$$

La notation $f(x)$ désigne une fonction d'intensité prise au point de coordonnées x.

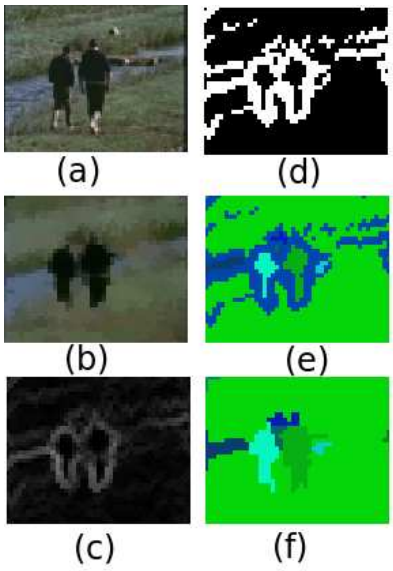


Fig. 5.3 – Illustration du processus de segmentation morphologique à BR. (a) Image originale (b) image pré-traitée (c) gradient morphologique (d) gradient morphologique seuillé (e) étiquetage des régions connexes de gradient nul (f) croissance de régions

5.2.2.2 Gradient morphologique

Le gradient est calculé pour chaque composante Y, U et V de la trame d'ondelettes filtrée $\tilde{LL}_t^K$ (cf. figure 5.3 (c)). Le gradient morphologique se calcule par différence entre la dilatation et l'érosion (5.4).

$$G_{\delta\epsilon}(\tilde{LL}_t^K) = \delta(\tilde{LL}_t^K) - \epsilon(\tilde{LL}_t^K) \tag{5.4}$$

Les trois gradients ainsi obtenus sont ensuite combinés pour obtenir une seule image de gradient G. Pour cela, la valeur maximale entre le gradient Y, le gradient U et le gradient V est sélectionnée en chaque point. Dans l'image ainsi obtenue les régions homogènes, qui sont typiquement les régions au milieu des objets, ont une faible valeur de gradient. A contrario, les pixels correspondants à des contours ont de fortes valeurs de gradient. L'utilisation des filtres lors du pré-traitement a permis d'obtenir des régions plus homogènes, rendant l'information de gradient plus fiable.
Il s'agit maintenant de définir les marqueurs à partir de cette image de gradient. D'abord, le gradient est binarisé (5.5).

$$G^b(x,y) = \begin{cases} 0 & \text{si } ||G(x,y)|| < T_G \\ 255 & \text{sinon} \end{cases} \tag{5.5}$$

Le seuil T_G est à fixer de façon empirique. Des expériences menées précédemment [Man05] montrent que le seuil le meilleur pour les vidéos génériques se situe entre 15 et 20[1] (cf. figure 5.3 (d)).

5.2.2.3 Croissance de régions

Chaque région connexe représentant les marqueurs est étiquetée individuellement (cf. figure 5.3 (e)). Pour chacune de ces régions, la couleur moyenne est calculée. Une carte des régions couleur est alors obtenue avec des zones d'incertitude correspondant aux zones de fort gradient, typiquement près des bords des objets. Afin d'assigner ces zones à la région connexe correspondante, un algorithme de croissance de région itératif avec un seuil adaptatif à la région est utilisé [Mah07]. Ainsi, un pixel considéré (lieu d'un fort gradient) est ajouté à la région de luminance moyenne $\bar{m}_Y$ (à laquelle il est adjacent) si la valeur de son coefficient d'ondelette dans la sous-bande LL^K pour la composante Y ($LLY_t^K(x,y)$) vérifie l'inégalité (5.6).

$$||LLY_t^K(x,y) - \bar{m}_Y||_{L_1} \leqslant T_{LPE}(\bar{m}_Y, i) \tag{5.6}$$

Le seuil T_{LPE} est calculé comme étant une fonction du niveau de gris moyen de la région considérée et d'un paramètre Δ^i qui croît avec l'itération i (5.7).

$$T_{LPE}(\bar{m}_Y, i) = F(\bar{m}_Y)\Delta^i \text{ avec } F(\bar{m}) = |\bar{m} - 127| + 128 \text{ et } \Delta^i = \Delta^{i-1} + 0.01 \tag{5.7}$$

La fonction $F(\bar{m})$ est la linéarisation [Mah07] du modèle de Chehdi [Che92]. Elle traduit la sensibilité au contraste du système visuel humain. Le paramètre Δ^i règle la relaxation du seuil, autrement dit, la "vitesse d'immersion" lors de la montée des eaux. La valeur initiale Δ^0 est calculée comme étant $1/F(127)$.
Dans un premier temps, les pixels qui diffèrent de la région à laquelle ils sont adjacent d'une valeur inférieure au seuil sont inclus dans la région. Quand plus aucun pixel ne peut être ajouté à aucune région, le seuil est re-calculé en utilisant le paramètre incrémental Δ^i. La procédure est itérée jusqu'à ce que tous les pixels de gradient fort (cf. section 5.2.2.2) soient affectés à une région (cf. figure 5.3 (f)).

5.2.3 Fusion des informations couleur et mouvement

La dernière étape de cette extraction d'objets à BR est la fusion du masque de mouvement et de la carte de segmentation couleur (cf figure 5.4). Le masque de mouvement renseigne sur la localisation des objets d'intérêt tandis que la segmentation couleur propose des régions homogènes dont les contours correspondent aux contours des objets. Il est donc intéressant de considérer les objets d'intérêt comme étant l'union des régions de la

[1] Dans ce chapitre, nous avons ramené les valeurs des coefficients d'ondelettes de l'intervalle $[-\frac{1}{2}; \frac{1}{2}]$ à l'intervalle $[0; 255]$ de façon à travailler avec des valeurs classiques en Traitement des Images.

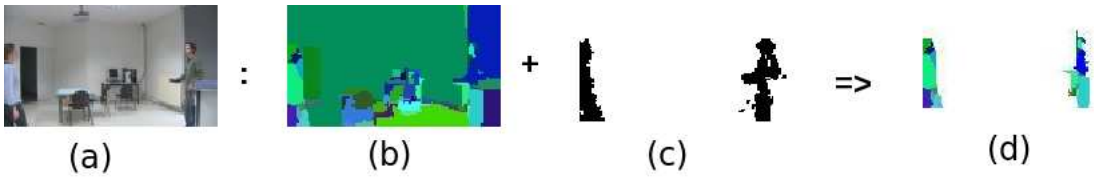


Fig. 5.4 – Illustration du principe de fusion des masques de segmentations couleur et mouvement. (a) trame courante (b) carte des étiquettes de la segmentation couleur (c) masque de mouvement (d) résultat de la fusion

segmentation couleur qui sont inclues dans le masque de mouvement. Comme il est plus que probable qu'une région ne soit pas complètement mais en grande partie incluse dans le masque de mouvement, un seuil de tolérance est proposé (5.8).

$$\frac{\text{nombre de pixels d'une région inclus dans le masque}}{\text{nombre de pixel total de la région}} \geqslant T_{fusion} \tag{5.8}$$

Le seuil T_{fusion} est de l'ordre de 80%, ce qui signifie que si la majorité des pixels de la région est contenue dans le masque, alors toute la région appartient à l'objet en mouvement.

5.3 Extraction multirésolution d'objets par projection spatiale

La segmentation à BR proposée donne déjà une première réponse sur la présence et la localisation de l'objet dans la vidéo. Nous allons maintenant l'adapter aux niveaux de résolution croissante. Pour cela, nous proposons d'utiliser une approche classique de projection/ajustement. Notre approche propose d'utiliser toutes les sous-bandes (LL, LH, HL et HH) dans des zones "d'incertitude" au voisinage des contours de l'objet à la résolution supérieure. La projection se déroule en trois temps. D'abord, le résultat de la segmentation à l'apex de la pyramide est projeté de façon grossière sur le niveau de résolution immédiatement supérieure en utilisant le principe de localisation des ondelettes. Cette projection induit des effets de bloc et conduit à une mauvaise segmentation sur les bords des objets. Pour corriger ce défaut, la deuxième étape consiste en la définition d'une zone d'incertitude dans laquelle les pixels vont être réaffectés suivant un critère fin au niveau courant de la pyramide. Cette étape de projection/affinement (figure 5.1) est appliquée successivement sur tous les niveaux de la pyramide pour obtenir les masques des objets à HR. Les étapes de la projection sont illustrées sur la figure 5.5.

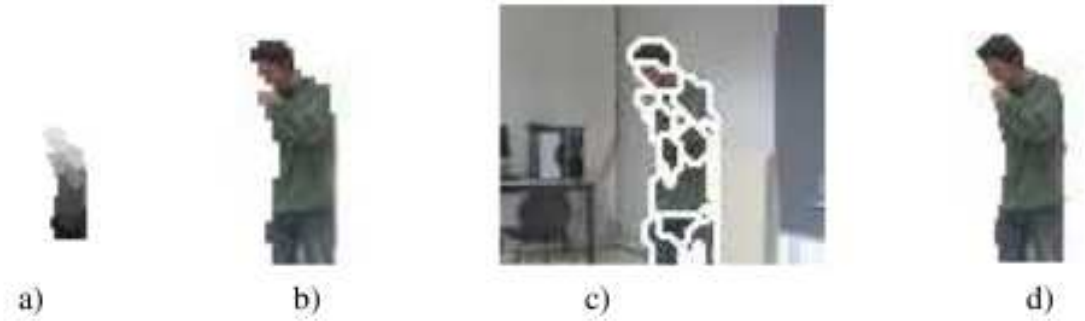


Fig. 5.5 – Etapes de projection/ajustement de l'objet extrait. (a) objet extrait au niveau k (b) projection brute (c) zone d'incertitude (d) résultat après ajustement markovien

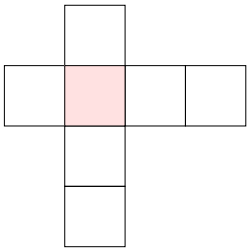


Fig. 5.6 – Elément structurant 4-connexe utilisé pour déterminer la zone d'incertitude

5.3.1 Projection brute et détermination de la zone d'incertitude

D'abord, la projection brute $\hat{S}_t^{k-1}$ sur les niveaux de résolution supérieure est obtenue à partir de S_t^k en utilisant le principe de localisation des ondelettes (cf section 2.4). Ce type de projection conduit à des effets de bloc sur les bords des objets (cf l'épaule gauche de "Vincent" figure 5.5(b)). Ainsi, seuls les pixels du bord des objets ont été mal assignés. Pour calculer l'ensemble des objets O_t^{k-1}, nous définissons une zone d'incertitude dans laquelle les pixels vont être ré-affectés suivant un critère d'ajustement au niveau courant de la pyramide. Nous définissons la zone d'incertitude (figure 5.5 (c)) comme étant la différence entre la dilatation et l'érosion du masque d'objet, obtenu grâce à la projection brute, par un élément structurant 4-connexe dissymétrique (Figure 5.6).

Nous avons déterminé cette zone d'incertitude par observation des conséquences de la projection suivant le principe de localisation des ondelettes sur le cas 1D (cf figure 5.7). La zone d'incertitude est déterminée de la manière présentée dans la 5.7 pour le cas 1D. Les deux premières lignes représentent le masque d'objet (deux étiquettes 1 et 2 correspondant à l'objet et au fond) au niveau k de la pyramide (figure 5.7 (a)) et sa projection brute au niveau $k-1$ (figure 5.7 (b)). Les trois dernières lignes (figure 5.7 (c)) décrivent les trois

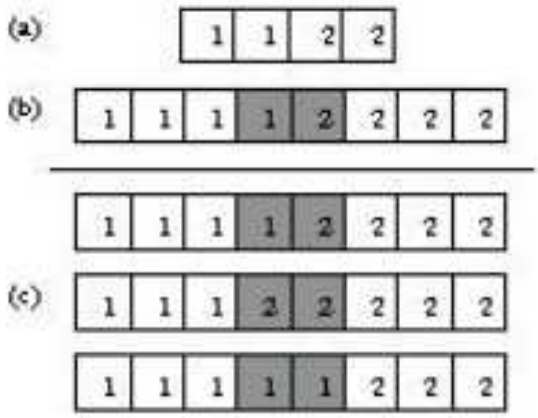


Fig. 5.7 – Illustration du principe de détermination de la Zone d'Incertitude dans le cas 1D

configurations du niveau $k-1$ qui, lors de la construction de la pyramide, auraient pu conduire à la configuration présentée au niveau k. La zone d'incertitude est celle qui est représentée en grisé dans la figure. Ainsi définie elle correspond donc bien à la différence entre la dilatation et l'érosion du masque d'objet projeté de façon "brute" par un élément structurant 4-connexe dissymétrique.

5.3.2 Ajustement

Une fois les deux étapes précédentes effectuées, il reste à ajuster les contours des objets. Cette troisième étape consiste en l'affectation de chaque pixel de la zone d'incertitude à la région dont il est le plus proche en fonction d'un critère d'attribution fondé sur l'homogénéité des régions. Les critères d'attribution possibles sont définis dans les sections suivantes. Nous avons défini deux types de critères. Le premier reprend et modifie l'algorithme de croissance de régions utilisé pour l'extraction à BR. Le second est fondé sur une modélisation de type markovien qui permet de tenir compte des configurations locales. Dans ces deux critères, notre soucis a été d'utiliser l'information de contour contenue dans les sous-bandes de HF de la décomposition en ondelettes.

5.3.2.1 Ajustement morphologique couleur et ondelettes

Cet ajustement reprend le principe de croissance de régions utilisée pour l'extraction des objets à BR (cf section 5.2.2.3). L'inégalité (5.6) régissant la croissance de régions est modifiée pour tenir compte de l'information apportée par les sous bandes de HF (5.9).

$$||LLY_t^K(x,y) - \bar{m}_Y||_{L_1} + \sum_{SB} || \, |SBY_t^K(x,y)| - \bar{m}_{Y,SB}||_{L_1} \leqslant T_{LPE}(\bar{m}_Y, i) \qquad (5.9)$$

où SB est pris successivement dans l'ensemble $\{LH, HL, HH\}$. Seule la valeur absolue des coefficients est significative. Pour tenir compte du degré de texture de la région nous ajoutons la différence entre le coefficient de HF et la valeur moyenne $\bar{m}_{Y,SB}$ des coefficients d'ondelettes décrivant la région dans la sous-bande considérée. Analysons cette inégalité. Si il n'y a pas de contour et que la région est plate, le terme

$$\sum_{SB} || \ |SBY_t^K(x,y)| - \bar{m}_{SB}||_{L_1} \tag{5.10}$$

est proche de 0, on se retrouve dans la situation de la croissance de régions à BR. C'est aussi le cas si la région R est une zone texturée et que le pixel (x,y) appartient à cette texture. Dans ce cas, le terme (5.10) est une caractérisation simplifiée de la distribution de la texture. Si il y a un contour et que la région R est plate, alors (5.10) exprime le contraste du contour. En relaxant le seuil définit comme dans (5.7), tous les pixels de la zone d'incertitude sont assignés progressivement aux régions avoisinantes au niveau $k-1$ de la pyramide. L'utilisant d'un critère sur la HF (5.10) ajoute une barrière dans la croissance de région qui améliore la définition des contours, en particuliers à la frontière entre deux régions ayant des valeurs moyennes proches.

5.3.2.2 Ajustement markovien

Les champs de Markov sont un outil très utilisé dans plusieurs domaines du Traitement d'Images tels que la détection de défauts rectilignes dans des images de chaussées [Del95] et la fermeture de contours [Ura95]. Pour une revue complète de l'utilisation des champs de Markov dans la segmentation d'images, nous invitons le lecteur à lire l'article de Dubes et Jain [Dub89] et le livre de Li [Li95]. Après avoir rappelé le principe de la segmentation d'images par modélisation markovienne, nous définirons notre nouvelle technique d'ajustement.

Principe de la modélisation markovienne en segmentation d'images

Nous rappelons ici le principe de la modélisation markovienne dans le cadre de l'estimation au sens du Maximum a Posteriori (MAP), mais d'autres cadres, tels que l'utilisation d'une fonction de coût, existent.

L'image à segmenter I est considérée comme étant la réalisation y d'une variable aléatoire Y. De même, l'image des étiquettes est la réalisation x d'une variable aléatoire X. L'objectif est de trouver l'image des étiquettes x associée à l'image observée y. Dans le cadre de la théorie bayesienne, le champ des étiquettes estimé $\hat{x}$ est le champ le plus probable en fonction de l'observation y. Le problème se formalise par :

$$\hat{x} = \underset{x}{argmax}\ p(X = x|Y = y) \tag{5.11}$$

Le théorême de Bayes relie la probabilité a posteriori p(X=x|Y=y) à la probabilité a priori p(Y=y|X=x) :

$$p(X = x|Y = y) = \frac{p(Y = y|X = x)p(X = x)}{p(Y = y)} \tag{5.12}$$

Le problème se ré-écrit alors

$$\hat{x} = \underset{x}{argmax}\ \frac{p(Y = y|X = x)p(X = x)}{p(Y = y)} \tag{5.13}$$

$p(Y = y)$ étant indépendante de x, cela revient à

$$\hat{x} = \underset{x}{argmax}\ [p(Y = y|X = x)p(X = x)] \tag{5.14}$$

et en passant au logarithme

$$\hat{x} = \underset{x}{argmax}\ [log(p(Y = y|X = x)) + log(p(X = x))] \tag{5.15}$$

Il faut ensuite modéliser la probabilité d'attache aux données $p(Y = y|X = x)$ et la probabilité a priori $p(X = x)$. Supposons que le champ X est markovien. Il suit alors, d'après le théorême de Hammersley-Clifford, la loi de Gibbs, autrement dit :

- $p(X = x)$ ne dépend que du voisinage immédiat du site du pixel
- $p(X = x) = \frac{1}{Z}e^{-U_c(x)}$

où Z est une constante de normalisation et $U_c(x)$ est une somme de potentiels :

$$U_c(x) = \sum_{c \in \mathcal{C}} V_c(x) \tag{5.16}$$

c est une clique de l'espace de cliques C choisi. La formulation la plus courante d'un tel potentiel est

$$V_c(x) = (1 - \delta(x, x_c))\alpha_c \tag{5.17}$$

avec x_c l'étiquette du voisin par rapport à la clique et α_c le facteur de potentiel associé.
La loi de probabilité a priori $p(Y = y|X = x)$ traduit les hypothèses que l'on fait sur la distribution des couleurs dans une région donnée. Conventionnellement, on fait l'hypothèse d'une distribution gaussienne autour de la couleur moyenne de la région :

$$p(Y = y|X = x) = \frac{1}{\sqrt{2\pi|\mathcal{E}_x|^2}} e^{(y-\nu_x)^T \mathcal{E}_x^{-1}(y-\nu_x)} \tag{5.18}$$

En utilisant dans (5.15) les expressions précédentes, le problème se simplifie en une minimisation de somme de potentiels. Le label $l(s)$ d'un site s est défini par :

$$l(s) = \underset{l \in [1,L]}{argmin} \left[\sum_i U_i(l, s) \right] \tag{5.19}$$

où i permet de désigner les différents potentiels.
Différentes méthodes existent pour résoudre ce problème de minimisation de fonction. Ces méthodes se divisent généralement en deux groupes : les méthodes stochastiques (recuit simulé [Gem84]) et les méthodes déterministes (ICM, "Iterated Conditional Modes", [Bes86], HCF, "Highest Confidence First", [Cho90]). Même si seules les méthodes stochastiques permettent de trouver le minimum global de la fonction étudiée, les méthodes déterministes sont le plus souvent utilisées. Elle permettent d'aboutir à une solution convenable, voire meilleure en qualité visuelle, tout en étant moins coûteuses en temps de calcul [Bla89]. Il n'est d'ailleurs pas surprenant de trouver plusieurs solutions puisque le problème de segmentation d'images que nous cherchons à résoudre est un problème mal-posé [Ber88].

Modélisation markovienne pour l'ajustement des pixels de la zone d'incertitude
Dans notre travail, nous avons choisi de considérer deux fonctions de potentiels : U_1 lié aux valeurs de couleur et U_2 lié au voisinage. Seuls les pixels appartenant à la région d'incertitude peuvent être ré-étiquettés. Le label d'un site s est donné par :

$$l(s) = \underset{l \in [1,L]}{argmin}[U_1(l,s) + U_2(l,s)] \tag{5.20}$$

Les potentiels sont définis dans la suite du paragraphe. Le potentiel U_1 est lié à la couleur et est issu de l'attache aux données :

$$U_1(l,s) = (\gamma_{LL} - \nu_l)^T \mathcal{E}_l^{-1} (\gamma_{LL} - \nu_l) \tag{5.21}$$

avec γ_{LL} le vecteur couleur des coefficients d'ondelettes pris dans la sous-bande LL, ν_l et $\mathcal{E}_l$ sont respectivement le vecteur couleur moyen et la matrice de covariance des vecteurs couleur de la région d'étiquette l.
Pour utiliser pleinement l'information de HF des sous-bandes LH, HL et HH, nous définissons un potentiel de régularisation modifié U_2 calculé sur les cliques. Le potentiel de cliques traditionnellement utilisé est exprimé par :

$$U_2^{class}(l,s) = \sum_{c \in C_s} [A(1 - \delta(l, l_s))] \tag{5.22}$$

Ici, C_s désigne l'ensemble des cliques de taille 2 (en travaillant en 8-connexité) et c désigne le pixel voisin de s dans une clique. La constante A est fixée expérimentalement, δ est le symbole de Kronecker. Ce potentiel privilégie (resp. pénalise) les configurations de deux pixels ayant deux labels identiques (resp. différents) dans une clique selon le résultat de la segmentation. Il traduit un a priori de compacité sur les régions homogènes en couleur. Or certaines régions peuvent avoir des formes plus complexes avec des enclaves. Le potentiel classique (5.22) ne permet pas d'en rendre compte et ne décrit pas correctement ce type de régions. Nous savons que dans les sous-bandes de HF, une information sur la position des

contours est disponible. Ainsi, si dans une clique horizontale (resp. verticale, diagonale), le coefficient HL (resp. LH, HH) est fort, un contour est présent. Aussi, dans ce cas favoriser une configuration de deux labels identiques paraît aberrant. Nous introduisons un premier terme correctif au potentiel (5.22) de la forme $\delta(l, l_c)|HF|_c^n$ aboutissant à U_2^{temp} (5.23).

$$U_2^{temp}(l, s) = \sum_{c \in C_s} (A(1 - \delta(l, l_s)) + \delta(l, l_c)|HF|_c^n) \tag{5.23}$$

Ici $|HF|_c^n$ désigne la valeur normalisée (cf équation (5.26)) du coefficient HF associé à la clique. Cette valeur $|HF|_c^n$ dépend du type de clique ; une clique horizontale (respectivement verticale, diagonale) utilise le coefficient d'ondelettes HL (resp. LH, HH). Si un contour est présent et bien marqué, le potentiel associé est fort dans le cas de deux étiquettes voisines identiques.
De façon analogue, nous introduisons un deuxième terme correctif $-(1 - \delta(l, l_c))|HF|_c^n$. L'objectif de ce terme est d'abaisser le potentiel classique dans le cas où les labels différents coïncident avec la présence d'un contour. Dans ce cas, il est naturel de trouver qu'il vaut mieux avoir deux labels différents. Le potentiel U_2 ainsi formé après correction s'exprime par :

$$U_2(l, s) = \sum_{c \in C_s} [A(1 - \delta(l, l_c) + (\delta(l, l_c)|HF|_c^n) - (1 - \delta(l, l_c))|HF|_c^n)) \tag{5.24}$$

ce potentiel s'écrit de façon simplifiée sous la forme :

$$U_2(l, p) = \sum_{c \in C_p} (A(1 - \delta(l, l_c) + (2\delta(l, l_c) - 1)|HF|_c^n)) \tag{5.25}$$

Le terme de correction, $A((2\delta(l, l_c) - 1)|HF|_c^n)$ nous permet de corriger les défauts de segmentation qui apparaissent quand la clique contient des labels identiques (respectivement différents) et un fort (resp. faible) coefficient HF.

Le coefficient $|HF|_c^n$ est obtenu par normalisation sur chaque sous-bande HL, LH et HH du maximum en valeur absolue de la sous-bande.

$$|HF|_c^n = \frac{|HFY|}{\underset{HFY \in HF}{argmax} |HFY|} \text{ avec } HF = LH, HL, HH \tag{5.26}$$

Dans notre travail, nous n'avons pas utilisé une des méthodes classiques de minimisation de fonction citée dans le paragraphe précédent. Nous avons considéré la minimisation de l'énergie comme une méthode déterministe de croissance de régions. Cela se justifie par le fait que la zone d'incertitude est très mince et donc l'optimisation stochastique serait plus coûteuse en temps de calcul pour une différence insuffisante dans les résultats.

5.3.2.3 Comparaison des ajustements

Deux ajustements ont été proposés. Nous allons maintenant les comparer sur un même exemple (figure 5.8) et montrer l'influence des coefficients de HF par rapport à un ajustement classique.

Ajustement Morphologique. Les figures 5.8 (c) et (d) représentent le résultat à pleine résolution de l'ajustement morphologique BF (c) et HF (d). Nous appelons ajustement morphologique HF l'ajustement présenté dans la paragraphe 5.3.2.1. Par analogie, nous définissons l'ajustement morphologique BF de la même manière, le critère de fusion (5.6) ne faisant plus intervenir les coefficients HF comme dans (5.9) ; on se retrouve dans le cas de la croissance de régions utilisée à BR (section 5.2.2.3). L'extraction d'objet à BR (cf. figure 5.8 (b)) présente un léger défaut de segmentation au niveau de l'objet que tient le personnage (signalé par un rond rouge) et un petit morceau de fond est fusionné avec l'objet. Avec l'approche morphologique BF ce défaut est amplifié le long de la pyramide lors des étapes de projection/ajustement (figure 5.8 (c)). L'utilisation des coefficients de HF permet de limiter cette expansion (figure 5.8 (d)). Nos expérimentations ont montré que l'utilisation des coefficients de HF est plus efficace aux niveaux intermédiaires, alors qu'une simple croissance de régions morphologique BF est appliquée au niveau basse résolution de la pyramide. A pleine résolution, il faut aussi utiliser l'approche BF car les coefficients des sous-bandes LH, HL et HH ne sont pas disponibles.

Ajustement markovien. Les figures 5.8 (e) et (f) représentent le résultat à pleine résolution de l'ajustement markovien BF (e) et HF (f). Nous appelons ajustement markovien HF l'ajustement présenté dans la paragraphe 5.3.2.2. Par analogie, nous définissons l'ajustement markovien BF de la même manière, en considérant $|HF|_c^n$ nul dans la définition du potentiel de clique (5.25). L'utilisation du formalisme markovien permet d'empêcher la croissance du fond par rapport à l'ajustement morphologique. Pour se rendre compte de l'influence des coefficients de HF nous avons agrandi le détail des mains (carré bleu sur les figures 5.8 (e) et (f)). L'utilisation du potentiel classique BF (e) tend à lisser les contours et favorise l'inclusion du fond autour du pouce. L'utilisation des coefficients HF permet de limiter cette tendance et le pouce est mieux distingué. Si ce genre de correction peu paraître anondin à BR, il prend tout son sens à HR où les détails de cet ordre sont visibles. Au vu de ces premiers résultats, nous avons décidé de poursuivre les expériences avec l'ajustement markovien uniquement.

5.4 Résultats

Le but de cette section est d'évaluer et de présenter quelques résultats de notre méthode d'extraction d'objets. Les figures 5.9 et 5.10 représentent les résultats de l'extraction d'objets multirésolution sur les séquences "man in restaurant" et "street with trees and bicycle".

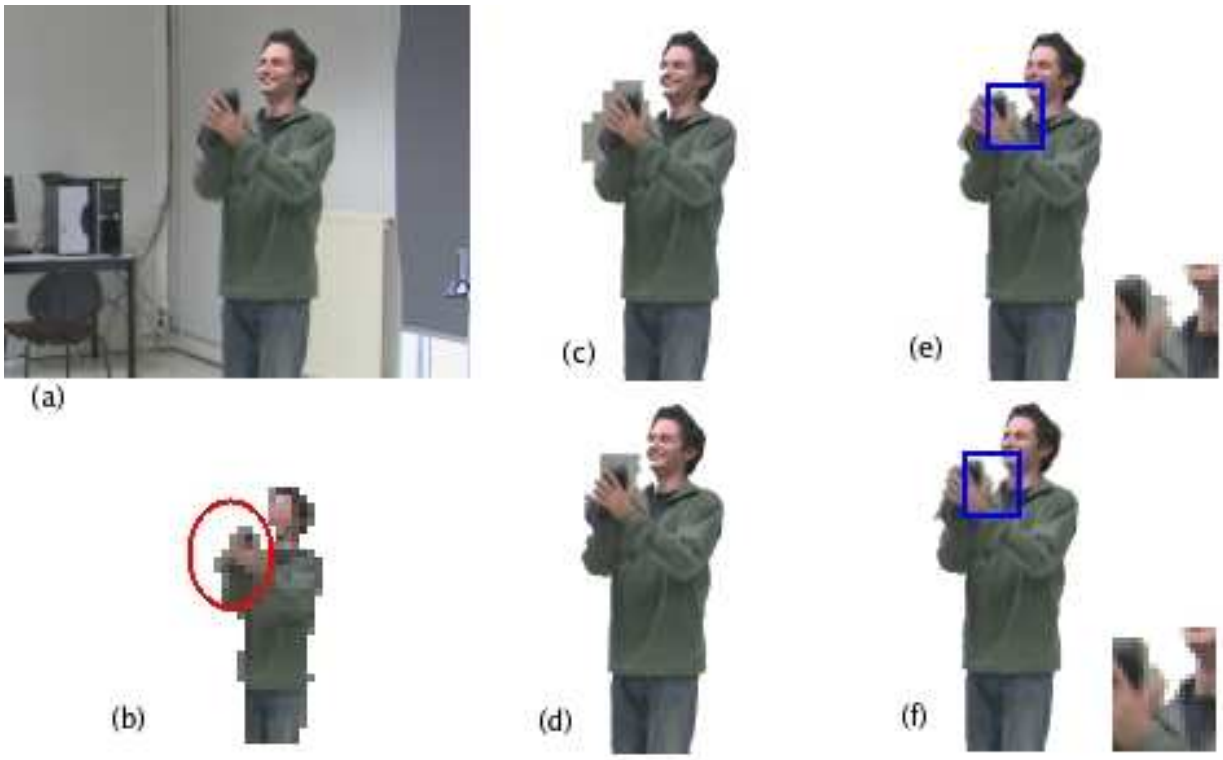


Fig. 5.8 – (a) Image originale, “Vincent”, LaBRI (b) Résultat de la segmentation à BR (grossi x3) (c) Segmentation morphologique LL (d) Segmentation morphologique HF (e) Régularisation markovienne classique (f) Régularisation markovienne HF

Fig. 5.9 – Exemple de résultat d'extraction d'objets obtenu avec notre méthode utilisant l'ajustement markovien, video "man in restaurant"

Les silhouettes des objets extraits sont représentées en foncé par rapport au fond. D'autres exemples de résultats sont présentés dans la figure 5.11 où les masques d'objets sont représentés en rouge.

Les méthodes d'évaluation de la segmentation de la littérature ont été présentées dans la section 3.2.1. L'évaluation de la qualité d'extraction d'objets dépend fortement de l'usage que l'on souhaite faire du résultat. Dans notre cas, il s'agit de bâtir un descripteur à partir des données extraites. Nous pensons que pour ce genre d'application, il n'y a pas besoin de récupérer avec précision les objets. Un objet est bien extrait dans une trame si :

Fig. 5.10 – Exemple de résultat d'extraction d'objets obtenu avec notre méthode utilisant l'ajustement markovien, video "street with trees and bicycle"

Fig. 5.11 – Exemples de résultats d'extraction d'objets obtenus avec notre méthode utilisant l'ajustement markovien

- Une part significative de l'objet, permettant sa reconnaissance par un être humain, est extraite ;
- Aucun élément du fond n'est considéré comme faisant partie de l'objet. Cette contrainte étant très forte, nous tolérons une petite proportion, par rapport à la taille de l'objet effectivement extrait, de fond.

C'est pourquoi nous proposons de ne pas utiliser les techniques classiques quantitatives de comparaison des objets avec un masque complet manuel. Notre méthode d'évaluation est visuelle. Pour chaque objet extrait, un opérateur humain détermine parmi trois classes possibles la catégorie à laquelle appartient l'image :

1. Pas d'objet extrait, ou extraction trop faible pour être reconnaissable (cf figure 5.12 (c))
2. Grande proportion de fond considéré comme objet (cf figure 5.12 (b))
3. Objet extrait reconnaissable sans adjonction de fond (cf figure 5.12 (a))

De même, cet opérateur humain crée une vérité terrain en indiquant pour chaque trame d'une vidéo si un objet dans une vidéo est présent ou non. Nous proposons de mesurer le **rappel global** comme étant le rapport du nombre total de détections correctes (catégorie 3) sur le nombre d'objets qui étaient à détecter. Sur la base de données présentée dans la

(a) Exemple de bonne détection

(b) Exemple de Mauvaise détection avec adjonction de fond

(c) Exemple de non détection

Fig. 5.12 – Exemple de décisions prises par un opérateur humain (a) bonne détection (b) mauvaise détection avec adjonction de fond (c) pas de détection

section 7.2, ce rappel est de $R_g = 0.31$. De même, nous mesurons la **précision globale** comme étant le nombre total de détections correctes (catégorie 3) sur le nombre total de détection (catégories 2 et 3). Sur la même base de données, on mesure $P = 0.67$. Le rappel global est faible du fait que la mesure suppose que l'objet est détecté dans chaque trame d'une même vidéo. Cependant, dans notre application d'indexation, nous verrons qu'il nous suffit de bien détecter l'objet au moins une fois dans une trame. Nous proposons de calculer un **rappel partiel** comme étant le nombre de vidéos pour lesquelles l'objet a été détecté correctement au moins une fois sur le nombre de vidéos où un objet devait être détecté. Ce rappel est de $R_p = 0.88$. Les vidéos où l'objet n'est pas récupéré sont typiquement "tractor" et "sunflower" où le fond très texturé ne permet pas d'obtenirs de bons résultats avec la mise en correspondance de blocs. De plus, pour la vidéo "sunflower", l'abeille représente en réalité moins de 6% de l'image, ce qui est très petit pour permettre une détection correcte à BR.
L'évaluation de l'impact de la qualité de la segmentation par rapport à l'indexation sera faite dans le chapitre 7, une fois les tâches à effectuer et le descripteur utilisé présentés.

5.5 Conclusion

Dans ce chapitre, nous avons proposé une méthode d'extraction d'objets en mouvement propre sur des paires de trames dans le domaine des ondelettes obtenu par compression JPEG2000. La méthode proposée est scalable dans le sens où elle suit la progression de transmission du flux JPEG2000. Autrement dit, elle extrait les objets en mouvement en utilisant l'information de Basse Résolution en premier lieu, puis en se servant des informations de moyenne et Haute Résolution au fur et à mesure de leur transmission. La représentation de l'objet obtenue est multirésolution (cf. figure 5.13) et peut être inclue dans un flux scalable.
Le principe de projection/ajustement des masques d'objets extraits est une technique classique du Traitement des Images et des vidéos. L'originalité de notre approche est de proposer de se servir explicitement des valeurs des coefficients de HF dans les méthodes de croissance de régions servant à l'ajustement. Ces coefficients servent de barrière à l'extension des régions dans le cas de la présence d'un contour. L'utilisation d'une modélisation markovienne permet de tenir compte de la direction privilégiée du contour indiquée par les coefficients d'ondelettes de HF. Une telle approche nous permet d'obtenir des contours fins des objets à pleine résolution à partir d'une extraction grossière effectuée à BR.

Fig. 5.13 – Résultat de l'extraction d'objets multirésolution

Troisième partie

Indexation scalable des vidéos HD par les objets

Chapitre 6

Indexation par histogrammes d'ondelettes

La méthode présentée dans la partie précédente fournit une représentation multirésolution des objets en mouvement de la vidéo. Dans ce chapitre, il s'agit de définir un indice scalable à partir de cette représentation à des fins d'indexation et de recherche dans des bases de données. Nous avons choisi d'utiliser les histogrammes d'ondelettes sur les coefficients décrivant l'objet. Après avoir rappelé la définition de l'histogramme et les travaux de la littérature l'utilisant, nous présentons notre descripteur d'histogrammes d'ondelettes en multirésolution. Les caractéristiques du descripteur sur la base de données sont ensuite présentées. Une évaluation plus complète de ce descripteur sera faite dans le chapitre suivant.

6.1 Présentation des histogrammes

Les histogrammes couleurs sont beaucoup utilisés dans la littérature (cf section 3.1.1) pour la recherche basée contenu des images et vidéos. Ils sont en effet peu couteux en terme de temps de calcul et généralement insensibles à de petits changements de la position de la caméra. Le principal défaut d'un histogramme couleur est qu'il fournit seulement une caractérisation grossière de l'image : deux images très différentes d'apparence peuvent avoir le même histogramme.

6.1.1 Définition : histogramme couleur

L'idée sous-jacente est de caractériser les images par la Fonction de Densité de Probabilité (PDF) des couleurs dont l'histogramme est une approximation. L'histogramme se définit

comme suit.
Soit une image (ou un objet) I quantifiée dans un espace réduit à M classes couleurs $(c_1, c_2, ..., c_M)$. L'**histogramme couleur** H est un vecteur à M composantes $(H_{c_1}, H_{c_2}, ..., H_{c_M})$ pour lequel H_{c_m} représente le nombre de pixels de couleur c_m dans l'image I. On a alors $\sum_{m=1}^{M} H_{c_m} = N$ où N est le nombre de pixels dans l'image (ou l'objet).
Précisons quelques notions liées à l'histogramme :

- Les couleurs c_m sont des vecteurs de couleurs (c_a, c_b, c_c) pris dans un espace couleur quelconque (par exemple RGB, YUV, HSV...).

- On appelle **dimension de l'histogramme** la dimension des vecteurs couleurs, c'est-à-dire qu'ici elle vaut 3. On appelle **classes de l'histogramme** les coefficients H_{c_m}.

- Un histogramme normalisé $h = (h_{c_1}, h_{c_2}, ..., h_{c_M})$ est un histogramme à valeurs comprises dans $[0, 1]$. Il s'obtient à partir de l'histogramme non normalisé par $h_{c_m} = \frac{H_{c_m}}{N}$. On remarque que par convention, l'histogramme non normalisé est noté H et l'histogramme normalisé h.

Les principales difficultés de la construction d'histogramme sont le choix de l'espace couleur et la quantification adoptée pour cet espace. Le choix de l'espace couleur est dans notre cas fixé par le processus de compression par JPEG2000. C'est pourquoi nous ne présentons pas dans la suite un état de l'art sur cette problématique. Notons toutefois que la plupart des auteurs choisissent de façon emprique l'espace qui leur semble le plus approprié. Le principe est que plus les composantes de l'espace couleur sont décorrélées, plus l'information apportée par l'histogramme est pertinente.
Dans la suite de cette section, nous donnons un bref aperçu de l'état de l'art en matière de choix du pas de quantification. Puis nous nous intéressons aux différentes métriques de similarité entre histogrammes existantes.

6.1.2 Choix de la taille des classes de l'histogramme

Le problème du choix de la taille des classes, autrement dit du pas de quantification de l'espace couleur, est un problème largement étudié dans le domaine de la statistique et non entièrement résolu. La quantification choisie peut conduire à une représentation trop lissée, correcte ou pas assez lissée de la PDF correspondante. La quantification de l'espace couleur a fait l'objet de plusieurs études dans les contextes du codage des images et de la vidéo couleur par quantification vectorielle, de l'évaluation de la qualité et de l'indexation ([MPE01, Lar04]). Deux cas de quantification peuvent être cités :

- *Quantification adaptative* : le pas de quantification s'adapte à la distribution observée de l'échantillon, les classes sont de tailles variées quel que soit le cadre (nombre de

classes fixé a priori ou non). Les algorithmes classiques comme SPLIT-LBG (utilisés en MPEG7), K-moyennes ou encore le plus récent K-moyenne++ peuvent être cités ici. Cette quantification, dans le cadre du codage par exemple, nécessite de faire de l'apprentissage sur un sous-ensemble de données.

– *Quantification uniforme* : toutes les classes sont de la même taille et de la même forme dans l'espace de représentation. Cette quantification ne nécessite pas d'apprentissage préalable et permet, comme nous le verrons par la suite, une comparaison simple des histogrammes.

Dans le cadre de cette étude, nous considérons ce deuxième cas pour des raisons de généricité de la description. Néanmoins se pose la question du choix de la taille et de la forme des cellules. Dans le cas de variables aléatoires scalaires, cette question a fait l'objet de nombreuses études statistiques et des règles de choix ont été développées.

– *Règle de Sturges* [Stu26] : donne le nombre de classes $b = 1+log_2 N$, avec N la taille de l'échantillon statistique (la taille de l'image ou de l'objet dans notre cas). On accède à la taille des classes par $\hat{l} = \frac{\Delta}{b}$ avec Δ l'intervalle des données. L'histogramme résultant est généralement trop lissé pour des échantillons de grande taille, mais cette règle est adéquate pour les petits ensembles. Elle est plus appropriée au cas d'une distribution unimodale.

– *Règle de Scott*[Sco79] : $\hat{l} = 3.49\hat{\sigma}N^{-1/3}$ où $\hat{\sigma}$ est une estimée de l'écart-type. C'est une version améliorée de la règle de Sturges qui est plus adaptée aux échantillons statistiques de grande dimension. Elle est elle aussi plus appropriée au cas d'une distribution unimodale.

– *règle de Shimazaki* [Shi07] : cette règle a été développée sous l'hypothèse d'une distribution de poisson. Elle consiste en la minimisation d'une fonction de coût.

6.1.3 Mesures de similarité entre histogrammes

Il existe plusieurs façons de comparer les histogrammes suivant qu'ils sont considérés comme des vecteurs ou des distributions. Deux types de mesure vont être décrits dans la suite : les mesures de distance et les mesures de similarité. Elles fonctionnent à l'inverse l'une de l'autre. Ainsi, pour une mesure de distance (resp. de similarité), plus la valeur mesurée est grande plus les histogrammes sont différents (resp. semblables).

Cas des vecteurs

Dans ce cas, les histogrammes sont supposés être normalisés. En considérant les histogrammes comme des vecteurs, il est possible de mesurer la différence entre histogrammes à l'aide des **métriques classiques de Minkowski** L_p définies par :

$$d_{L_p}(h,g) = (\sum_{m=1}^{M}(h_{c_m} - g_{c_m})^p)^{\frac{1}{p}} \tag{6.1}$$

Les plus utilisées sont les normes L_1 (norme de Manhattan) et L_2 (norme euclidienne). La mesure d'**intersection des histogrammes de Swain et Ballard** [Swa91] est une des plus utilisée. Elle permet de mesurer le recouvrement entre deux histogrammes :

$$d_{CH}(h,g) = \sum_{m=1}^{M} min(h_{c_m}, g_{c_m}) \tag{6.2}$$

Cette mesure est efficace dans un très grand nombre de cas. Elle possède entre autres propriétés que :

- La distinction d'une grande variété d'objets est possible.
- La mesure est peu sensible aux rotations et à des changements modérés de la distance par rapport à la caméra.
- L'identification peut se faire même si l'objet est partiellement occulté.

Cette mesure est une mesure de similarité à valeurs comprises entre 0 et 1. Enfin, le **coefficient de Bhattacharyya** [Bha43] est défini par :

$$d_B(h,g) = \sum_{m=1}^{M} \sqrt{h_{c_m} g_{c_m}} \tag{6.3}$$

Il s'agit là aussi d'une mesure de similarité comprise entre 0 et 1.

Cas des distributions

Nous l'avons dit, l'histogramme est une approximation de la PDF couleur. A ce titre, il peut être considéré comme une distribution et les mesures de divergence entre les lois de probabilité ainsi que les autres tests d'hypothèses statistiques sont applicables. La loi théorique de l'histogramme est inconnue mais peut être estimée. Le principe du test statistique est de savoir s'il faut accepter ou rejeter une hypothèse H_0 sur la loi de probabilité associée à une distribution. Dans le cas de la comparaison d'histogrammes, l'hypothèse H_0 est que les deux histogrammes représentent la même loi de probabilité. Notons que de tels tests travaillent sur les effectifs et non les fréquences, c'est-à-dire que les histogrammes ne doivent pas être normalisés. Le **test du** χ^2 (aussi appelé **test de Pearson**) est à l'origine défini pour comparer l'adéquation d'une loi de probabilité empirique (l'histogramme) à une loi de probabilité théorique. Elle a été étendue à la comparaison de deux lois empiriques. La distance du χ^2 est donnée par :

$$d_{\chi^2}(H,G) = \sum_{m=1}^{M} \frac{(H_{c_m} - G_{c_m})^2}{G_{c_m}} \tag{6.4}$$

Cette distance est en fait une variable aléatoire qui, d'après le théorême de Pearson, suit une loi χ^2_r à r degrés de liberté lorsque N tend vers l'infini. Le nombre de degrés de liberté se calcule par $r = M - 1$[1]. L'hypothèse H_0 est acceptée, avec une certaine valeur de confiance, si d_{χ^2} est inférieure à un certain seuil, déterminé par la fonction de répartition de la loi du χ^2. On notera que cette mesure n'est pas une "vraie" mesure de distance car elle n'est pas symétrique. Pour le **test de Kolmogoroff-Smirnov**, se sont les fonctions de répartition empiriques qui sont comparées. Dans le cas des histogrammes, elles correspondent aux histogrammes cumulés. L'histogramme cumulé $\tilde{h}$ de h est le vecteur à M dimensions $(\tilde{h}_{c_1}, \tilde{h}_{c_2}, .., \tilde{h}_{c_M})$ avec $\tilde{h}_{c_m} = \sum_{i=1}^{m} h_{c_i}$. La distance entre deux fonctions de répartition empiriques est donnée par :

$$d_{KS}(h, g) = \underset{m}{max} |\tilde{h}_{c_m} - \tilde{g}_{c_m}| \tag{6.5}$$

Là encore, si d_{KS} excède un certain seuil (le niveau de coupure) alors les distributions sont différentes. Enfin, évoquons la **divergence de Kullbach-Leibler** définie, dans sa version discrète, par :

$$d_{KL}(H, G) = \sum_{m=1}^{M} H_{c_m} log \frac{H_{c_m}}{G_{c_m}} \tag{6.6}$$

qui est elle aussi à comparer à un seuil pour prendre une décision sur H_0. Il a été montré que cette mesure pouvait être calculée à l'aide des algorithme de k plus proches voisins, ce type d'approche ayant été utilisé avec succès par Boltz et al [Bol09]. Cependant, dans les objectifs de recherche des vidéos les plus similaires, les décisions de rejet ou d'acceptation de H_0 ne sont pas prises et les métriques précédemment définies sont utilisées comme les mesures de similarité des vidéos.

La complexité du processus de mise en correspondance peut être réduite par la quantification plus fine de l'espace couleur [Swa91], l'utilisation de l'histogramme des caractéristiques dominantes [Str95], l'utilisation d'histogrammes de plus faible dimension en représentant l'histogramme couleur à différentes résolutions [vel95] (c'est-à-dire que l'histogramme est requantifié) et l'utilisation d'une métrique de complexité moindre [Haf95]. Les performances de recherche peuvent être améliorées en tenant compte de la localisation des couleurs dans la représentation couleur de l'image [Str95]. Cependant cette technique nécessite une segmentation efficace et une représentation des sous-images.

[1] Dans le cas de la comparaison à une distribution théorique, $r = M - 1 - l$ où l est le nombre de paramètres qui ont été estimés pour définir la loi théorique

6.2 Descripteur scalable proposé : histogrammes d'ondelettes en multirésolution

Pour tenir compte de la diversité des formes des objets articulés le long de la séquence vidéo et des fortes auto-occultations des objets, nous proposons d'utiliser un descripteur fondé sur les histogrammes dans le domaine de la TOD. En effet, nous avons vu dans les nombreux travaux de la littérature que les histogrammes, bien que n'étant pas parfaitement discriminatifs, sont un outil efficace d'indexation, en particulier robuste à de nombreuses modifications de l'image. Nous allons considérer conjointement l'information de couleur (contenue dans les sous-bandes de BF) et de texture (contenue dans les sous-bandes de HF). Après avoir rappelé les différentes modélisations statistiques de la distribution des coefficients d'ondelettes de l'état de l'art, nous présentons notre descripteur. Comme ce descripteur repose sur des histogrammes, nous préciserons la stratégie de quantification adoptée et les mesures de similarité proposées.

6.2.1 Etat de l'art : histogrammes des ondelettes

L'utilisation de descripteurs sous la forme d'histogrammes d'ondelettes a été abordée dans des travaux précédents. Ainsi, Wang et al [Wan97] proposent un histogramme de vecteurs binaires construit à partir de la représentation d'une image en ondelettes. Dans leurs travaux, l'image est décomposée sur K niveaux d'ondelettes. Seules les valeurs absolues des coefficients des sous-bandes de HF sont utilisées par la suite. Les sous-bandes de HF sont sur-échantillonnées à la taille de l'image et constituent les canaux de texture ; un "point de texture" est défini, caractérisé par un vecteur de dimension $3 * K$ en utilisant le principe de localisation des ondelettes. Chaque élément du vecteur est binarisé en valeur élevée (d'étiquette 1) et valeur faible (d'étiquette 0). L'histogramme d'ondelettes est alors l'histogramme de ces vecteurs binarisés. Mandal et al. [Man99a] reprennent cette méthode et proposent de construire ces histogrammes de texture à une résolution qui peut être différente de la pleine résolution. De plus, ils incluent les coefficients de BF préalablement centrés en 0. Les histogrammes obtenus présentent des pics à intervalles réguliers qui sont caractéristiques de la texture. De telles approches ne peuvent être adaptées à un contexte scalable car toutes les résolutions de la pyramide d'ondelettes doivent être utilisées. De plus la binarisation des coefficients de HF ne caractérise que de façon grossière la distribution des coefficients en ne tenant pas compte des nuances entre texture (coefficient de HF de valeur moyenne) et contour (coefficient de HF fort).
L'histogramme étant une mesure de la distribution des coefficients, l'étude des travaux de modélisation de la distribution des coefficients d'ondelettes s'avère nécessaire. Rappelons que la sous-bande de BF est assimilable à une image et que sa distribution est donc classiquement modélisable par une ou plusieurs gaussiennes. Les coefficients de HF ont une

distribution plus atypique. Ainsi, par étude empirique, Buccigrossi et Simoncelli [Buc99] ont montré que les distributions des coefficients des sous-bandes de HF suivaient une loi laplacienne plutôt qu'une loi gaussienne. La différence est que le "pic" correspondant au maximum est plus resserré dans le cas d'une distribution de Laplace. Pour tenir compte de ce pic, Yuan et Zhang [Yua04] modélisent la distribution des coefficients dans chaque sous-bande individuellement par la combinaison de deux gaussiennes, la première gaussienne correspondant aux coefficients de faible amplitude. Sa variance est choisie relativement faible pour permettre de capturer le "pic" de valeurs autour de zéro. La deuxième gaussienne permet de caractériser les coefficients de forte amplitude. La valeur de variance est plus élevée que pour la précédente afin de capter la traîne de la distribution.

6.2.2 Définition

Nous proposons de définir notre descripteur à partir des histogrammes normalisés des coefficients d'ondelettes décrivant l'objet extrait. Pour une trame de la vidéo W_t, le descripteur scalable $H_{t,i}$ de l'objet extrait $O_{t,i}$ est défini par :

$$H_{t,i} = \{(h^k_{LL,t,i}(O^k_t, \gamma_{LL}), h^k_{HF,t,i}(O^k_t, \gamma_{HF})), k \in [0, K]\} \tag{6.7}$$

Nous avons modifié la notation des histogrammes par rapport au paragraphe de l'état de l'art afin de rendre compte du contexte. Ainsi, l'histogramme normalisé est toujours noté h. L'exposant k indique le niveau de résolution de la décomposition en ondelettes où l'histogramme est calculé, l'indice t désigne l'instant de temps de la trame d'ondelettes considérée. Nous précisons pour chaque histogramme le domaine spatial sur lequel il est calculé à savoir l'objet O^k_i et le type de couleur utilisé γ_{LL} ou γ_{LH}. Nous précisons la signification de ces couleurs dans la suite. Nous noterons dans la suite $H^k_{t,i} = \{(h^k_{LL,t,i}(O^k_t, \gamma_{LL}), h^k_{HF,t,i}(O^k_t, \gamma_{HF}))\}$ le descripteur $H_{t,i}$ considéré au niveau k.
Pour chaque objet, le couple d'histogrammes $H^k_{t,i}$ est calculé à chaque niveau de résolution de la pyramide d'ondelettes. Considérons une quantification uniforme sur chaque axe de l'espace couleur considéré (YCrCb dans notre cas, cf section2.3). Supposons également que le pas de quantification est propre à chaque axe. L'histogramme joint h représente le nombre des vecteurs couleurs (6.8) ou (6.9). Pour la sous-bande LL, la couleur associée est le simple triplet de valeur des coefficients d'ondelettes pris dans les composantes Y, U et V de la sous-bande LL :

$$\gamma_{LL} = (Y_{LL}, U_{LL}, V_{LL}) \tag{6.8}$$

Pour les sous-bandes HF, la couleur est calculée comme étant la moyenne des valeurs des coefficients des sous-bandes HF prises sur les trois composantes soit :

$$\gamma_{HF} = \begin{pmatrix} \frac{1}{3}(|Y_{LH}| + |Y_{HL}| + |Y_{HH}|) \\ \frac{1}{3}(|U_{LH}| + |U_{HL}| + |U_{HH}|) \\ \frac{1}{3}(|V_{LH}| + |V_{HL}| + |V_{HH}|) \end{pmatrix}^T \tag{6.9}$$

Seule la valeur absolue des coefficients de HF est significative. Nous conservons la distinction entre sous-bande de BF et sous-bandes de HF, toujours motivée par le fait qu'elles sont de signification physique différentes (LL représente la couleur et HF les contours). Nous avons décidé de regrouper les trois sous-bandes de HF en un seul histogramme afin d'être plus robuste aux rotations. En effet, considérons un objet simple constitué de traits verticaux uniquement. Dans la représentation en ondelettes, cette information va se retrouver dans la sous-bande HL. Si maintenant cet objet subit une rotation dans le même plan de 90^o, les traits vont apparaître horizontaux et l'information sera portée par la sous-bande LH. En moyennant les sous-bandes, l'information se retrouve au même endroit et peut donc être comparée efficacement.

6.2.3 Choix du pas de quantification

Le pas de quantification de l'espace couleur choisi influence l'allure de l'histogramme. Considérons une quantification uniforme sur chaque axe de l'espace couleur considéré (YCrCb dans notre cas). Supposons également que le pas de quantification est propre à chaque axe.

Suivant que ce pas est adapté ou non à chaque axe, la distribution des couleurs estimée par l'histogramme sera ou non pertinente pour caractériser l'objet. Il s'agit donc de trouver pour chaque objet le pas de quantification adéquat. Cependant, il s'agit de garder à l'esprit que les histogrammes devront être comparés. Afin que cette comparaison se fasse de façon rapide, nous avons décidé d'adopter une quantification uniforme des histogrammes. Dans un tel cas, le pas de quantification doit être identique pour tous les objets. Ces deux objectifs (pas de quantification unique mais adapté à l'objet) sont antagonistes. Il s'agit donc de trouver un compromis. Nous proposons de définir pour chaque objet le pas de quantification l qui lui est le plus adapté. Cette étape nous permet d'attribuer un ordre de grandeur à ce paramètre. Ensuite, nous constatons qu'il est possible de re-quantifier de façon exacte un histogramme d'une taille de classe K à une taille de classe L si K et L sont des puissances de 2 et $K < L$. L'idée est alors de choisir pour pas de quantification la puissance de 2 la plus proche de l.

Détermination de l'ordre de grandeur du pas de quantification

Compte tenu du fait que l'histogramme est défini sur l'objet et non sur l'image entière, il faut s'attendre à ce que le nombre de coefficients à disposition soit relativement faible, en particulier à BR. Nous décidons donc de déterminer le pas de quantification optimal à l'aide de la règle de Sturges (section 6.1.2). En utilisant cette règle, nous faisons l'hypothèse que la distribution des couleurs est unimodale et se rapproche d'une loi gaussienne. Dans le cas de la sous-bande BF, l'approximation de la distribution des couleurs par une loi gaussienne ou un mélange de lois gaussiennes est classique. Comme dans notre cas peu de point sont considérés, une gaussienne suffit. Pour les coefficients de HF, nous avons indiqué

que leur répartition correspond à une loi de Laplace (section 6.2.1). En approximant par une gaussienne, la description du pic sera peu précise. Nous pensons cependant qu'une telle approximation est acceptable. En effet, de part leur nature, les sous-bandes de HF contiennent de nombreux coefficients nuls ou quasi-nuls. L'information importante permettant de distinguer deux objets sera alors plutôt portée par les coefficients d'amplitude forte. Les coefficients de valeur faible n'ont donc pas besoin d'être quantifiés avec précision.
La loi de Sturges donne le nombre de classes à utiliser en fonction du nombre d'écantillons statistiques disponibles.

$$b = 1 + log_2(N) \tag{6.10}$$

Le pas de quantification $\hat{l}$ est déduit en fonction de l'intervalle de définition des coefficients Δ par :

$$\hat{l} = \frac{\Delta}{b} \tag{6.11}$$

Ici, le nombre d'échantillons statistiques disponibles est le nombre de pixels décrivant l'objet, soit $N = Card(O_i^k)$. Le nombre de classes b^k dans l'histogramme est déduit des équations précédentes :

$$b^k = 1 + log_2(Card(O_i^k)) \tag{6.12}$$

Comme nous l'avons indiqué dans le paragraphe 6.2.2, le descripteur considéré est composé d'histogrammes joints dans le domaine couleur YCrCb. Avec une telle définition, b^k est le nombre de classes dans l'espace 3D YCrCb. L'étape suivante est alors le calcul du nombre de classes marginal b_Y^k (resp. bk_{Cr}, bk_{Cb}) pour la composante Y (resp. Cr, Cb). La relation entre le nombre de classes des différentes composantes et le nombre de classes total est donnée par :

$$b^k = b_Y^k * b_{Cr}^k * b_{Cb}^k \tag{6.13}$$

En nous appuyant sur le fait que l'oeil humain est moins sensible à une petite variation des coefficients Cr et Cb que du coefficient Y, nous avons choisi de prendre $b_{Cr}^k = b_{Cb}^k = 0.5 b_Y^k$. On en déduit $b_Y^k = (4b^k)^{\frac{1}{3}}$. Etant donné l'intervalle de définition de coefficients de l'objet pour chaque composante, les largeurs de classes marginales $\hat{l}_Y^k$, $\hat{l}_{Cr}^k$ et $\hat{l}_{Cb}^k$ sont obtenues comme étant l/b.

Approximation à la puissance de 2 la plus proche

La largeur de classe arrondie choisie est définie comme étant la puissance de 2 la plus proche, c'est-à-dire :

$$\breve{l}_Y^k = 2^s, \text{ avec } s = \underset{s}{argmin} |2^s - \hat{l}_Y^k| \tag{6.14}$$

$\breve{l}_{Cr}^k$ et $\breve{l}_{Cb}^k$ sont définies de façon identique.

Il est certain que le pas de quantification ainsi choisi n'est pas complétement adapté à la distribution des coefficients d'ondelettes d'un objet. Cependant, il semble suffisamment précis pour permettre des comparaisons fiables, comme on le verra dans la partie résultats.

6.2.4 Métrique de similarité

Nous avons envisagé de tester deux mesures de similarité : l'intersection d'histogrammes de Swain et Ballard et le coefficient de Bhattacharyya (cf section 6.1.3). Ces deux mesures se calculent de façon très rapide puisque notre descripteur basé objet $H_{t,i}$ a un nombre raisonnable de classes (par exemple, les histogrammes LL et HF ont un nombre moyen de 20 classes en haut de la pyramide).
Etant donné deux histogrammes représentant les objets O_i et O_j respectivement aux niveaux k et k' de la pyramide d'ondelettes, nous rappelons la définition des métriques d'intersection d'histogrammes et de Bhattacharyya, après recalage en taille, avec les notations complètes. Pour une sous-bande $SB = \{LL, HF\}$, l'intersection d'histogrammes est :

$$d_{CH}^{SB}(O_i^k, O_j^{k'}) = \sum_{\gamma_{SB}} min(h_{SB,i}^k(O_i^k, \gamma_{SB}), h_{SB,j}^{k'}(O_j^{k'}, \gamma_{SB})) \tag{6.15}$$

et le coefficient de Bhattacharyya :

$$d_{B}^{SB}(O_i^k, O_j^{k'}) = \sum_{\gamma_{SB}} \sqrt{h_{SB,i}^k(O_i^k, \gamma_{SB}) h_{SB,j}^{k'}(O_j^{k'}, \gamma_{SB})} \tag{6.16}$$

Les mesures de similarité pour le descripteur H_i sont définies par combinaison linéaire des mesures de similarité sur les sous-bandes.

$$d_{CH} = \alpha d_{CH}^{LL} + (1-\alpha) d_{CH}^{HF} \text{ avec } \alpha \in [0,1] \tag{6.17}$$

$$d_{B} = \beta d_{B}^{LL} + (1-\beta) d_{B}^{HF} \text{ avec } \beta \in [0,1] \tag{6.18}$$

Les deux mesures varient de 0 (non similaire) à 1 (similaire). Ainsi deux objets sont similaires si leur mesure de similarité est supérieure à un seuil donné. Notons que les mesures de similarité d_{CH} et d_B peuvent considérer des objets au même niveau de décomposition $k = k'$ ou à des niveaux de décomposition différents $k \neq k'$.

6.3 Résultats et conclusion

Dans ce chapitre, nous avons défini un descripteur scalable des objets en mouvement extraits sur une pyramide d'ondelettes. Ce descripteur est défini à l'aide des histogrammes d'ondelettes. Nous ne présenterons pas d'évaluation du descripteur. En effet, cette évaluation se fera en fonction des tâches d'indexation à réaliser. Elle est donc indiquée dans

	Masques Manuels				Masques Automatiques			
	Sous-Bande BF		Sous-Bandes HF		Sous-Bande BF		Sous-Bandes HF	
Niveau	Espace	descript.	Espace	descript.	Espace	descript.	Espace	descript.
0	121	51	NS	NS	113	46	NS	NS
1	117	45	73	25	113	41	108	35
2	105	38	73	23	108	35	73	22
3	96	32	72	22	98	28	71	20
4	93	26	70	20	90	21	68	17

Tab. 6.1 – Tableau des nombres de classes moyens en fonction des niveaux de la pyramide. Les résultats sont donnés pour les objets obtenus par segmentation automatique et segmentation manuelle. Nous présentons séparément les nombres de classes obtenus pour chaque type de sous-bande. Enfin nous indiquons le nombre de classes de l'espace de représentation (espace) et le nombre de classes non vides représentant effectivement le descripteur (descript).

le chapitre 7, une fois les tâches d'indexation scalable définies. Le tableau 6.1 illustre le nombre de classes d'histogramme utilisées en moyenne sur la base de données de test. Nous donnons les résultats pour deux types d'obtention des masques : par segmentation automatique et par segmentation manuelle. La segmentation manuelle représente le résultat idéal de la segmentation et nous permet de montrer les résultats obtenus si la segmentation était parfaite. On constate que le nombre de classes décroît avec le niveau de résolution. De plus, comme attendu, il faut moins de classes pour décrire les coefficients de HF que de BF. Nous comparons aussi le nombre de classes pavant l'espace de représentation au nombre de classes effectivement non vides utilisées pour le descripteur. Ce second est nettement plus faible que le premier. Enfin notons que le descripteur est compact. Il faut au maximum un descripteur de dimension 2x121 pour décrire un objet à pleine résolution. Une approche de type indexation locale par points SIFT nécessite plusieurs descripteurs de taille 128.

Nous proposons de décrire les objets sur chaque niveau de décomposition afin de fournir une description scalable en résolution des objets. Nous proposons de construire chaque niveau de descripteur à partir de deux histogrammes, un histogramme décrivant les coefficients de HF et un histogramme décrivant les sous-bandes de BF. Une attention particulière a été apportée au choix du nombre de classes nécessaires à la construction de l'histogramme pour permettre une description pertinente des objets et des comparaisons d'histogrammes rapides.

Chapitre 7

Evaluation et résultats

L'objectif des méthodes développées dans les chapitres précédents est de fournir un descripteur scalable des vidéos à des fins d'indexation. L'objet de ce chapitre est de tester les capacités du descripteur proposé à répondre à des tâches d'indexation. Une seule tâche d'indexation est choisie pour tester le descripteur : il s'agit des requêtes par similarité dans des bases de données vidéo. Après avoir établi la procédure permettant à notre descripteur de répondre à une requête, nous présentons la BD que nous avons utilisée pour nos tests. Les scénarios de requête envisagés seront détaillés à cette occasion. Nous évaluons ensuite les réponses de notre descripteur à ces deux scenarios. Enfin, nous comparerons les performances de notre descripteur global basé objet avec celles d'un descripteur local basé objet.

7.1 Procédure de réponse à une requête scalable basée objet

Nous souhaitons utiliser notre descripteur scalable pour répondre à des tâches de requêtes par similarité dans des bases de données. Pour répondre à une telle tâche, un système d'indexation doit, étant donné une vidéo requête fournie par l'utilisateur, renvoyer toutes les vidéos de la BD qui lui sont similaires et rien que ces vidéos. Nous définissons dans la suite une procédure de comparaison des vidéos par notre descripteur.
Une constatation s'impose, bien que les résultats de l'extraction d'objets soient acceptables, le processus de segmentation reste un processus instable qui dépend fortement de l'amplitude relative du mouvement de l'objet par rapport à celui du fond. A moins d'utiliser des outils plus complexes tels que le suivi d'objet, l'extraction de l'objet sur chaque trame de la séquence n'est pas garantie. Ainsi, dans l'état actuel de notre travail, nous proposons d'effectuer une recherche par trame.

Soit un clip C_{BD} pris dans la BD video. L'ensemble des masques d'objet $O_{BD} = \{O_{t,i}, t = t_0, t_0 + \Delta t, ..., \}$, où Δt est le facteur de sous-échantillonage dans le temps, est extrait pour chaque objet à chaque niveau de la pyramide d'ondelettes. Les descripteurs d'histogrammes H_{BD} correspondants sont calculés et enregistrés comme des méta-données. Soit un clip de requête C_Q et l'ensemble des descripteurs H_Q des objets extraits de ce clip. L'utilisateur est invité à sélectionner une trame de C_Q dans laquelle le résultat de l'extraction d'objets est le plus satisfaisant. Le descripteur associé à cette trame est comparé individuellement à chaque descripteur associé à une trame dans le clip C_{BD} en utilisant une des métriques de similarité proposées dans la section 6.2.4. Les différentes valeurs de ressemblance ainsi mesurées sont comparées et seule la plus forte ressemblance est conservée et est prise comme mesure de similarité entre le clip C_Q et le clip C_{BD}. Le clip C_Q est comparé à toutes les vidéos de la BD et seules les vidéos qui donnent un score de similarité supérieur à un certain seuil sont considérées comme étant similaire au clip requête.
La procédure présentée partage la philosophie de la recherche par image-clé. En effet, la recherche s'effectue d'après le choix d'un résultat d'extraction d'objet en mouvement représentatif de la vidéo. Cependant, contrairement à la recherche par image-clé, aucune notion de trame représentative n'est utilisée pour les vidéos de la base. La sélection des images-clés n'est donc pas une source d'erreur dans notre approche.

Le descripteur que nous proposons est multirésolution, ce qui nous permet de choisir le niveau de résolution dans la pyramide d'ondelettes à utiliser pour faire la comparaison. Deux types de recherche sont considérés dans le cadre de ce manuscrit de thèse : la recherche mono-niveau et la recherche à niveaux croisés. Dans le cas de la recherche mono-niveau, seule la portion H_Q^k du descripteur H_Q qui décrit le niveau k est utilisée et comparée à chaque portion H_{BD}^k de descripteur des trames de la BD au même niveau de résolution. Dans la recherche à niveaux croisés, les portions de descripteurs H_Q^k et $H_{BD}^{k'}$, avec $k \neq k'$ sont comparées. Nous pensons que ces deux types de requête décrivent complètement toutes les situations occasionnées par l'utilisation de flux compressés scalables. Nous présentons ici quelques situations qui nous sont venues à l'esprit au moment de rédiger ce manuscrit.

1. Supposons que la requête soit faite par un client à un serveur au travers d'un réseau. Le serveur a en charge la BD. Supposons que le client a une version originale de la vidéo (c'est-à-dire qu'elle n'a pas subi de transformations frauduleuses et n'a pas été ré-encodée) mais que cette version originale ne représente qu'une part du flux scalable original correspondant à la capacité du client. Alors une recherche en mono-résolution est parfaitement adaptée, le niveau k de recherche étant fixé par la capacité en résolution du client.

2. Supposons, en plus des conditions du cas 1, que la vidéo servant de requête n'a pas de descripteur embarqué. Le client doit alors calculer le descripteur et l'envoyer au

serveur. Afin de réduire le coût calculatoire, le descripteur n'est calculé qu'à la plus Basse Résolution. Le serveur peut choisir d'effectuer une recherche en mono-résolution comme dans le cas 1 ou une recherche à niveaux croisés en espérant que l'utilisation de la HR améliore la qualité des résultats.

3. Considérons maintenant une tâche de détection de copies. Etant donnée une vidéo originale, le but est de retrouver toutes les versions transformées de cette vidéo dans la BD. Les transformations subies peuvent être géométriques (rotation, rognage...) ou dues à un processus de décodage ré-encodage multiple d'une portion du flux scalable. La recherche à niveaux croisés est utilisée où le descripteur requête est utilisé à la pleine résolution et le descripteur de la base à une résolution quelconque.

4. Le cas inverse du cas 3 est que l'utilisateur a une version modifiée de la vidéo et veut retrouver la séquence originale dans la base de données. Là encore, c'est un cas de recherche à niveaux croisés où le descripteur requête est à BR et les descripteurs de la base sont à HR.

7.2 Présentation de la base de données

Pour pouvoir attester de la robustesse et de la qualité des descripteurs proposés à diverses transformations de la vidéo (transformations affines du plan image, masquage partiel, post-production frauduleuse...), il faut avoir à disposition une base de données de vidéos transformées et annotées manuellement. Dans le cas de la recherche par le contenu des images, de nombreuses bases de données libres de droits existent. Par exemple, le banc de test utilisé dans [Nis06] contient quatre versions de la même image fixe obtenues par transformations affines. Dans le cas des contenus vidéo, de telles bases n'existent pas, particulièrement dans le cadre des vidéos HD avec lesquelles nous travaillons. Les quelques flux HD disponibles sont de courtes vidéos de test pensées pour évaluer la qualité du codage vidéo mais pas la qualité de l'indexation. Parmi ce type de bases de données, on peut citer les corpus VQEG [VQE], TUM [TUM]. De telles bases de données ne contiennent pas de séquences vidéos transformées utiles pour faire des tests d'indexation et de recherche. Dans le cadre du projet ICOS-HD, nous avons développé un corpus. Ce corpus est constitué de vidéos tournées au LaBRI[1] au format 1080p ou en transformant des séquences des corpus VQEG/SVT, TUM et autres corpus publics.
La base de données utilisée dans nos tests est constituée de 18 clips originaux de 3s à un taux de 25fps. La première image de chacun de ces clips est présentée dans la figure 7.1.

Les clips sont disponibles en format brut. Les 14 séquences (a-k,n,o,r) sont au format 1080psf, les 4 séquences restantes (l,m,p,q) sont au format 1920x896psf. La séquence (r) est

[1]Laboratoire Bordelais de Recherche en Informatique

(a) visu_Eliana, clip1

(b) visu_ChrisDanielRemi, clip1

(c) visu_ZoomChris, clip1

(d) pedestrian_area, clip1

(e) sunflower, clip1

(f) tractor, clip1

(g) parkjoy, clip1

(h) kungFu, clip1

(i) kungFu, clip2

(j) lancer_trousse, clip1

(k) lancer_trousse, clip2

(l) man_in_restaurant, clip1

(m) man_in_restaurant, clip2

(n) pieton_exterieur, clip1

(o) pieton_exterieur, clip2

(p) street, clip1

(q) street, clip2

(r) montage_frauduleux, clip1

Fig. 7.1 – Présentation des 18 clips originaux du corpus ICOS HD utilisés comme BD

une séquence produite par montage frauduleux. L'objet Vincent a été extrait manuellement de toutes les trames de la séquence "lancer_trousse", clip1 (j), et intégré dans un nouveau fond. Certains clips sont issus de la même vidéo à deux instants de temps différents. Chacun des clips a subi une série de transformations géométriques, à savoir rognage à la taille 960x540, symétries axiales verticale et horizontale, redimensionnement aux tailles 960x540 et 480x270, rotations d'angle 10^o, 20^o, 30^o, 40^o, 45^o et 190^o ; ces transformations étant appliquées sur chaque trame individuellement. Une illustration de ces transformations sur la séquence (d) "pedestrian_area" est donnée dans la figure 7.2.

La BD complète contient donc 216 clips dans lesquels, selon notre méthode, 3240 trames ont été traitées. Pour chacune de ces trames, une segmentation manuelle des masques d'objets a été réalisée par rotoscoping pour pouvoir évaluer la robustesse du descripteur proposé par rapport aux bruits de segmentation dans les tâches de recherche dans les bases de données.

La BD ainsi construite nous permet de définir deux scénarios de recherche selon la réponse attendue à la requête.

- Scenario 1 : Recherche d'un clip contenant une copie exacte de l'objet requête. Le but est de retrouver toutes les vidéos de la BD qui contiennent le même objet, dans la même posture et vu par le même angle de caméra que dans la vidéo requête. Pour la video (j), il s'agira de retrouver toutes les versions de (j) et de (r) (originale et transformées)

- Scenario 2 : Recherche d'un clip contenant le même objet que l'objet requête. Le but est de retrouver toutes les vidéos de la BD qui contiennent le même objet. Contrairement au scenario 1, il n'y a pas de contraintes sur la posture ou la position de la caméra. Les copies exactes du scenario 1 sont clairement des réponses à ce type de requête. Cependant, plus de vidéos doivent être retournées à l'utilisateur. Ainsi, pour la video (j), il faudra aussi récupérer la vidéo (k) et ses transformées.

7.3 Evaluation du descripteur

Les données de tests et les scenarios de requêtes ayant été établis, il est maintenant possible d'évaluer la robustesse et l'efficacité du descripteur. Les performances du descripteur sont mesurées par les courbes de rappel/précision interpolées moyennes (cf. section 3.1.4). Ces évaluations permettent de choisir entre les métriques de similarité entre histogrammes (intersection d'histogrammes de Swain et Ballard et coefficient de Bhattacharyya, section 6.2.4) laquelle est la plus performante. Nous évaluerons ensuite la robustesse de notre méthode par rapport au bruit de segmentation. Puis, nous montrerons la scalabilité du

Fig. 7.2 – Illustration des transformations géométriques appliquées à une image. Dans l'ordre lexicographique : original, symétrie d'axe horizontal, symétrie d'axe vertical, rognage à la taille 960x540, redimensionnement à la taille 960x540, redimensionnement à la taille 480x270, rotation d'angle 10^o, rotation d'angle 20^o, rotation d'angle 30^o, rotation d'angle 40^o, rotation d'angle 45^o et rotation d'angle 190^o.

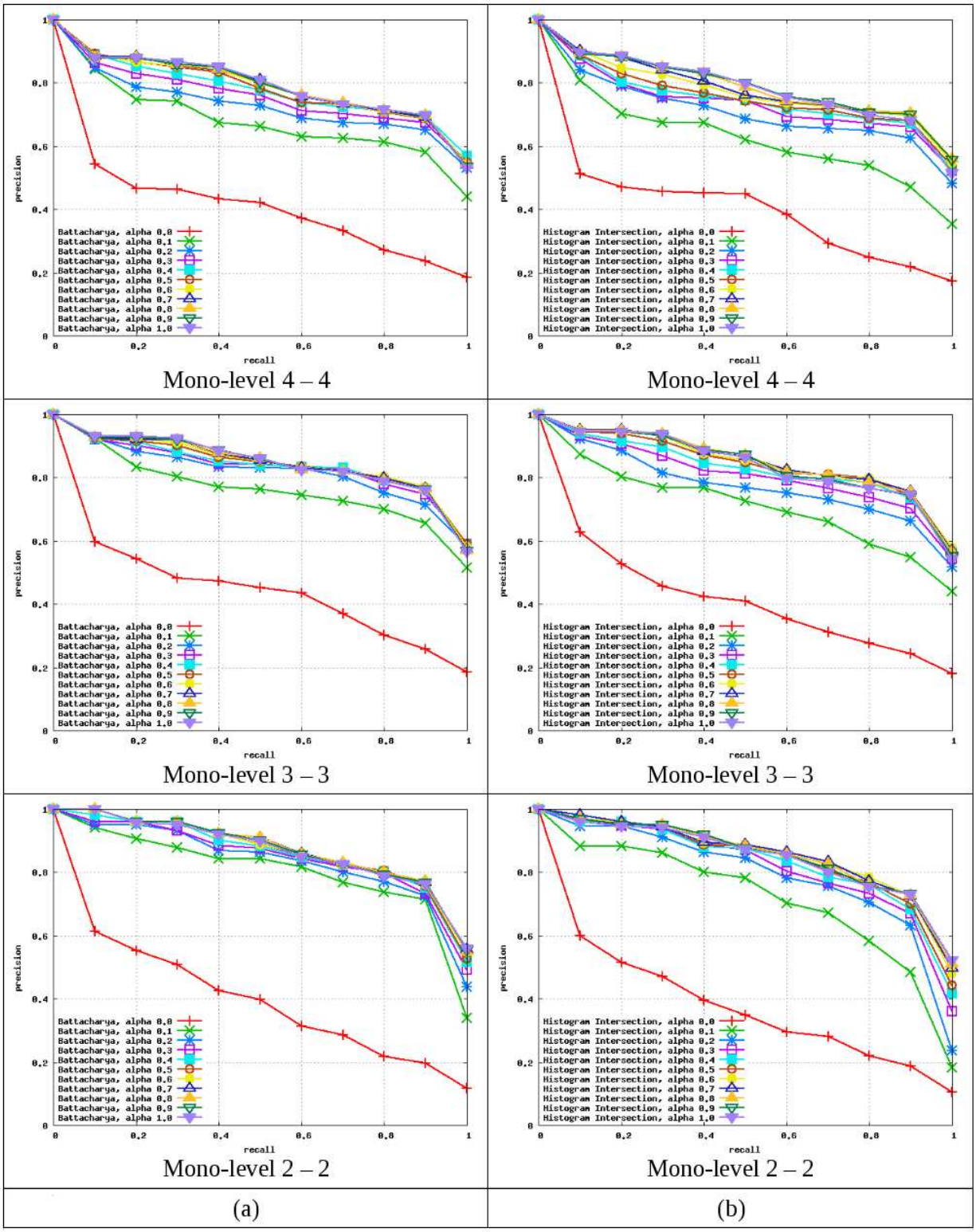
descripteur pour des tâches de requêtes. Enfin, nous comparerons les résultats de notre descripteur avec ceux d'un descripteur basé sur les points SIFT des objets.

7.3.1 Choix de la métrique de similarité et des proportions de mélange pour le descripteur d'objets basé histogramme

Nous comparons ici le comportement des mesures par coefficient de Bhattacharyya et intersection d'histogrammes de Swain et Ballard dans le cas de la recherche mono-niveau. D'abord, comme les masques d'objets extraits ne sont pas parfaits et la plupart du temps des morceaux de l'objet ne sont pas détectés, il faut déterminer quelle est la métrique la plus robuste à ce type de perte d'information. Ensuite, le choix du coefficient de mélange α (cf equations (6.17) et (6.18)) le plus approprié doit être effectué. Les figures 7.3 et 7.4 décrivent le comportement de la mesure de similarité du coefficient de Bhattacharyya (a) et de l'intersection d'histogrammes (b) pour différentes valeurs de $\alpha = 0, 0.1, ..., 1$ pour les deux scenarii sur les masques d'objet extraits de façon automatique.

Il est à noter que pour avoir des valeurs équilibrées de Rappel-Précision, le coefficient de Bhattacharyya est systématiquement meilleur que l'intersection d'histogrammes. L'utilisation seule des coefficients de HF ne permet pas d'obtenir des résultats satisfaisants (courbes rouges sur les figures 7.3 et 7.4). Comme il s'agit uniquement de l'information de HF, celle-ci n'est pas suffisamment riche pour permettre une reconnaissance complète. A partir de $\alpha = 0.5$, le coefficient de HF améliore systématiquement le résultat, la précision interpolée est meilleure pour les mêmes valeurs de rappel. Le comportement des métriques de similarité en fonction de la pondération α est identique quelle que soit la métrique utilisée et le scénario envisagé. Cela traduit une certaine robustesse du descripteur aux diverses situations qui se présentent à lui.
Afin de mettre en avant le fait que la mesure de Bhattacharrya est meilleure en terme de l'équilibre Rappel/Précision, nous l'avons comparée à la méthode d'intersection d'histogrammes de Swain et Ballard pour $\alpha = 0.7$ uniquement dans le cas du scénario 1 (cf figure 7.5) et du scénario 2 (cf figure 7.6). On constate que quel que soit le niveau, la courbe de rappel-précsion moyenne interpolée est meilleure pour Bhattacharyya que pour l'intersection d'histogrammes.

Notons enfin que les courbes de Rappel-Précision interpolées sont plus élevées dans le cas du scénario 1 que dans le cas du scénario 2. Cela est dû au fait que l'exigence de recherche est moins forte dans le scénario 2 que dans le scénario 1.
La mesure de similarité choisie dans la suite est la mesure de Battacharya avec $\alpha = 0.7$.



Mono-level 4 – 4 | Mono-level 4 – 4

Mono-level 3 – 3 | Mono-level 3 – 3

Mono-level 2 – 2 | Mono-level 2 – 2

(a) | (b)

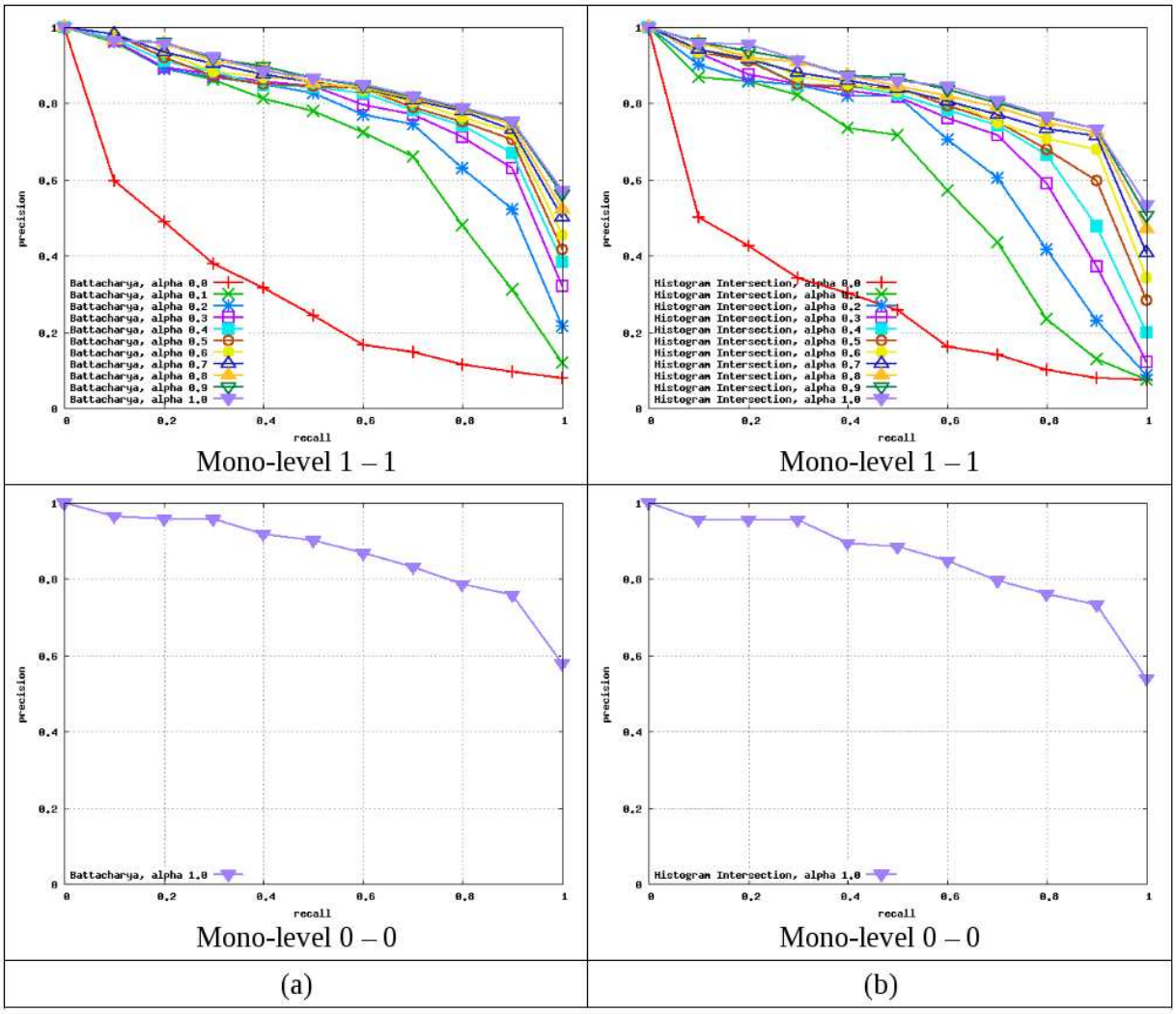
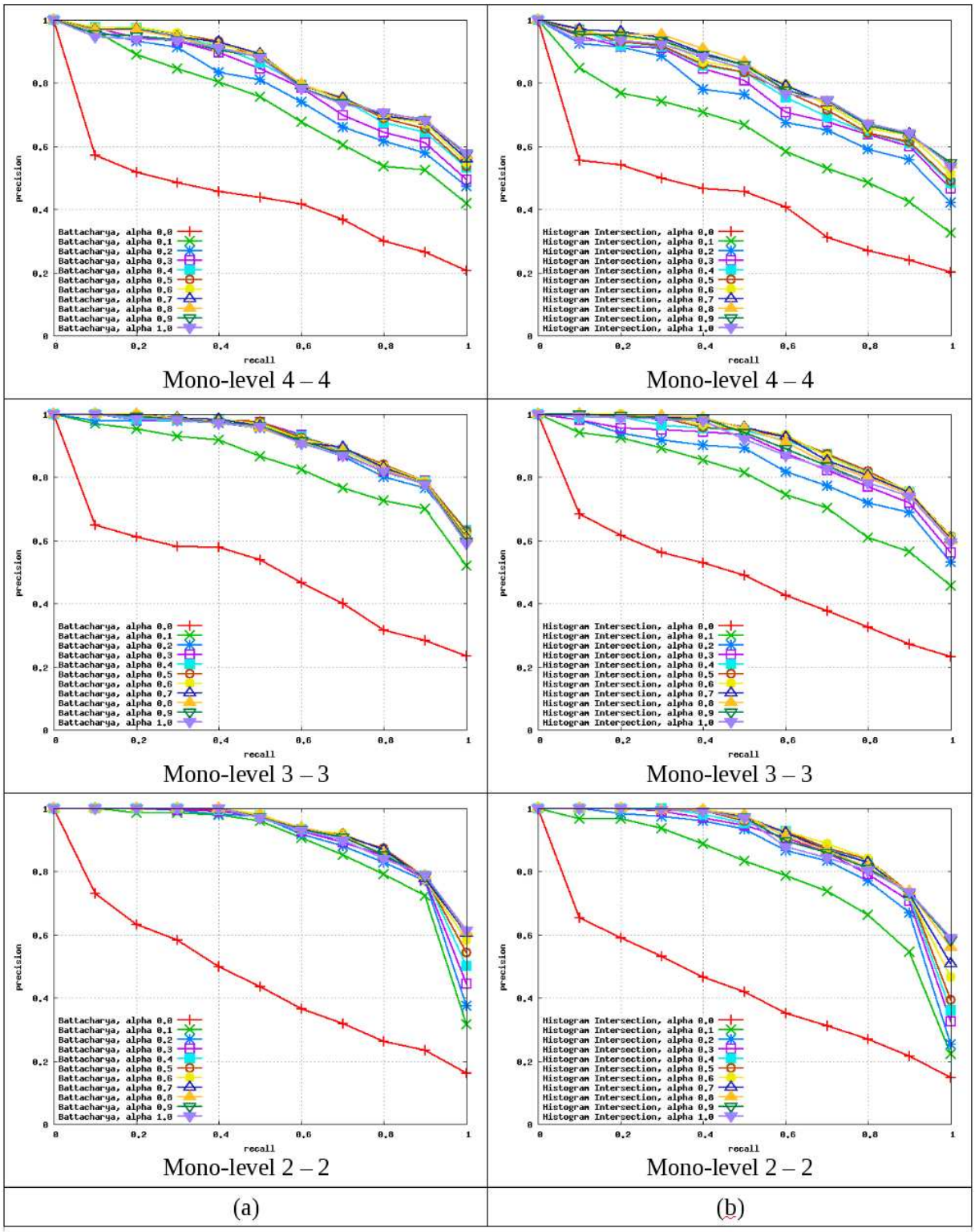


Fig. 7.3 – Comparaison des métriques de similarité pour le descripteur basé histogramme sur les masques d'objets extraits (a) Bhattacharyya, (b) Swain and Ballard. Scenario de requête 1.


precision
recall
Battacharya, alpha 0.0
Battacharya, alpha 0.1
Battacharya, alpha 0.2
Battacharya, alpha 0.3
Battacharya, alpha 0.4
Battacharya, alpha 0.5
Battacharya, alpha 0.6
Battacharya, alpha 0.7
Battacharya, alpha 0.8
Battacharya, alpha 0.9
Battacharya, alpha 1.0
Histogram Intersection, alpha 0.0
Histogram Intersection, alpha 0.1
Histogram Intersection, alpha 0.2
Histogram Intersection, alpha 0.3
Histogram Intersection, alpha 0.4
Histogram Intersection, alpha 0.5
Histogram Intersection, alpha 0.6
Histogram Intersection, alpha 0.7
Histogram Intersection, alpha 0.8
Histogram Intersection, alpha 0.9
Histogram Intersection, alpha 1.0
Mono-level 4 – 4
Mono-level 4 – 4
Mono-level 3 – 3
Mono-level 3 – 3
Mono-level 2 – 2
Mono-level 2 – 2
(a)
(b)

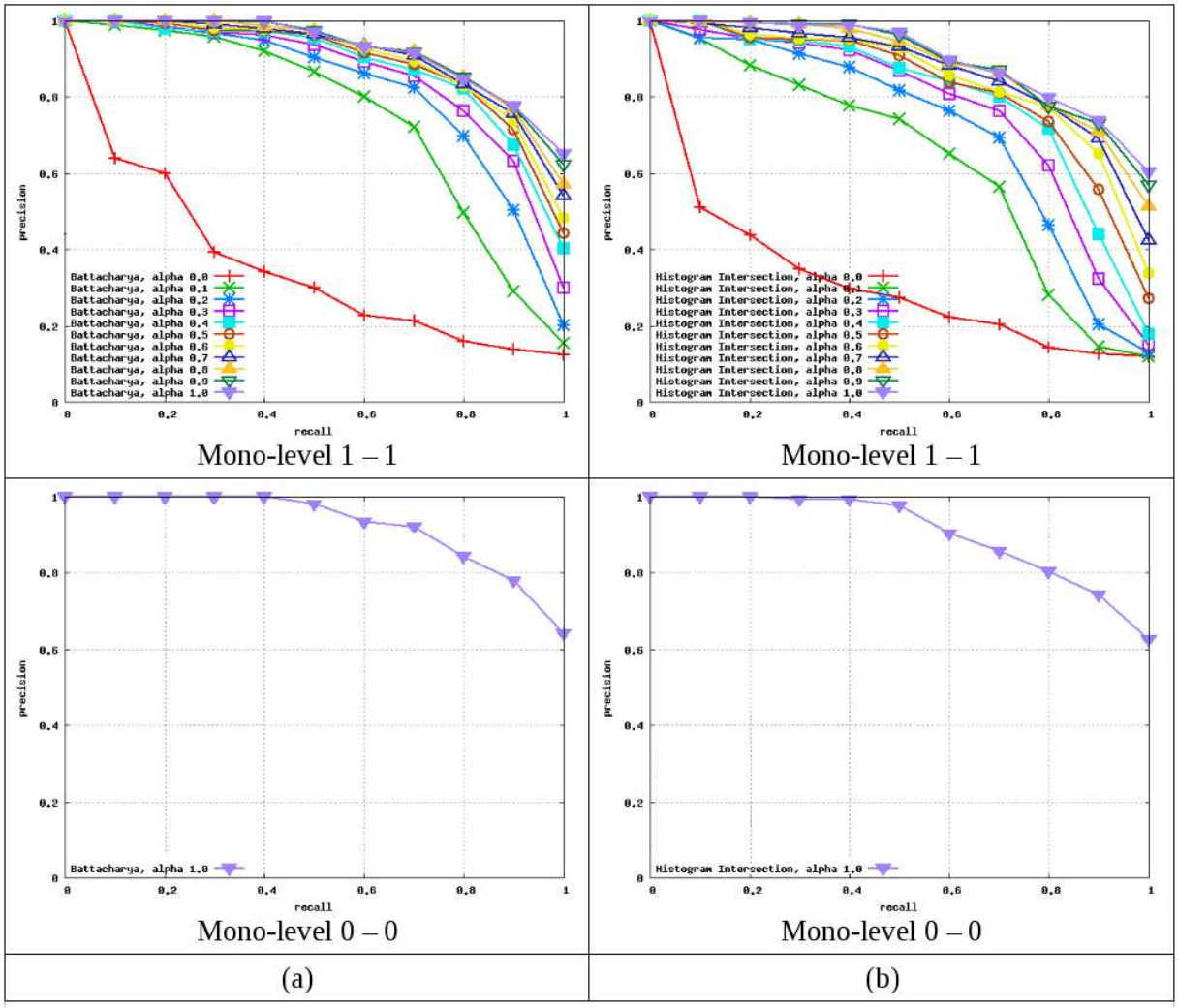


Fig. 7.4 – Comparaison des métriques de similarité pour le descripteur basé histogramme sur les masques d'objets extraits (a) Bhattacharyya, (b) Swain and Ballard. Scenario de requête 2.

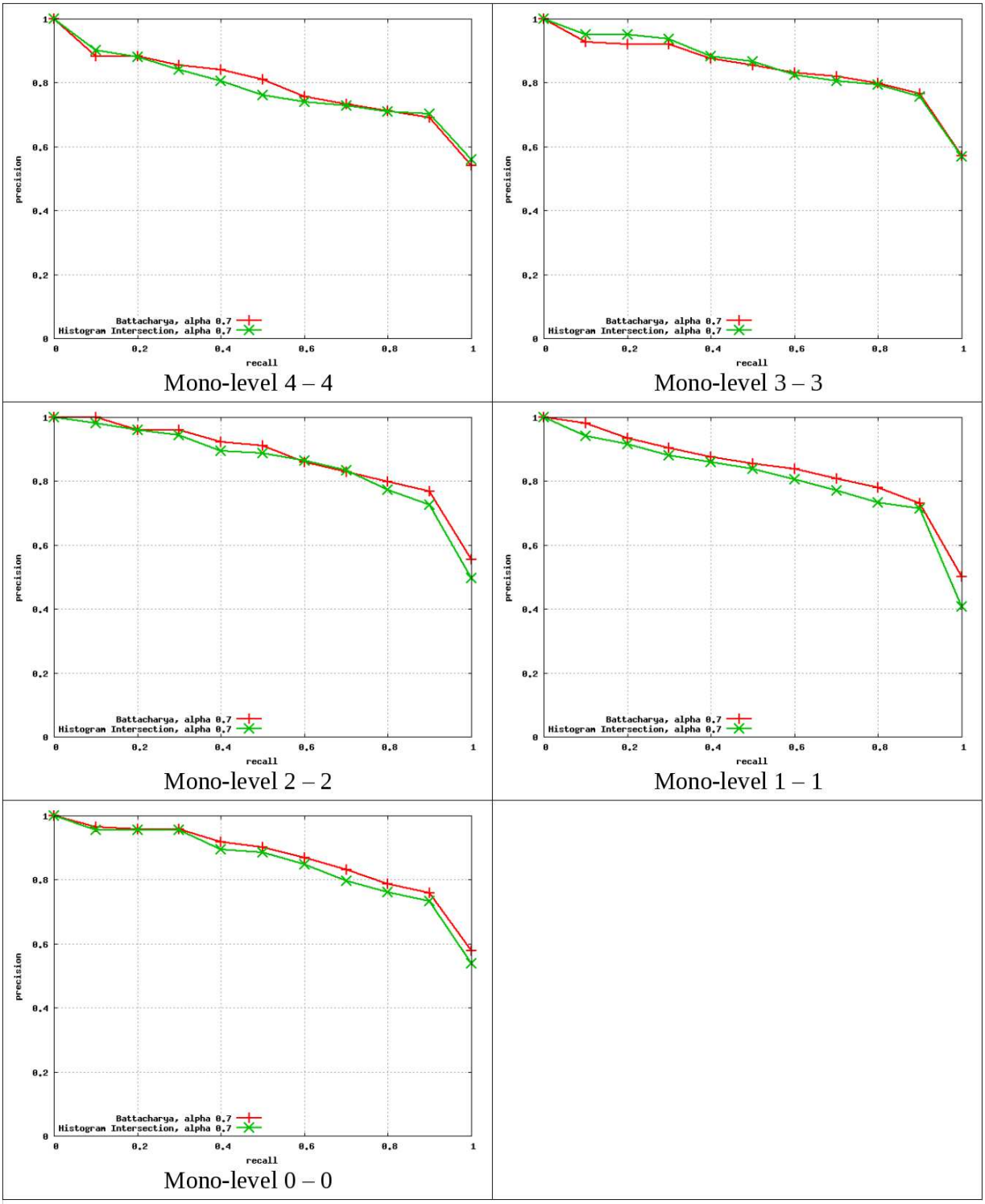


Fig. 7.5 – Comparaison des métriques de similarité pour le descripteur basé histogramme sur les masques d'objets extraits automatiquement. Scenario de requête 1.

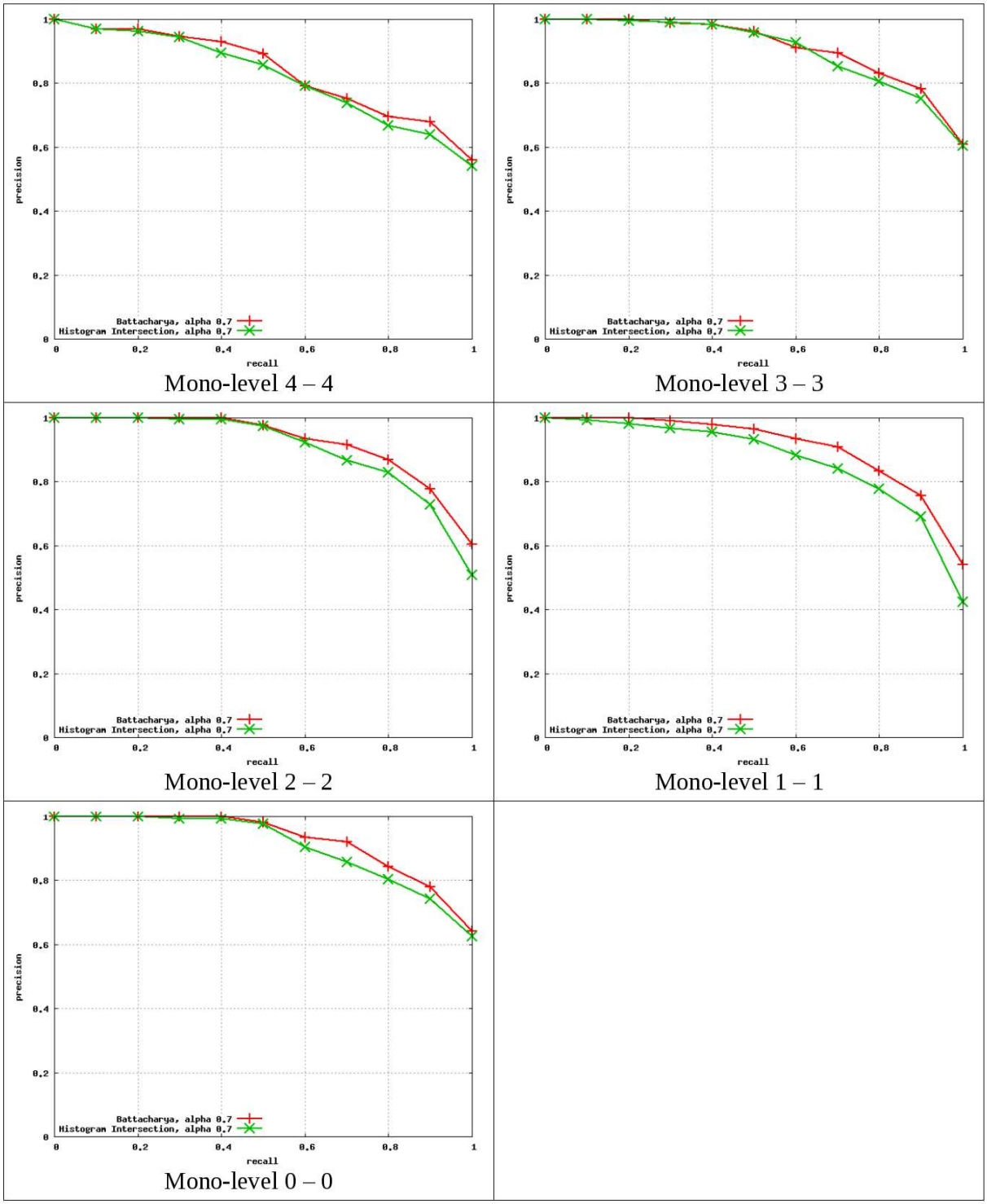


Fig. 7.6 – Comparaison des métriques de similarité pour le descripteur basé histogramme sur les masques d'objets extraits automatiquement. Scenario de requête 2.

7.3.2 Influence du bruit de segmentation sur les performances du descripteur

Il est bien connu que le comportement de la segmentation est instable lorsque l'on change les paramètres de la segmentation. Aussi, il est impossible de faire décroitre progressivement la qualité de la segmentation et mesurer la performance de la recherche dans les bases de données. A la place, nous comparons les performances de la méthode sur les masques d'objet extraits *manuellement et automatiquement*. Des exemples de ces deux types d'extraction pour la séquence "Zoom Chris" du LaBRI sont donnés dans la figure 7.7.

La figure 7.8 représente les courbes de Rappel/Précision interpolées moyennes pour le scénario 1 dans le cas où les masques sont extraits manuellement - donc supposés "parfaits" - (courbe verte) et automatiquement - présentant de nombreux défauts - (courbe rouge). La figure 7.9 présente le même type de courbes dans le cadre du scénario 2. Commençons par analyser les courbes correspondant aux masques manuels. Elles sont proches de l'idéal. Cela nous amène à conclure que les descripteurs basés objets sont efficaces pour les tâches de recherche dans les bases de données. De plus, ces résultats étant de qualité identique à tous les niveaux de la pyramide, le descripteur est multirésolution ce qui est une propriété désirable pour arriver à la propriété de scalabilité. L'utilisation d'un descripteur global tel que les histogrammes d'ondelettes permet de conserver cette qualité de résultats lorsque le résultat de la segmentation est incertain. Ainsi, même si les courbes sont moins satisfaisantes dans le cas des masques automatiques que dans le cas des masques manuels, le couple Rappel-Précision reste très correct, autour de 0.8 pour le résultat le mieux équilibré.

Fig. 7.7 – Masques automatiques et masques manuels obtenus pour la séquence “Zoom Chris” du LaBRI

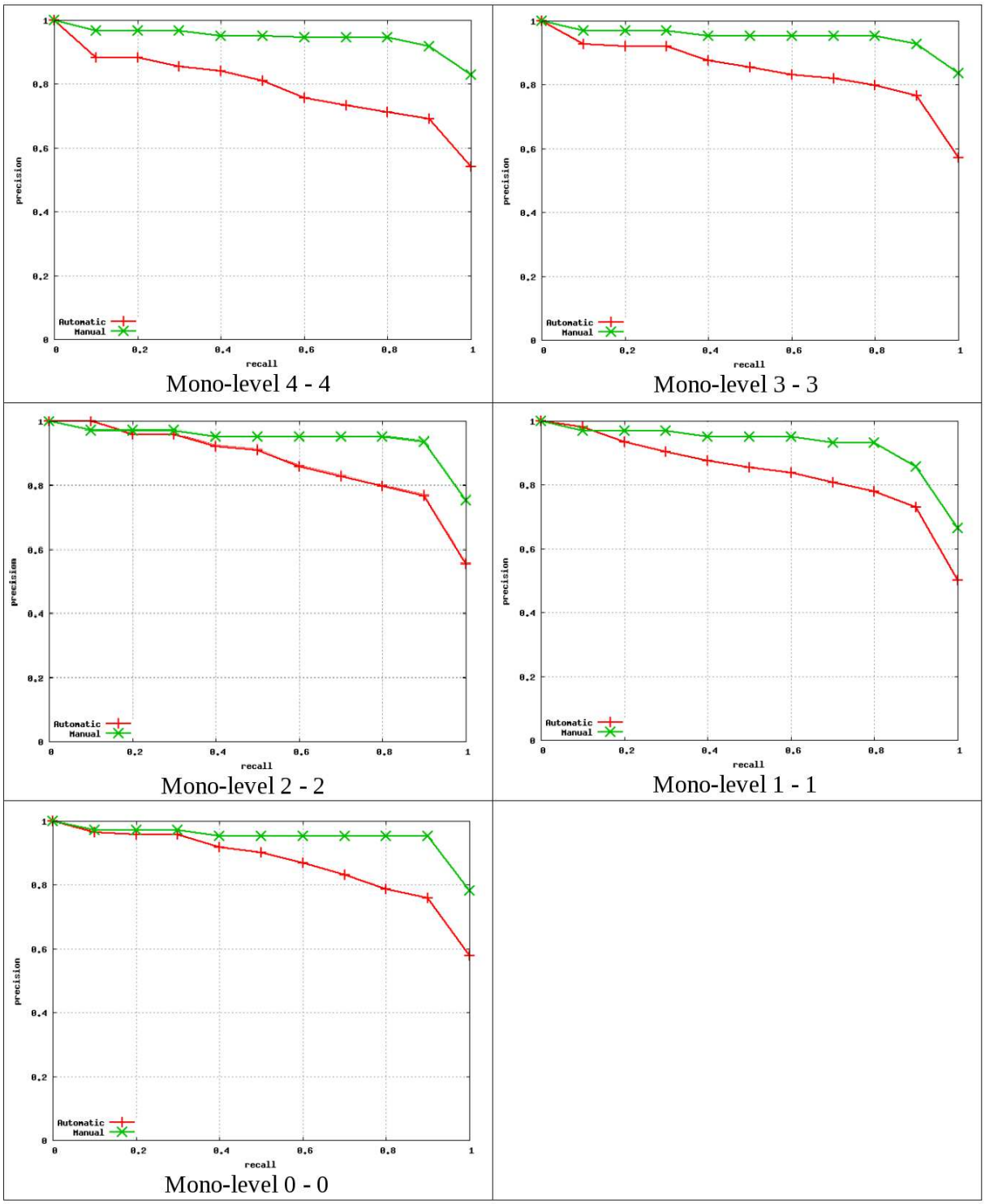


Fig. 7.8 – Comparaison des performances de la mise en correspondances du descripteur sur les masques extraits manuellement et automatiquement. Scenario de requête 1.

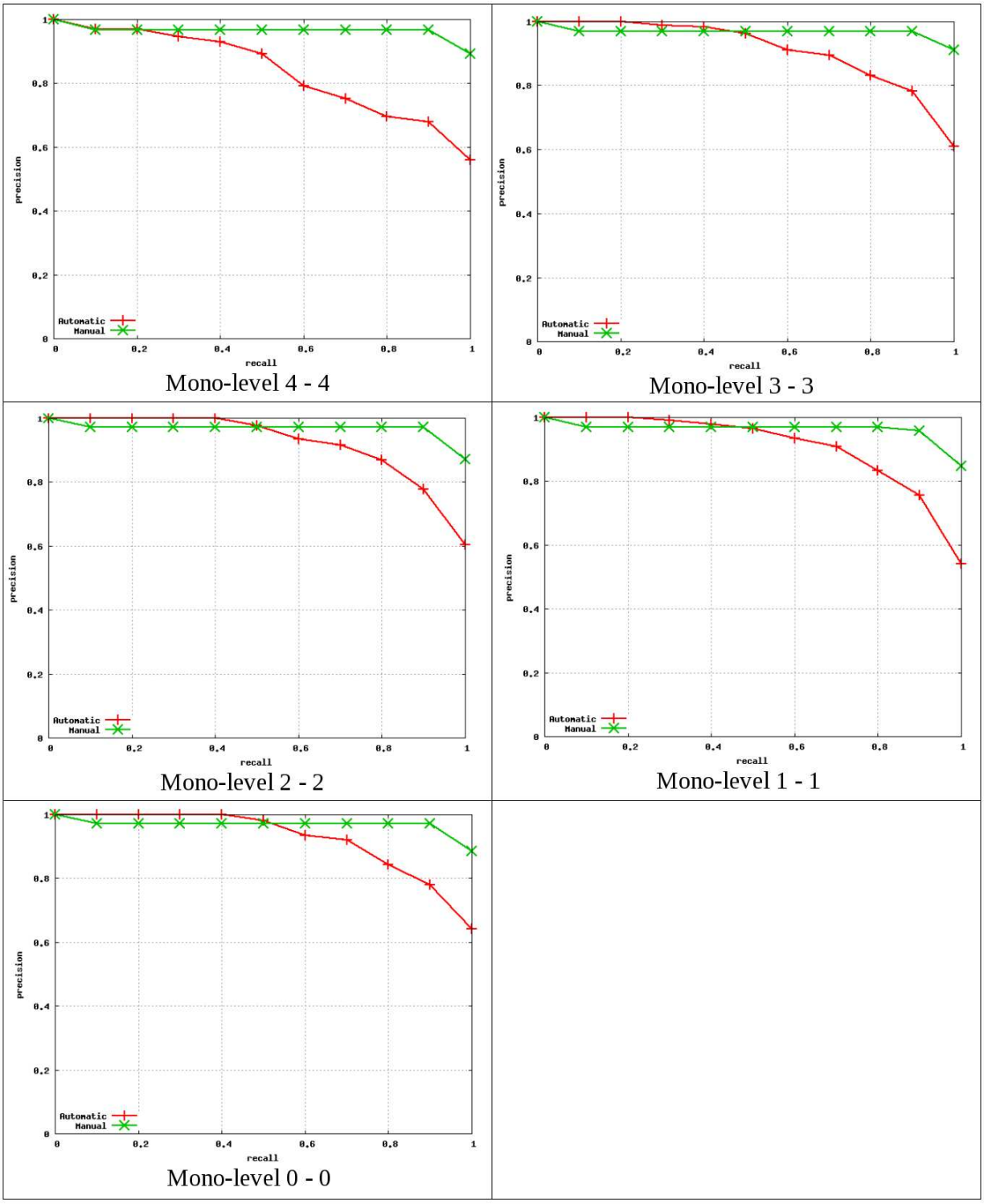


Fig. 7.9 – Comparaison des performances de la mise en correspondances du descripteur sur les masques extraits manuellement et automatiquement. Scenario de requête 2.

7.3.3 Scalabilité des tâches de requêtes

Dans cette expérimentation, nous appliquons la méthode pour des recherches à niveaux croisés avec les masques d'objet extraits automatiquement. Les résultats de requêtes de basse résolution vers haute résolution sont donnés dans la figure 7.10 pour le scénario 1 et dans la figure 7.11 pour le scénario 2. De même les résultats de requête de haute résolution vers basse résolution sont donnés dans les figures 7.12 et 7.13. Afin de permettre des comparaisons, nous avons proposé de comparer à la pleine résolution et au niveau de résolution 1. En effet, à la pleine résolution, le descripteur ne contient pas d'information de HF et la comparaison ne s'effectue qu'entre histogrammes couleurs. Pour faire intervenir les coefficients d'ondelettes de HF, il faut utiliser un niveau de résolution supérieur ou égal à 1. De même, pour conserver des écarts entre niveaux comparables, nous avons considéré la Basse Résolution comme étant le niveau 3 ou le niveau 4.
Toutes les courbes présentées dans les sections précédentes (en mono-niveau) et dans cette section (en niveaux croisés) ont un profil quasiment identique. Le descripteur proposé apporte des réponses identiques et de haute qualité quel que soit le type de requête considéré. Autrement dit, notre descripteur est scalable. au sens de la définition donnée dans le chapitre 1.
Ainsi, par construction, le descripteur d'objets par histogrammes d'ondelettes peut être extrait à partir de n'importe quelle portion de flux scalable et uniquement à partir des données disponibles dans le flux compressé. De plus, les réponses qu'il apporte dans les tâches de recherche par similarité sont quasi-indépendantes de la version de la vidéo utilisée pour l'extration de l'indice. Avec un tel descripteur, il est possible de répondre au problème des requêtes scalables telles que définies par quelques exemples concrets dans la section 7.1.

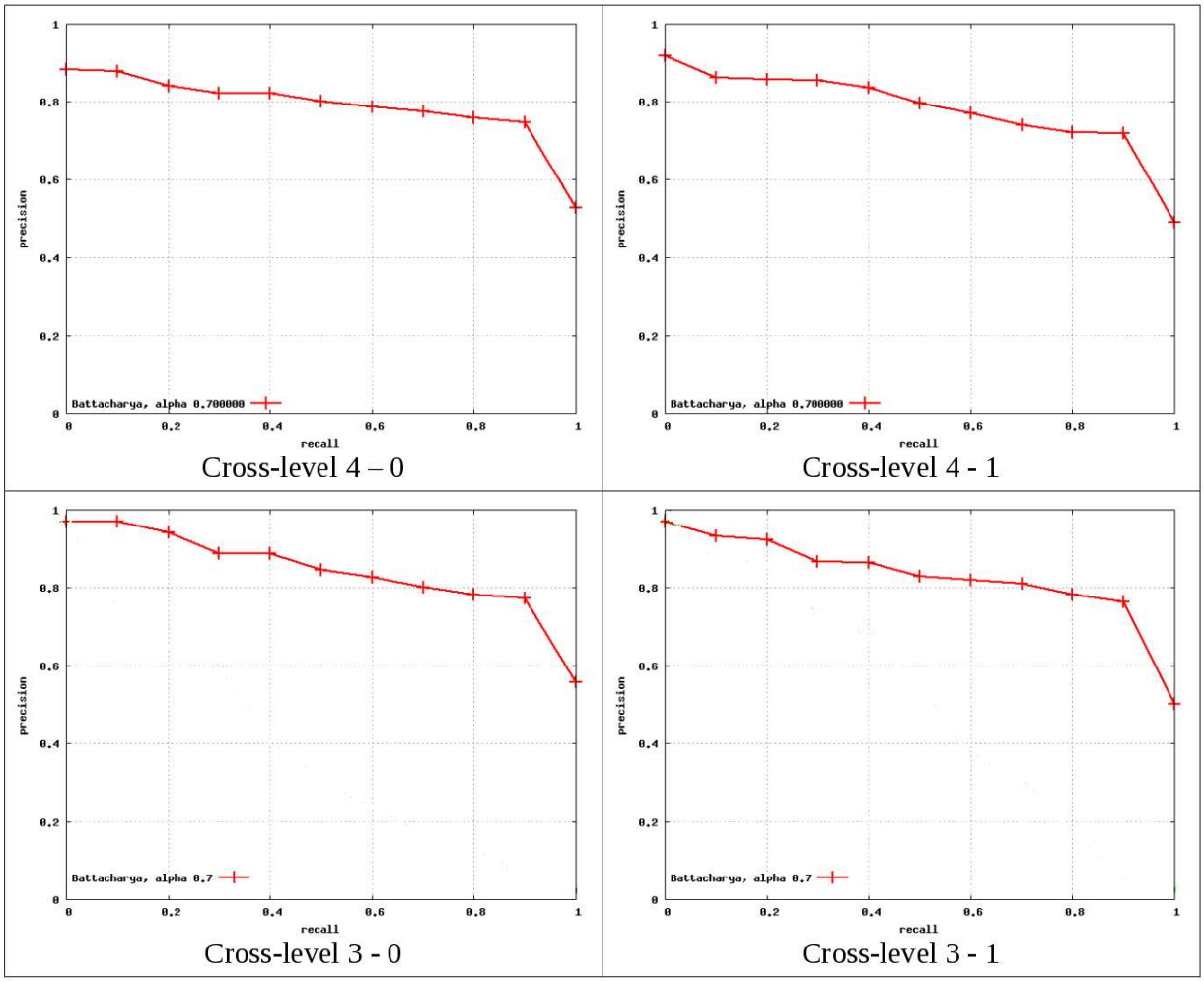


Fig. 7.10 – Courbes de Rappel-Précision interpolées moyennes pour les recherches à niveaux croisés de la BR (requête) vers la HR (BD) sur les masques extraits automatiquement. Scenario de recherche 1.

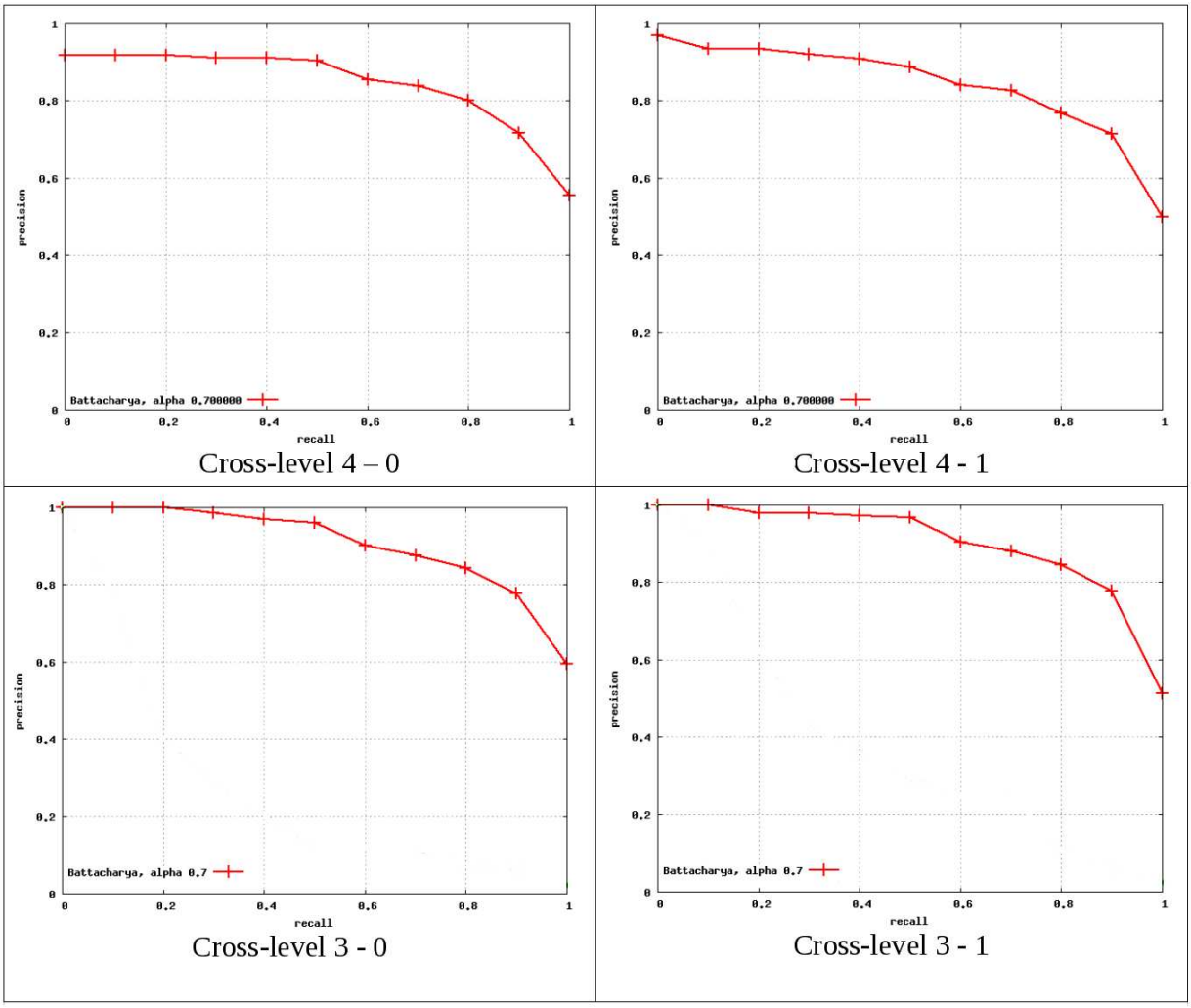


Fig. 7.11 – Courbes de Rappel-Précision interpolées moyennes pour les recherches à niveaux croisés de la BR (requête) vers la HR (BD) sur les masques extraits automatiquement. Scenario de recherche 2.

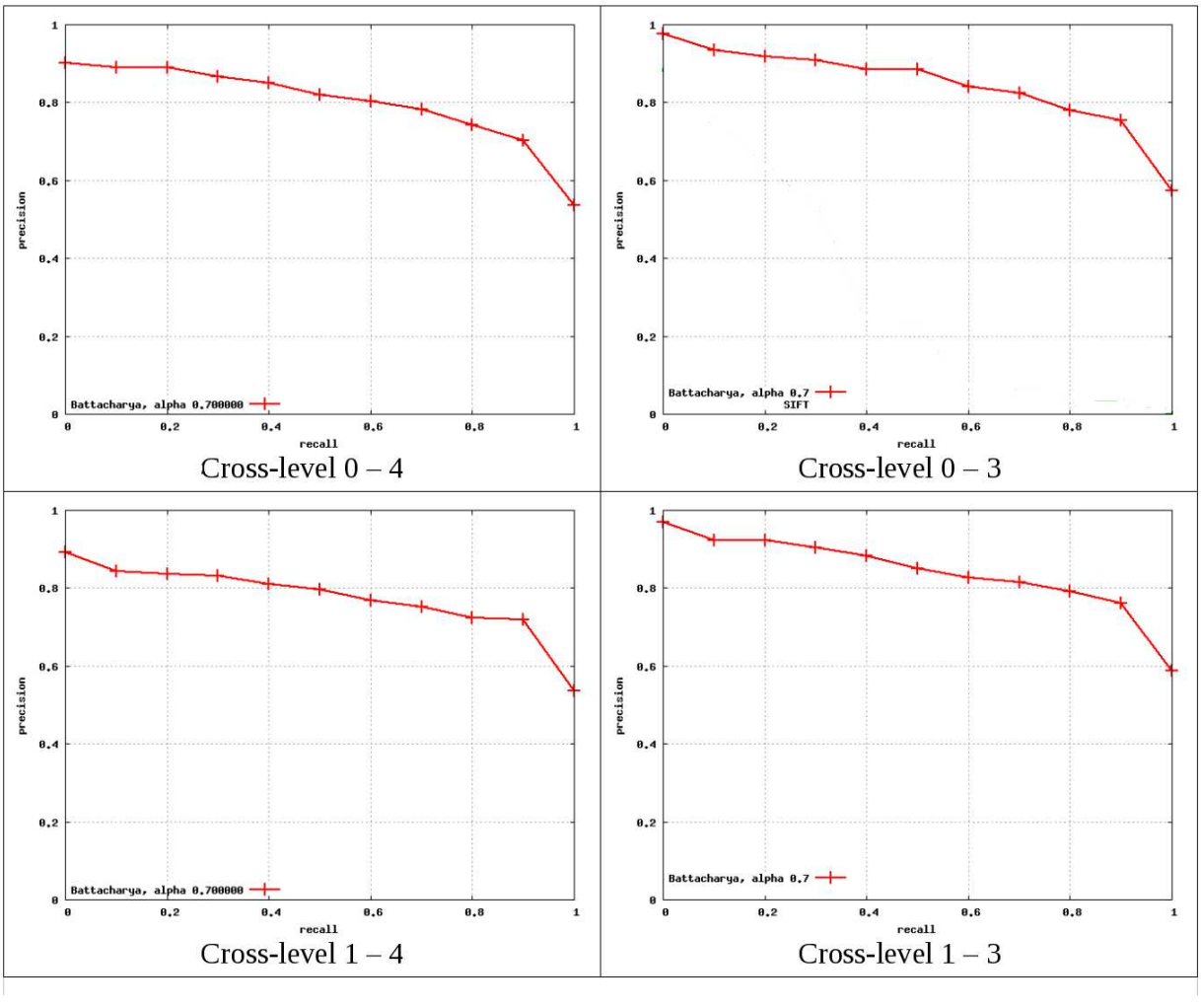


Fig. 7.12 – Courbes de Rappel-Précision interpolées moyennes pour les recherches à niveaux croisés de la HR (requête) vers la BR (BD) sur les masques extraits automatiquement. Scenario de recherche 1.

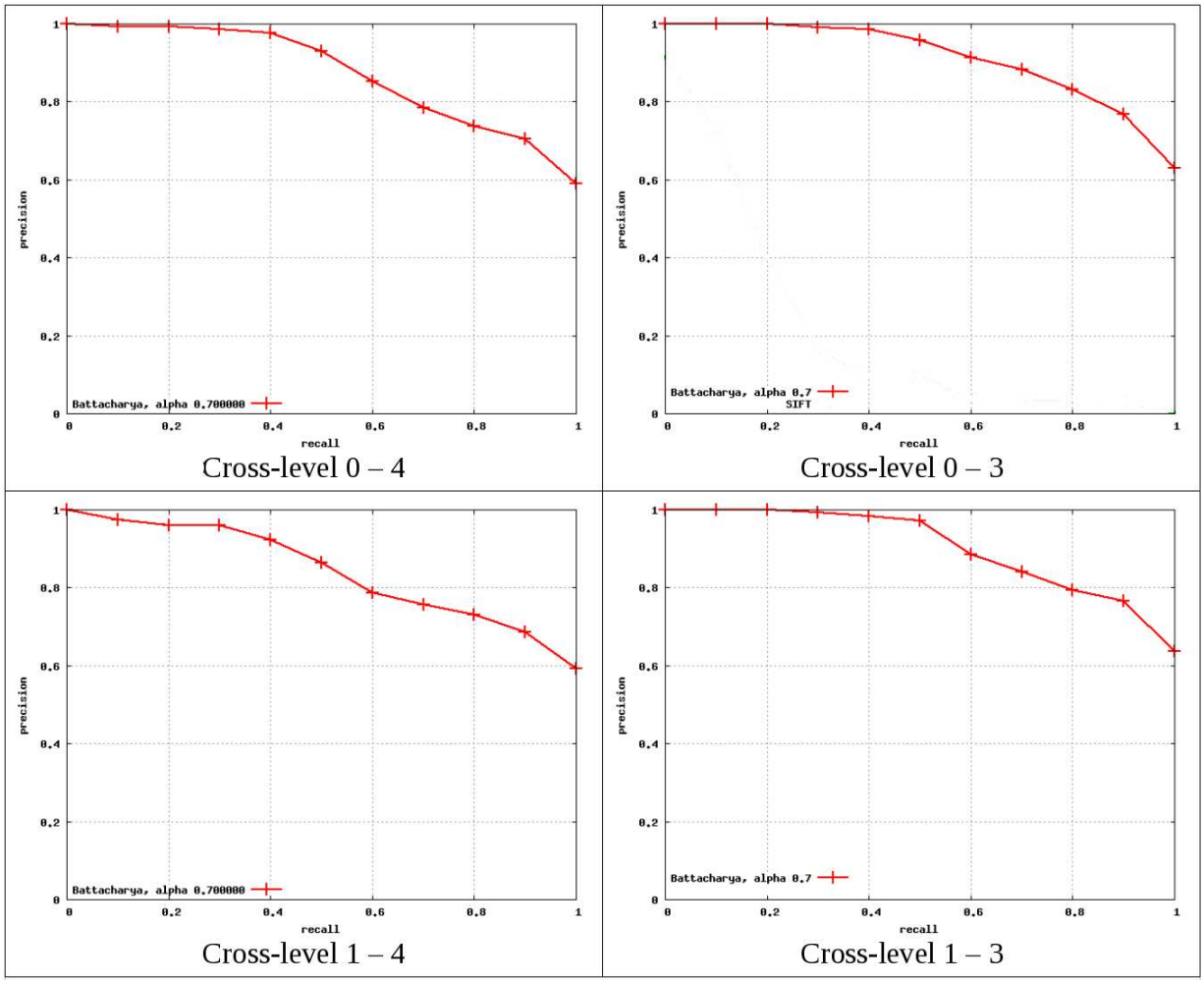


Fig. 7.13 – Courbes de Rappel-Précision interpolées moyennes pour les recherches à niveaux croisés de la HR (requête) vers la BR (BD) sur les masques extraits automatiquement. Scenario de recherche 2.

7.3.4 Comparaison avec un descripteur local basé objet et points SIFT

La comparaison des performances de notre descripteur statistique scalable avec les descripteurs de l'état de l'art (SIFT) a été possible grâce au travail collaboratif avec l'IRISA/INRIA Centre de Rennes dans le cadre du projet ICOS-HD.

7.3.4.1 Définition du descripteur local basé objet et points SIFT

Come nous l'avons souligné dans l'état de l'art en indexation vidéo, l'axe de recherche le plus populaire consiste en l'utilisation des points SIFT. Nous avons donc pensé à utiliser un descripteur local pour décrire les coefficients d'ondelettes des objets. Pour rester dans le cadre de notre schéma d'indexation par objets, les masques d'objet O_t^k sont utilisés pour déterminer la zone de détection des points SIFT proposés par Lowe [Low04]. Pour prendre en compte les contours des objets pour la détection des points caractéristiques, les masques O_t^k sont dilatés comme cela a été fait dans le même but par [Gar07]. Les régions de support pour le calcul des points SIFT sont donnés par l'algorithme MESR [Mat02]. Il calcule des régions affines covariantes qui consistent en un groupement de pixels adjacents ayant des niveaux de gris qui sont distincts de ceux des pixels environnants. Les régions résultantes sont des ellipses qui sont ajustées autour de chaque région. L'allure et l'orientation des ellipses sont une mesure de la transformation affine appliquée à l'image. Les régions sont ensuite normalisées en appliquant une transformation affine conduisant à un cercle normalisé, résultant en une invariance à l'échelle. Les régions MSER sont extraites sur toutes les images et seules les régions partageant au moins 20% de leur surface avec le masque d'objet sont conservées. Les descripteurs SIFT sont ensuite calculés pour chacun des centres des régions MSER normalisées ainsi sélectionnées. Ce processus est répété pour toutes les sous-bandes LL^k, $k = K, ..., 0$ de la pyramide d'ondelettes, ce qui donne un ensemble de descripteurs locaux pour chaque objet à chaque niveau de la pyramide. La mesure utilisée pour mettre en correspondance deux points SIFT pris dans deux objets différents est la distance euclidienne ; les descripteurs sont appariés si la distance est plus petite qu'un seuil σ. La similarité entre deux objets est donnée par un algorithme standard de vote qui comptent le nombre de descripteurs qui ont été appariés, sous l'hypothèse d'unicité de l'appariement (un descripteur peut n'être mis en correspondance qu'une seule fois).

7.3.4.2 Comparaison descripteur d'objet local et global

Nous comparons ici les courbes de rappel précision interpolées moyennes obtenues avec les deux types de descripteur. Pour cette expérimentation, nous avons utilisé les masques d'objets extraits manuellement afin de ne pas observer l'impact de la qualité de la segmentation.

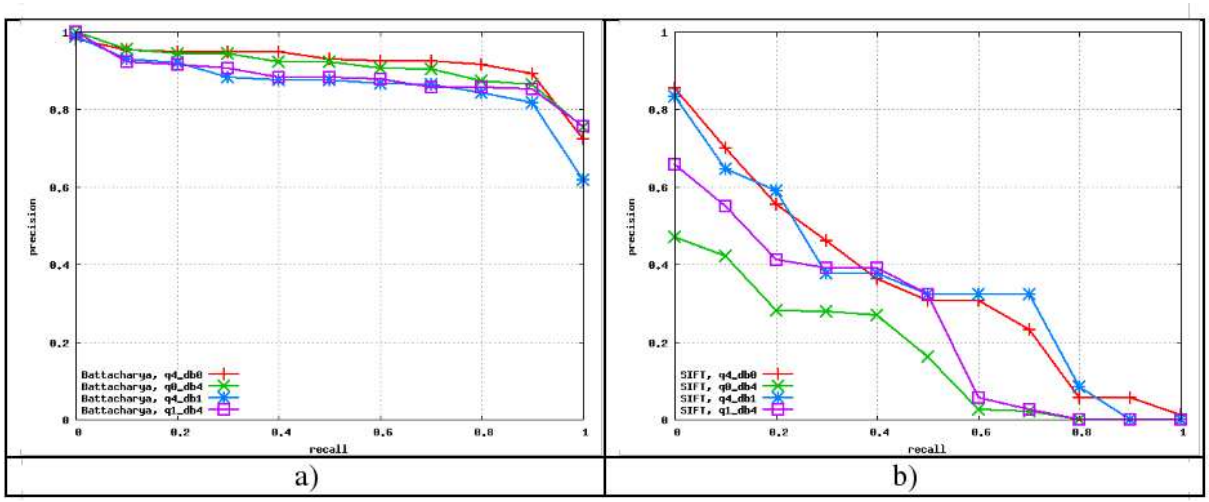


Fig. 7.14 – Courbes de Rappel-Précision interpolées moyennes pour des requêtes à **niveaux croisés** avec le descripteur d'objet global basé histogrammes (a) et le descripteur d'objet local basé SIFT (b). Les masques d'objets sont extraits **manuellement**. Scenario de requête 1.

Les résultats obtenus pour les deux descripteurs pour des recherches en mono-niveau sont données dans la figure 7.15 et pour des recherches à niveaux croisés dans la figure 7.14. Dans tous les cas, on constate qu'aux très bas niveaux de rappel, les points SIFT donnent une meilleure précision Dans tous les autres cas, le descripteur basé histogramme est systématiquement plus performant.
Des résultats similaires sont obtenus pour le scénario 2, nous ne les développerons pas ici. Le même comportement est observé pour les masques extraits automatiquement avec bien évidemment une nette dégradation due au bruit de segmentation pour les requêtes mono-niveau (Figure 7.16) et à niveaux croisés (Figure 7.17).

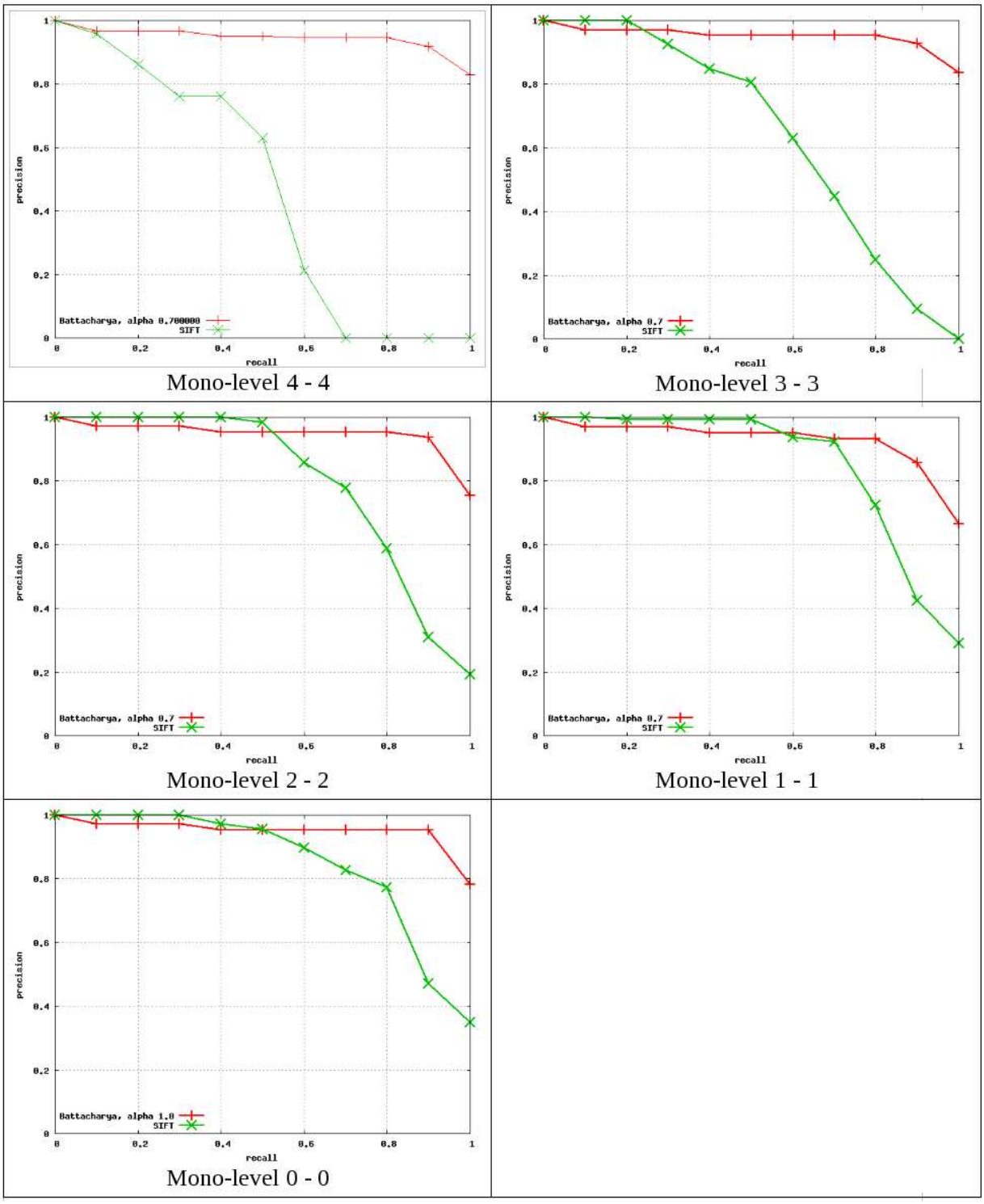
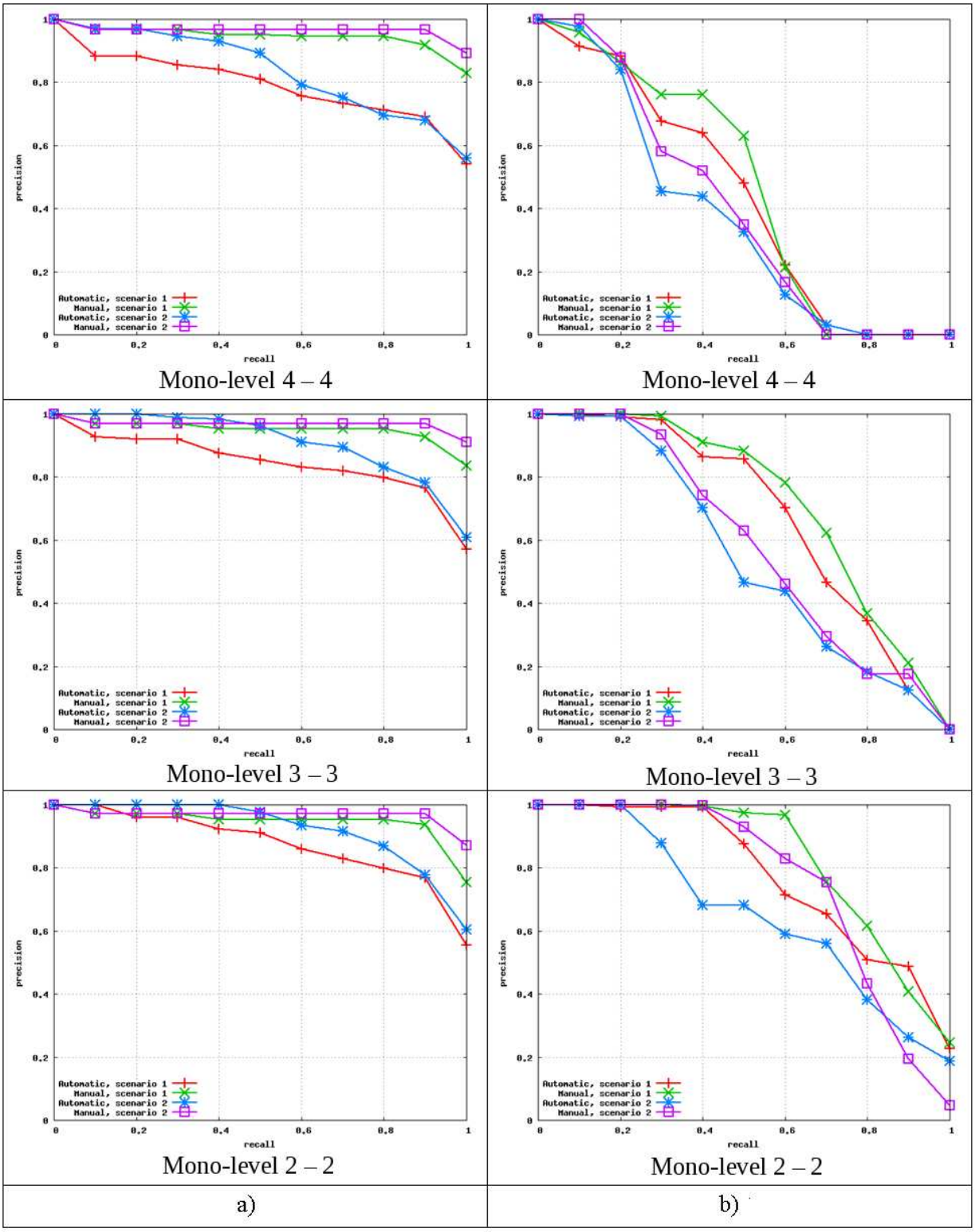


Fig. 7.15 – Courbes de Rappel-Précision interpolées moyennes pour des requêtes **mono-niveaux** avec le descripteur d'objet global basé histogramme et le descripteur d'objet local basé SIFT. Les masques d'objets sont extraits **manuellement**. Scenario de requête 1.



Mono-level 4 – 4 | Mono-level 4 – 4

Mono-level 3 – 3 | Mono-level 3 – 3

Mono-level 2 – 2 | Mono-level 2 – 2

a) | b)

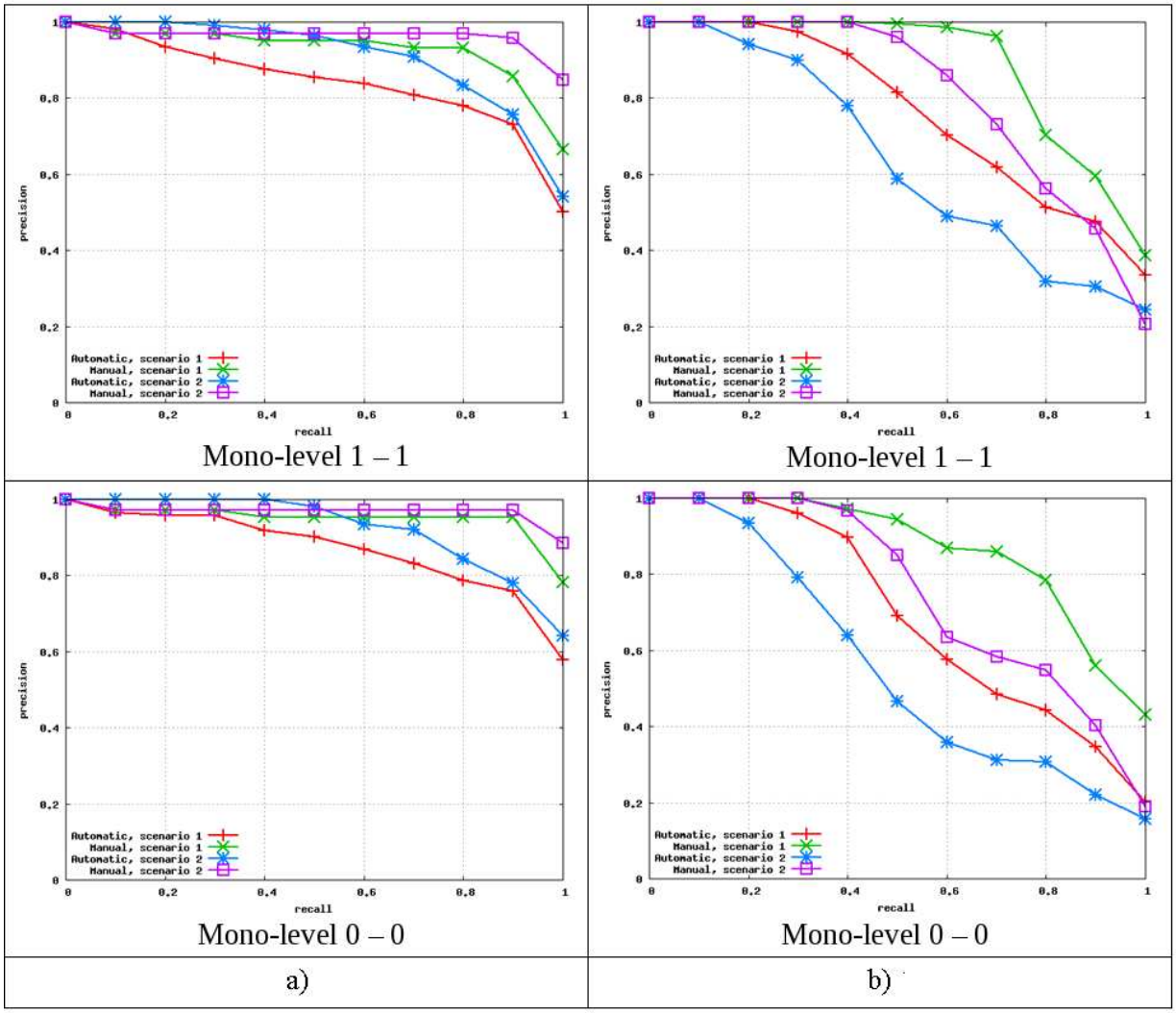
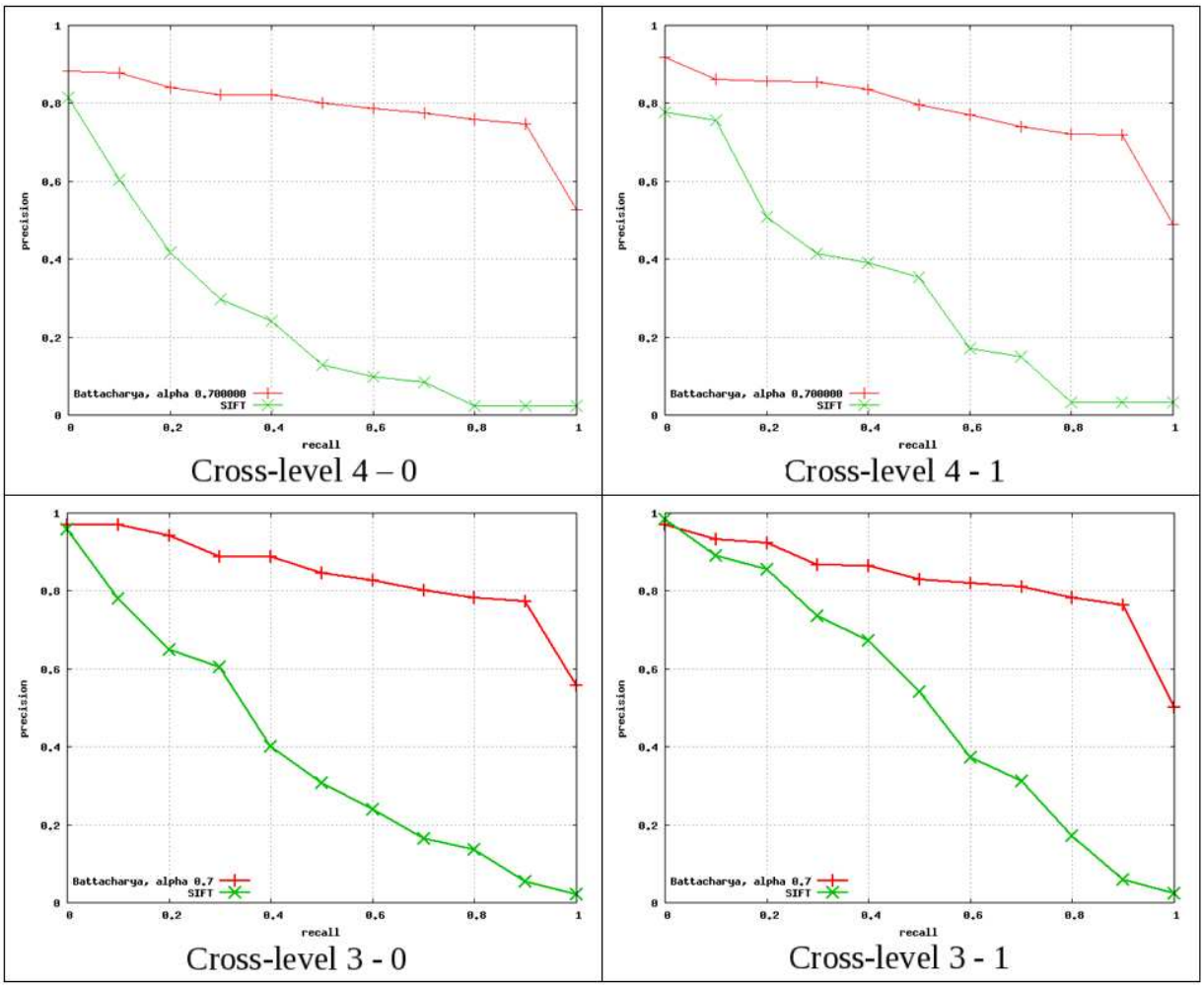


Fig. 7.16 – Courbes de Rappel-Précision interpolées moyennes pour des requêtes **mono-niveaux** avec le descripteur d'objet global basé histogramme (a) et le descripteur d'objet local basé SIFT (b). Les masques d'objets sont extraits **manuellement et automatiquement**.


precision
recall
Battacharya, alpha 0.700000
SIFT
Cross-level 4 – 0
Cross-level 4 - 1
Battacharya, alpha 0.7
Cross-level 3 - 0
Cross-level 3 - 1

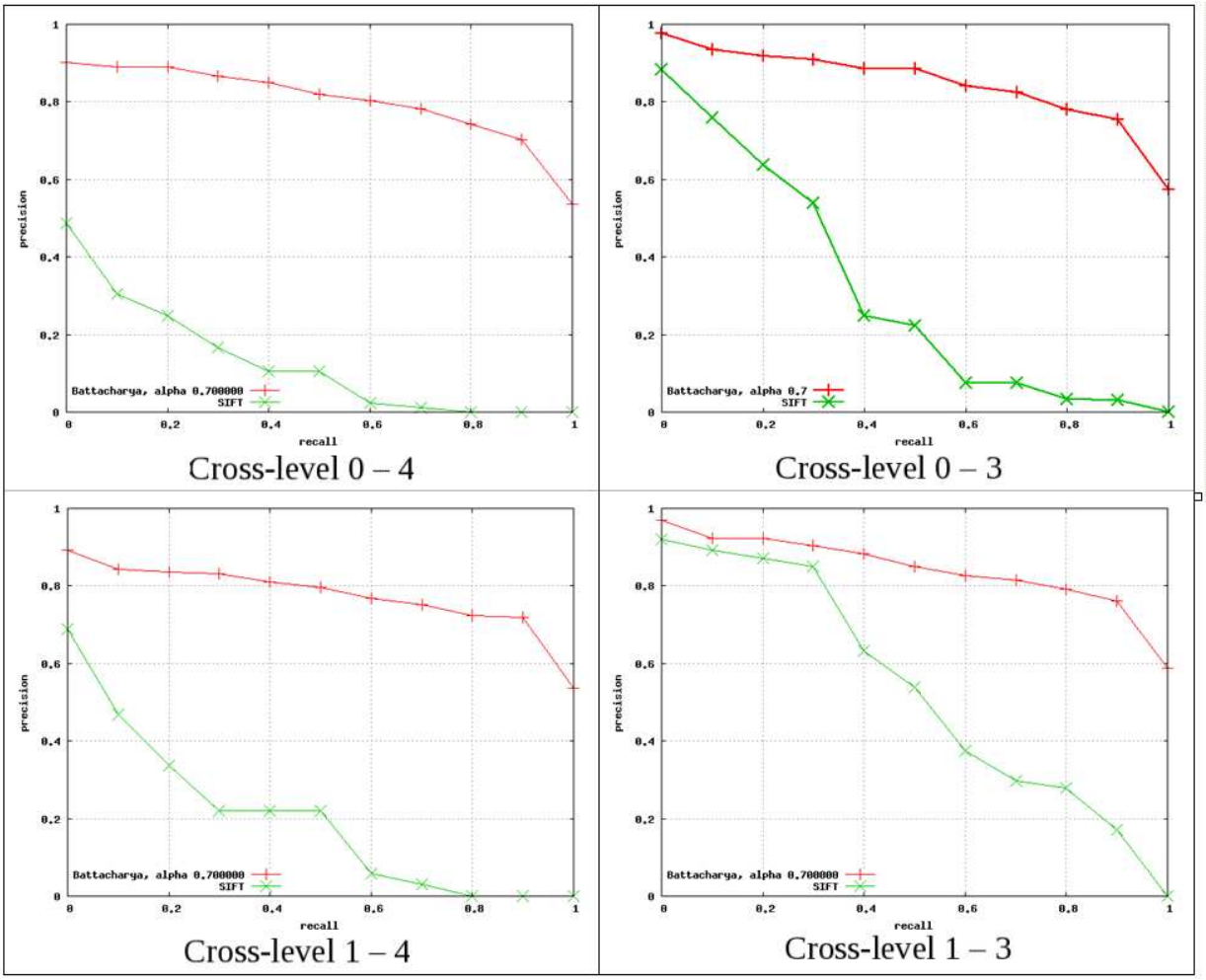


Fig. 7.17 – Courbes de Rappel-Précision interpolées moyennes pour des requêtes à **niveaux croisés** avec le descripteur d'objet global basé histogrammes (a) et le descripteur d'objet local basé SIFT (b). Les masques d'objets sont extraits **automatiquement**. Scenario de requête 1.

Chapitre 8

Conclusion et perspectives

Dans ce manuscrit, nous avons proposé une méthode complète d'indexation des vidéos de HD orientée objet et scalable. Cette méthode est définie à partir des données du flux compressé MJPEG2000. Une telle approche a nécessité que nous définissions une méthode scalable d'extraction des objets en mouvement à partir du flux compressé.

Depuis les années 1990, un travail de recherche significatif a été effectué dans le domaine de l'indexation des vidéos par le contenu. Même si le courant majeur actuel tend à la définition et l'utilisation de descripteurs locaux, l'indexation par objet reste une approche prometteuse. L'utilisation des objets permet en effet d'apporter une information sémantique de niveau moyen par rapport à des points d'intérêt qui sont de bas niveau sémantique. Utiliser une telle approche présente une difficulté supplémentaire. Il faut extraire les objets d'intérêt. Il est bien connu que ce problème est mal posé et qu'il est difficile de le résoudre. Cependant l'information apportée est d'une importance majeure. C'est pourquoi nous avons décidé de travailler sur ce type d'indexation. De plus, la distinction des objets d'intérêt est de plus en plus présente au niveau des standards de compression. Ainsi, les nouveaux standards tels que H264/SVC et MJPEG2000 proposent de coder avec plus de précision les régions d'intérêt. Cela nous a conforté dans notre choix de travailler avec des objets d'intérêt.

De nombreuses méthodes d'indexation par les objets ont été et sont toujours proposées dans le domaine pixel. Cependant aujourd'hui très peu de vidéos de HD sont disponibles au format non compressé. Pour appliquer de telles méthodes, il faut procéder au décodage complet de la vidéo. Vu la quantité des données à traiter, spécifiquement dans le cas HD, ce type d'opération est très coûteux en temps de calcul. Aussi des techniques d'indexation dans le domaine compressé ont fait leur apparition. La philosophie d'une telle approche est décrite par le Paradigme de l'Indexation Primaire proposé au LaBRI. Le principe d'une telle démarche est d'utiliser les données du flux compressé pour faire de l'indexation. La

difficulté réside dans le fait que les transformations utilisées ont été optimisées dans un objectif de codage et non d'analyse. Dans le cadre des vidéos de HD, de nouveaux standards de compression dits scalables sont apparus. Il s'agit d'une nouvelle façon de penser le stockage et la transmission des vidéos. L'indexation de flux compressés par de tels standards est un domaine de recherche très jeune. Nous avons choisi de travailler avec des vidéos encodées en MJPEG2000.

L'originalité de ce standard réside entre autres dans l'utilisation de la TOD pour décorréler les données. Cela permet à MJPEG2000 d'offrir un meilleur compromis débit/distorsion que le standard H264 pour les vidéos de très grandes dimensions spatiales, typiquement la HD. Ainsi, même si H264 semble être actuellement le standard de compression scalable le plus populaire pour les applications commerciales grand public, MJPEG2000 offre des perspectives intéressantes et est déjà utilisé dans des applications professionnelles telles que le cinéma numérique et l'archivage des contenus audiovisuels. C'est pourquoi nous pensons qu'il est primordial de développer des solutions d'indexation pour ce standard. A notre connaissance, il n'existe pas d'autres travaux proposant l'indexation scalable orientée objet des vidéos à partir des données du flux compressé JPEG2000.

L'approche que nous avons proposée se décompose en deux parties. Dans un premier temps, nous avons défini une méthode d'extraction des objets en mouvement dans le domaine compressé. Ce domaine compressé, nous l'avons considéré comme étant le domaine de la TOD utilisée dans MJPEG2000. Dans un deuxième temps, un descripteur global sur ces objets a été défini en utilisant toujours les données du domaine compressé. La TOD est une transformée très riche car elle apporte une information localisée à la fois en espace et en fréquence. Ainsi, des informations sur la localisation des contours des objets et les textures sont directement définies par les coefficients obtenus par la TOD. L'originalité de notre démarche a consisté en l'utilisation de cette information.

La méthode d'extraction des objets a suivi un processus classique de segmentation spatio-temporelle. Une segmentation en mouvement a été combinée à une segmentation couleur. Le résultat de cette segmentation a été projeté et affiné le long de la pyramide de multirésolution du domaine de la TOD. Notre contribution a porté sur l'amélioration des performances des algorithmes par l'utilisation directe des informations de HF apportées par la TOD.

Ainsi, la segmentation en mouvement nécessite de faire une estimation de mouvement global. La technique que nous avons utilisée passe par l'estimation d'un modèle à partir de vecteurs de mouvement obtenus par la méthode classique de MCB. Le modèle est estimé par un schéma robuste permettant de rejeter les valeurs aberrantes. Afin d'améliorer les performances de l'algorithme, nous avons proposé d'utiliser l'activité des coefficients de HF pour définir des vecteurs non fiables. Ces vecteurs ont été considérés a priori comme des valeurs aberrantes et ne faussaient plus l'estimation du modèle global.

La segmentation couleur des trames à BR a suivi un schéma morphologique classique. Notre contribution a été apportée au moment de la projection aux niveaux supérieurs des résultats de l'extraction d'objet. Nous avons proposé un ajustement fin des contours des objets extraits grâce aux coefficients de HF. Les techniques classiques utilisent un a priori de compacité de l'objet et ne peuvent suivre les formes d'objets complexes. Nous avons proposé d'introduire la connaissance a priori sur la forme et la position des contours apportée par les coefficients de HF de la TOD. Nous avons trouvé qu'une modélisation markovienne permettait de rendre compte de façon pertinente de cette information. Dans le cadre de l'extraction markovienne, nous avons introduit un nouveau potentiel exploitant l'information HF et servant de barrière de propagation des étiquettes à travers les contours.
L'extraction des objets proposée est scalable dans la mesure où : i) elle s'adapte à la nature d'un flux compressé scalable et ii) elle fournit des objets en multirésolution ayant des contours précis et proche de la perception humaine.

Nous avons défini ensuite un descripteur scalable à partir de la représentation multirésolution des objets obtenus. Nous avons choisi d'utiliser un descripteur global sur l'objet afin d'être robuste aux erreurs inévitables de segmentation. De plus, l'utilisation d'un descripteur global suit la philosophie du Paradigme de l'Indexation Primaire. En effet, les caractéristiques globales sont particulièrement adaptées à la reconnaissance d'objets, et plus encore si un prototype est disponible, ce qui est le cas dans la recherche par le contenu de vidéos. Le descripteur que nous avons proposé est fondé sur les distributions statistiques des coefficients d'ondelettes des sous-bandes de BF et HF. Ces distributions sont caractérisées par deux histogrammes (un pour la sous-bande LL et un pour les sous-bandes HF) à chaque niveau de la pyramide. L'utilisation des coefficients de HF est une façon de décrire la texture des objets. Conformément à notre intuition, les expériences que nous avons menées ont montré que l'utilisation de l'information de HF assure une meilleure robustesse par rapport à l'utilisation de la BF uniquement. D'autre part, nos expériences ont montré la robustesse de notre descripteur global au bruit de segmentation par rapport à un descripteur local de type SIFT.
Le descripteur proposé est scalable et répond à des requêtes par similarité scalables. Nos expérimentations ont principalement porté sur la démonstration de la scalabilité en résolution. Ce descripteur permet d'effectuer des recherches de vidéos par le contenu scalables, en mono-niveau ou à niveaux croisés, de façon plus équilibrée que SIFT.

Ce travail offre de nombreuses perspectives de poursuite de recherche. D'une part, la méthode d'extraction des objets peut être améliorée en intégrant une mémoire du résultat de l'extraction dans les trames précédentes. Nous étudions la possibilité d'utiliser également une approche markovienne. Ce type d'approche intègre en effet la philosophie de l'ajustement markovien utilisant les coefficients de HF que nous avons proposé. Dans une

perspective plus éloignée, la méthode d'extraction d'objet présentée dans ce manuscrit peut être utilisée pour déterminer des ROI dans les vidéos déjà encodées en JPEG2000. La région d'intérêt peut être ainsi calculée à la volée lors d'opérations de transcodage de la vidéo. La nouvelle vidéo ainsi codée contient alors la représentation de l'objet sur laquelle il est possible de définir notre descripteur.
Une autre perspective de notre travail est donc d'étudier la scalabilité de notre descripteur dans d'autres conditions telles que la scalabilité en qualité et la scalabilté en temps.
Une étape supplémentaire dans la scalabilité de la requête peut être considérée. Il s'agirait de laisser à l'utilisateur la liberté de définir la similarité qu'il envisage (par exemple : exactement similaire, simplement similaire ou moins précis comme similaire en forme mais pas en couleur et texture ou de même contenu couleur mais pas forcément de texture ou de forme). Notre descripteur combine deux caractéristiques couleur et texture qui peuvent être pondérées et s'adapteraient à ce genre de requête.
D'autre part, la nature scalable du descripteur permet d'envisager son intégration à un flux compressé scalable.
Enfin, nous envisageons d'utiliser notre descripteur pour réaliser d'autres tâches d'indexation telles que la création de résumés vidéos. Dans la perspective du projet ICOS-HD, l'étude de la coopération entre plusieurs descripteurs est aussi une perpective à notre travail.

Annexe A

Logiciels codant le standard JPEG2000

Il existe plusieurs implémentations du standard JPEG2000. L'objet de cette annexe est de citer quelques uns de ces logiciels dont le code source est disponible et ré-utilisable à des fins de recherche. Ces logiciels sont récapitulés dans le tableau A.1 .

Nom	Développeurs	Site Internet	Langage	Remarques
Jasper	University of British Columbia et Image Power, Inc.	http://www.ece.ubc.ca/~mdadams/jasper	C	Référencé dans la norme [ISO04a]. Ne prend pas en charge le codage de la ROI.
JJ2000	Canon, Ecole Polytechnique Fédérale de Lausanne et Ericsson	http://JJ2000.epfl.ch	Java	Référencé dans la norme [ISO04a]. Lent.
Open JPEG2000	Laboratoire de Communication et Télédétection, Université Catholique de Louvain, Belgique	http://www.tele.ucl.ac.be/PROJECTS/OPENJPEG	C++	Issu du projet "Open JPEG".
Kakadu	David Taubman, University of New South Wales	http://www.kakadusoftware.com	C++	Commercial ©Unisearch Ltd. Encodage des ROI. Rapide.

Tab. A.1 – Exemples de logiciels codant JPEG2000 disponibles en code source.

Nous avons opté pour le logiciel Kakadu. Bien que commercial et non référencé par la norme, il est le plus abouti des logiciels présentés. C'est non seulement le plus rapide, mais

c'est aussi le seul à implémenter entièrement la partie 1 du standard. Il est aussi largement utilisé dans la littérature.

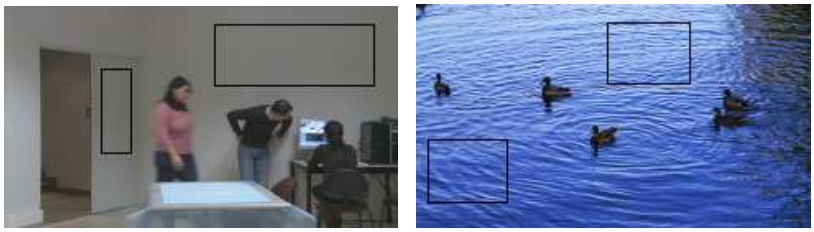
Annexe B

Détermination empirique de seuils pour l'estimation de mouvement

L'objet de cette annexe est de présenter la méthode empirique utilisée pour fixer les seuils de rejet des blocs de faible activité HF (section 4.2.3) et de qualité de reconstruction par compensation de mouvement (section 4.2.4) utilisés dans la méthode d'estimation du mouvement hiérarchique (chapitre 4).
Les seuils de rejet interviennent dans la méthode d'estimation du mouvement global à partir des valeurs estimées des vecteurs de mouvement par bloc (section 4.2.3). Cette méthode repose sur un schéma de rejet préalable des valeurs non fiables en deux étapes. La deuxième étape consiste ainsi en l'exclusion des vecteurs de mouvement des blocs ayant une faible activité HF ; la mesure de l'activité étant définie par le vecteur d'écart-types des coefficients de HF du bloc. La décision de rejet se fait alors par seuillage (4.18). Les seuils de qualité de reconstruction par compensation de mouvement interviennent dans la définition des valeurs non conformes (section 4.2.4) Nous avons proposé d'estimer si les valeurs non fiables définies précédemment sont conformes ou non conformes au modèle de mouvement global en utilisant un critère de qualité. La décision de conformité se fait par seuillage (4.19).

B.1 Seuils de rejet des blocs de faible activité HF

Afin d'établir une vérité terrain, un opérateur humain repère dans les vidéos du corpus ICOSHD des régions rectangulaires correspondant à des régions homogènes de faible activité HF (zones plates) et de forte activité HF (zones texturées). Un exemple d'une telle extraction est donné dans la figure B.1.



(a) Zones plates (b) Zones texturées

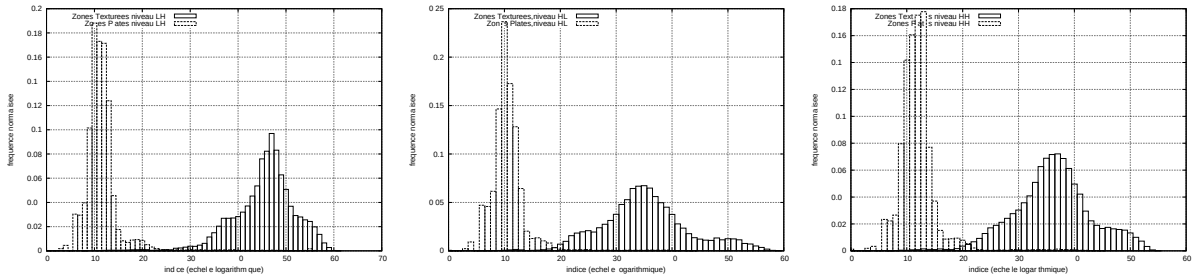
Fig. B.1 – Détermination manuelle des zones plates et texturées dans deux trames issues du corpus ICOSHD

Niveau k	Taille des blocs n^k
4	4
3	8
2	16
1	32

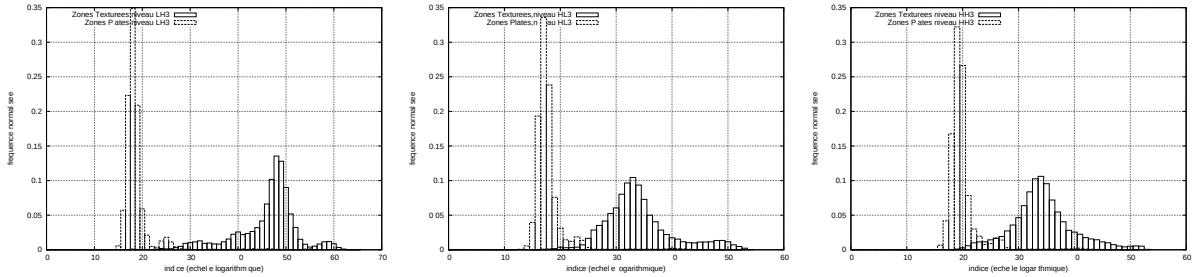
Tab. B.1 – Tailles des blocs utilisées pour l'estimation des seuils de rejet de faible activité HF en fonction du niveau de résolution dans la pyramide d'ondelettes

Dans les zones homogènes définies manuellement, des blocs de taille n^kxn^k sont découpés et le vecteur d'écart-type σ^k des coefficients de HF de chaque bloc est calculé. La taille des blocs est fonction du niveau de résolution considéré et de la taille des blocs choisie pour l'étape de MCB à BR (section 4.2.2). Ces tailles sont récapitulées dans le tableau B.1.

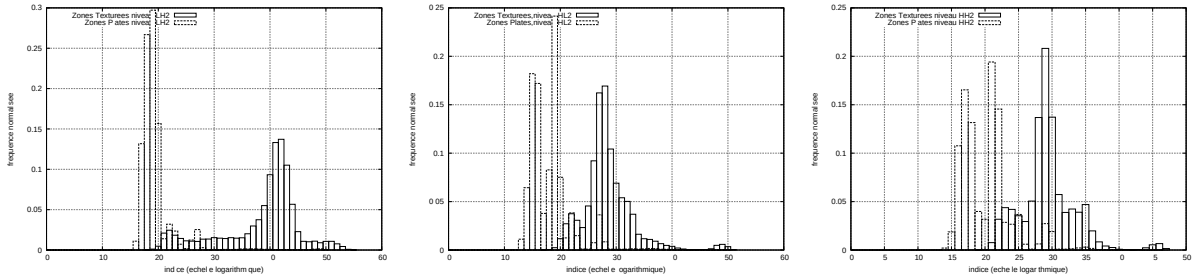
La masse d'écarts-types ainsi collectée est visualisée sous la forme d'histogrammes (figure B.2). Pour chaque type de sous-bande et chaque niveau de décomposition, les histogrammes des écarts-types des coefficients d'ondelettes pour les zones plates et pour les zones texturées sont superposés. A chaque fois, les deux histogrammes sont bien distincts. Cependant, à HR (niveau 1), l'histogramme correspondant aux zones plates devient bimodal. Cela est dû à la présence de bruit et de détails fins qui n'existent pas à BR du fait du filtrage de construction de la pyramide d'ondelettes. Le seuil de distinction est déterminé manuellement à partir de ces figures à l'endroit de séparation des deux histogrammes.
Notons que l'abscisse des histogrammes est représentée en échelle logarithmique. En effet, la gamme des valeurs de variances (de 10^{-8} à 10^{-2}) est trop étendue pour être clairement affichée en échelle linéaire. Les indices de classe d'histogramme sont donnés par la formule (B.1a). La valeur de variance associée à un indice de classe est donnée par la formule (B.1b)



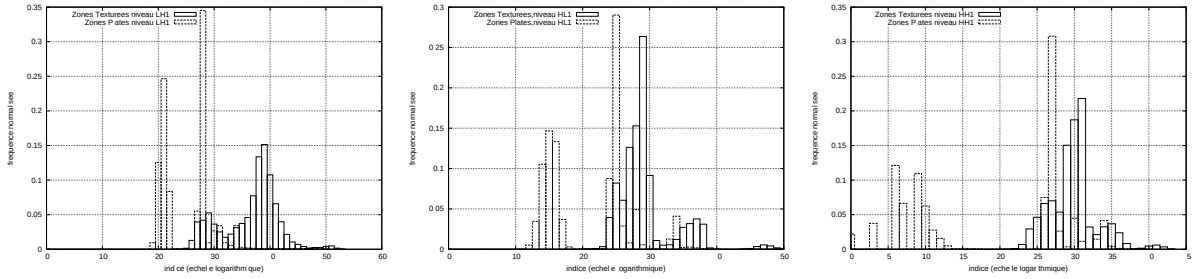
(a) Niveau 4



(b) Niveau 3



(c) Niveau 2



(d) Niveau 1

Fig. B.2 – Histogrammes des écarts-types des coefficients de HF. La colonne de gauche représente la sous-bande LH, la colonne du milieu la sous-bande HL et la colonne de droite la sous-bande HH. Sur chaque figure sont superposés les histogrammes des zones plates et des zones texturées.

suivante

$$ind = \lfloor log_{10} \lfloor \frac{\sigma}{10^{-8}} \rfloor * 10 \rfloor \tag{B.1a}$$

$$\sigma = 10^{ind/10} * 10^{-8} \tag{B.1b}$$

où σ est la valeur de variance, ind est l'indice de classe de l'histogramme, 10^{-8} est le pas de quantification utilisé dans l'histogramme et 10 le pas de quantification pour l'affichage en logarithme. Les seuils déterminés sont récapitulés dans le tableau 4.1. Rappelons que nous travaillons sur les coefficients d'ondelettes extraits du flux compressé juste après la déquantification. Ces coefficients ont donc été calculés à partir d'images à valeurs normalisées dans $[-\frac{1}{2}, \frac{1}{2}[$, ce qui explique l'ordre de grandeur des variances calculées.

B.2 Seuils de qualité de reconstruction

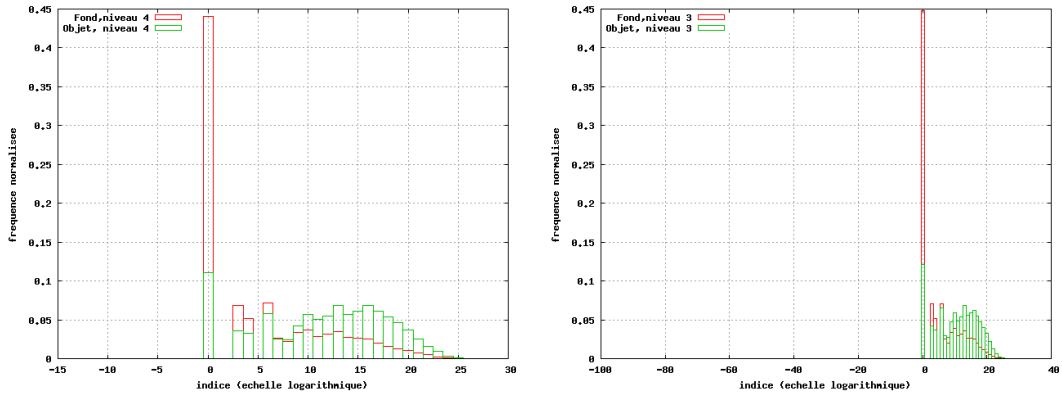
La vérité terrain donnant les masques des objets est à notre disposition. Le calcul de l'estimation de mouvement est conduit comme décrit dans le chapitre 4. La seule différence se situe pour le calcul du rejet que nous sommes en train d'évaluer. Dans le cadre de cette annexe, les masques d'objets manuels sont connus. Nous évaluons donc la quantité de qualité de reconstruction $|MAD_{B_i}(v) - MAD_{B_i}(v(\theta))|$ pour tous les blocs. Grâce à la vérité terrain, nous savons si le bloc considéré décrit un objet en mouvement ou s'il décrit le fond. La valeur de qualité de reconstruction est alors ajoutée dans l'histogramme correspondant. La figure B.3 représente ces histogrammes à chaque niveau de la pyramide. A tous les niveaux, la valeur de la qualité de reconstruction est proche de zéro pour le fond, ce qui signifie que l'on peut approximer le vecteur de déplacement par le vecteur obtenu par calcul du mouvement global. Quelques valeurs de reconstruction pour le fond sont non nulles et se confondent avec l'objet. Tous les blocs du fond (mais une grande partie) ne pourront pas être éliminés par ce traitement. La valeur commune de seuil à tous les niveaux est de 10^{-3}.

Notons que, là encore, l'abscisse des histogrammes est représentée en échelle logarithmique. En effet, la gamme des valeurs de variances est trop étendue pour être clairement affichée en échelle linéaire. Les indices de classe d'histogramme sont donnés par la formule (B.2a). La valeur de différence de MAD associée à un indice de classe est donnée par la formule (B.2b) suivante

$$ind = \lfloor log_{10} \lfloor \frac{mad}{10^{-3}} \rfloor * 10 \rfloor \tag{B.2a}$$

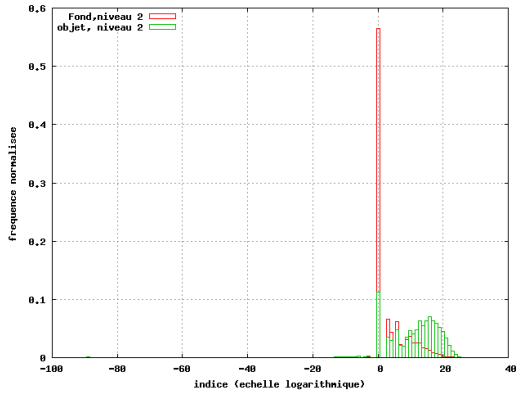
$$mad = 10^{ind/10} * 10^{-3} \tag{B.2b}$$

où $mad = MAD_{B_i}(v) - MAD_{B_i}(v(\theta))$ est la valeur de qualité de reconstruction, ind est l'indice de classe de l'histogramme, 10^{-3} est le pas de quantification utilisé dans l'histogramme et 10 le pas de quantification pour l'affichage en logarithme.

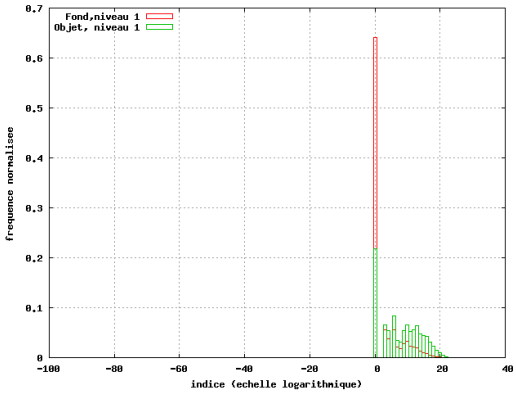


(a) Niveau 4

(b) Niveau 3



(c) Niveau 2



(d) Niveau 1

Fig. B.3 – Histogrammes des valeurs de qualité de reconstruction par le vecteur de mouvement global sur l'objet et sur le fond pour chaque niveau de résolution de la pyramide d'ondelettes.

Bibliographie

[Ada09] N. Adami, A. Boschetti, R. Leonardi et P. Migliorati. Embedded indexing in scalable video coding. Dans *7th International Workshop on Content-based Multimedia Indexing (CBMI)*, p. 101–106. Chania, Crète, 3-5 juin 2009.

[Akh05] F. AkhlaghianTab. *Multiresolution Scalable Image and Video Segmentation*. Thèse de doctorat, Université de Wollongong, New South Wales, Australie, 2005.

[Alb01] E. Albuz, E. Kocalar et A. Khokhar. Scalable color image indexing and retrieval using vector wavelets. *IEEE Transactions on Knowledge and Data Engineering*, 13(5) :p. 851–861, Septembre/Octobre 2001.

[Ant92] M. Antonini, M. Barlaud, P. Mathieu et I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2) :p. 205–220, Avril 1992.

[Aug05] B. Augereau, B. Tremblais et C. Fernandez-Maloigne. Vectorial computation of the optical flow in color image sequences. Dans *13th Color Imaging Conference*, p. 130–134. Scottsdale, Arizona, Etats-Unis, 7-11 November 2005.

[Bag07] A.D. Bagdanov, L. Ballan, M. Bertini et A. DelBimbo. Trademark matching and retrieval in sports video databases. Dans *Workshop on Multimedia Information Retrieval*, p. 79–86. Augsburg, Allemagne, 24-29 Septembre 2007.

[Bar02] Michel Barlaud et Claude Labit, réds. *Compression et Codage des images et des vidéos*. Traité IC2. Hermes, Lavoisier, 2002.

[Bay06] H. Bay, T. Tuytelaars et L. Van Gool. SURF : Speeded Up Robust Features. *Lecture Notes in Computer Science*, 3954 :p. 404–417, 2006.

[Ben01] J. BenoisPineau, W. Dupuy et D. Barba. Recovering of visual scenarios in movies by motion analysis and grouping spatio-temporal colour signatures of video shots. Dans *2nd EUSFLAT Conference*, p. 385–389. Leicester, UK, 5-7 Septembre 2001.

[Ben03] J. BenoisPineau, W. Dupuy et D. Barba. Outils de structuration des documents vidéo en vue d'indexation basée sur une approche du signal 1D. *RSTI/ série Techniques et Sciences Informatiques, Recherche d'informations par le contenu visuel*, 22(9) :p. 1167–1200, 2003.

[Ber88] M. Bertero, T.A. Poggio et V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8) :p. 869–889, Août 1988.

[Ber95] P. Bertolino. *Contribution des pyramides irrégulières en segmentation d'images multirésolution.* Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG, 1995.

[Bes86] J. Besag. On the statistical analysis of dirty pictures. *Journal of The Royal Statistical Society, Series B*, 48 :p. 259–302, 1986.

[Bha43] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc*, 35 :p. 99–109, 1943.

[Bha98] D. Bhate et S. Nayar. Ordinal measures for image correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4) :p. 415–423, 1998.

[Bla89] A. Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(1), janvier 1989.

[Bla92] M.J. Black. Combining intensity and motion for incremental segmentation and tracking over long image sequences. *Proceedings of the Second European Conference on Computer Vision, Lecture notes in computer science*, 588 :p. 485 – 493, 1992.

[Bol09] S. Boltz, E. Debreuve et M. Barlaud. High-dimensional statistical measure for region-of-interest tracking. *IEEE Trans. On Image Processing*, 18(6) :p. 1266–1283, juin 2009.

[BP98] J. Benois-Pineau, F. Morier, D. Barba et H. Sanson. Hierachical segmentation of video sequences for content manipulation and adaptive coding. *Signal Processing, elsevier*, 66(2), 1998.

[Bro08] O. Brouard, F. Delannay, V. Ricordel et D. Barba. Spatio-temporal segmentation and regions tracking of high definition video sequences based on a markov random field model. Dans *15th IEEE International Conference on Image Processing (ICIP)*, p. 1552–1555. San Diego, CA, 12-15 Octobre 2008.

[Bru02] E. Bruno et D. Pellerin. Video structuring, indexing and retrieval based on global motion wavelet coefficients. Dans *16th International Conference on Pattern Recognition (ICPR)*, t. 3, p. 302–387. Quebec, Canada, 11-15 Août 2002.

[Buc99] R.W. Buccigrossi et E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, 8(12) :p. 1688–1701, Decembre 1999.

[Cha98] S.F. Chang, W. Chen, H.J. Horace, H. Sundaram et D. Zhong. A fully automated content based video search engine supporting spatio-temporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5) :p. 602–615, 1998.

[Che92] K. Chehdi. *Traitement et Analyse d'images en vue de la reconnaissance des formes*. Thèse de doctorat, HDR, Université de Rennes 1, France, mai 1992.

[Che06] L. Chen et F.W.M. Stentiford. Video sequence matching based on temporal ordinal measurement. Rap. tech. 1, UCL Adastral, 2006.

[Che07] F. Chevalier, M. Delest et J-Ph. Domenger. A heuristic for the retrieval of objects in low resolution video. Dans *Workshop on Content-Based Multimedia Indexing (CBMI)*, p. 144–151. Bordeaux, France, 25-27 Juin 2007.

[Cho90] P.B. Chou et C.M Brown. The theory and practice of bayesian image labeling. *International Journal of Computer Vision*, 4 :p. 185–210, 1990.

[Col98] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tommiver, E. Enomoto, O. Hasegawa, P. Burt et L. Wixson. A system for video surveillance and monitoring : VSAM final report. Rap. tech. CMU-RI-TR-00-12, Carnegie Mellon University, mai 1998.

[Cos06] B. Coskun, B. Sankur et N. Mamon. Spatio-temporal transform-based video hashing. *IEEE Transactions on Multimedia*, 8(6) :p. 1190–1208, 2006.

[Cou99] F. Coudert, J. Benois-Pineau et D. Barba. Binkey a system for video content analysis on the fly. Dans *International Conference on Multimedia Computing and Systems*, p. 679–684. Florence, Italie, 7-11 Juin 1999.

[Dau92] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[DCIL] (DCI) Digital Cinema Initiatives LLC. Specifications. http ://www.dcimovies.com. Consulté le 30/04/2008.

[Del94] P. Delagnes, J. Benois et D. Barba. Adjustable polygons : a novel active contour model for object tracking on complex background. *Journal on Communications*, 45 :p. 83–85, Juillet/Août 1994.

[Del95] P. Delagnes. Détection de défauts rectilignes f=dans des images de chaussées par une approche markovienne. Dans *15ème colloque GRETSI*, p. 1249–1252. Juan-les-pins, France, septembre 1995.

[Del04] M. Delest, A. Don et J. Benois-Pineau. Intuitive color-based visualization of multimedia content as large graphs. Dans *Conference on Visualization and Data Analysis*, t. 5295, p. 65–74. San Jose, CA, Etats-Unis, 19-20 Janvier 2004.

[DG09] E. Drelie Gelasca et T. Ebrahimi. On evaluating video object segmentation quality : A perceptually driven objective metric. *IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING*, 3 :p. 319–335, 2009.

[Dub89] R.C. Dubes et A.K. Jain. Random field in image analysis. *Journal of Applied Statistics*, 16(2), 1989.

[Dum09] E. Dumont et B. Merialdo. Rushes video parsing using video sequence alignment. Dans *7th International Workshop on Content-Based Multimedia Indexing (CBMI)*, p. 44–49. Chania, Grèce, 3-5 juin 2009.

[Dur01] M. Durick et J. Benois-Pineau. Robust global motion characterisation for video indexing based on mepg2 optical flow. Dans *CBMI2001*, p. 57–64. Brescia, Italy, September 2001.

[ElS03] M.A. ElSaban et B.S. Manjunath. Video region segmentation by spatio-temporal watersheds. Dans *International Conference on Image Processing (ICIP)*, t. 1, p. 349–352. Barcelone, Espagne, 14-17 Septembre 2003.

[For07] P-E. Forssen et D. Lowe. Shape descriptor for maximally stable extremal regions. Dans *International Conference on Computer Vision (ICCV)*, p. 1–8. Rio de Janeiro, Brésil, 14-21 Octobre 2007.

[Gar95] U. Gargi, S. Oswald, D. Kosiba, S. Devadiga et R. Kasturi. Evaluation of video sequence indexing and hierarchical video indexing. *Proceedings of SPIE, Storage and Retrieval in Image and Video Databases III*, 2420 :p. 144–151, 1995.

[Gar07] V. Garcia, S. Boltz, E. Debreuve et M. Barlaud. Outer-layer based tracking using entropy as a similarity measure. Dans *IEEE international Conference on Image Processing ICIP*, t. 6, p. 309–312. San Antonio, Texas, 2007.

[Gem84] D. Geman et S. Geman. stochastic relaxations, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 6 :p. 721–741, 1984.

[Gil06] Ian Gilmour et R. Justin Dáliva. Lossless video compression for archives : Motion JPEG2k and other options. Rap. tech., Media matters, llc, janvier 2006.

[Gou06] M. Gouiffès, C. Collewet, Ch. Fernandez-Maloigne et A. Trémeau. Feature points tracking : Robustness to specular highlights and lighting changes. *Lecture Notes in Computer Science*, 3954 :p. 82–93, Mai 2006.

[Gra07] K.L. Gray. The JPEG2000 standard, Janvier 2007. Cours (lecture).

[Haf95] J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner et W. Niblack. Efficient color histogram indexing for quadratic form distance function. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(7) :p. 381–392, ? 1995.

[Ham01] A. Hampapur, K-H. Hyun et R. Bolle. Comparison of sequence matching techniques for video copy detection. Dans *Proc. SPIE Conference on Storage and Retrieval for Media Databases*, t. 4676, p. 194–201. Decembre 2001.

[Han01] G. Hanjalic, A. ans Kakes, R.L. Lagendijk et J. Biemond. Dancers : Delft advanced news retrieval system. Dans *SPIE/IST ELECTRONIC IMAGING, Storage and Retrieval for Media Databases*. San Jose, USA, 2001.

[Han04] A. Hanjalic. *Content-Based Analysis of Digital Video*. Kluwer Academic Publishers (now Springer), Norwell, MA, 2004.

[Han08] A. Hanjalic, R. Lienhart, W.-Y. Ma et J.R. Smith. The holy grail in multimedia information retrieval : So close or yet so far away ? *Proceedings of the IEEE, Special Issue on Advances in multimedia information retrieval*, 96(4) :p. 541–547, Avril 2008.

[Har79] R.M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5) :p. 786–804, 1979.

[Har88] C. Harris et M. Stevens. A combined corner and edge detector. Dans *4th Alvey Vision Conference*, p. 153–158. 1988.

[Hor81] B. Horn et B. Schunck. Determining optical flow. *Artificial Intelligence*, p. 185–203, 1981.

[Hsu02] C.-T. Hsu et S.-J. Teng. Motion trajectory based video indexing and retrieval. Dans *Proceedings of International Conference on Image Processing (ICIP)*, p. 605–608. Rochester, NY, Etats-Unis, 22-25 Septembre 2002.

[Hua97] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu et R. Zabih. Image indexing using color correlograms. Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 762–768. San Juan, Porto Rico, 17-19 Juin 1997.

[Hua04] X-S. Hua, X. Chen et H-J. Zhang. Robust video signature based on ordinal measure. Dans *International Conference on Image Processing (ICIP)*, p. 685–688. Singapore, Singapoure, 24-27 Octobre 2004.

[ICO] ICOSHD. Indexation et Compression Scalables et Conjointes pour la Gestion de Contenus Vidéos de Haute Définition, ANR-06-MDCA-010 ICOSHD. http://icos-hd.irisa.fr.

[ISO93] (MPEG1) ISO/IEC 11172-1 :1993. Information technology : Coding of moving picture s and associated audio for digital storage media at up about 1,5mbits/s, 1993.

[ISO00] (MPEG2) ISO/IEC 13818-2 :2000. Technologies de l'information : Codage générique des images animées et du son associé : Données vidéo, 2000.

[ISO02] (MPEG7) ISO/IEC. 15938-3 :2002. Information technology : Multimedia content description interface. part 3 : Visual, 2002.

[ISO04a] (JPEG2000) ISO/IEC. 15444-1 :2004. Technologies de l'information. Système de codage d'images JPEG2000. Partie 1, 2004.

[ISO04b] (MPEG4) ISO/IEC 14496-1 :2004. Technologies de l'information : Codage des objets audivisuels, 2004.

[ISO06] (Amendement Jpeg2000) ISO/IEC. 15444-1 :2004 amd1 :2006. Technologies de l'information. Profiles for digital cinema applications, 2006.

[ISO07] (MJPEG2000) ISO/IEC. 15444-3 :2007. Information technology. JPEG2000 image coding system : Motion JPEG2000, 2007.

[ISO09] (h264) ISO/IEC. 14496-10 :2009. Coding of audio-visual objects – part 10 : Advanced video coding, 2009.

[ITRH93] (H.261) ITU-T Recommandation H.261. Codec vidéo pour services audiovisuels à p x 64 kbits/s version 3, 1993.

[ITRH05] (H.263) ITU-T Recommandation H.263. Codage vidéo pour communications à faible débit version 3, 2005.

[Jol07] A. Joly, O. Buisson et C. Frelicot. Content-based copy detection using distorsion-based probabilistic similarity search. *IEEE Transactions on Multimedia*, 9(2) :p. 293–306, 2007.

[Jun03] C.R. Jung. Multiscale image segmentation using wavelets and watersheds. Dans *IEEE Proceedings of the XVI Brazilian symposium on computer graphics and image processing*, p. 278–283. 2003.

[Kim01] H.S. Kim et H.W. Park. Wavelet-based moving picture coding using shift invariant motion estimation in wavelet domain. *Signal Processing : Image Communication*, 16 :p. 669–679, 2001.

[Kim03] J.B. Kim et H.J. Kim. Multiresolution-based watersheds for efficient image segmentation. *Pattern recognition letters, Elsevier*, 24 :p. 473–488, 2003.

[Kim05] C. Kim et B. Vasudev. Spatiotemporal sequence matching techniques for video copy detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 1(15) :p. 127–132, janvier 2005.

[Kra05] P. Kramer et J. Benois-Pineau. Camera motion identification in the rough indexing paradigm. Dans *on-line proceedings of TRECVID05*. 2005. Http ://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[Lap03] I. Laptev et T. Lindeberg. Space-time interest point. Dans *International Conference on Computer Vision (ICCV)*, p. 432–439. Nice, France, 13-16 Octobre 2003.

[Lar04] M.C. Larabi, C. Montagne, S. Lelandais, A. Smolarz, Ch. Fernandez-Maloigne et Cornu Ph. Quantification couleur : comparaisons objectives et subjectives de différents algorithmes. *Traitement du signal*, 21, 2004.

[Law06] J. LawTo, O. Buisson, V. Gouet-Brunet et N. Boujemaa. Robust voting algorithm based on labels of behaviour for video copy detection. Dans *14th ACM International Conference on Multimedia (ACM MM)*, p. 835–844. Santa Barbara, CA, Etats-Unis, 23-27 Octobre 2006.

[Law07] J. LawTo, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa et F. Stentiford. Video copy detection : A comparative study. Dans *5th ACM International Conference on Image and Video Retrieval*, p. 371–378. Amsterdam, Pays-Bas, Juillet 2007.

[Lew06] M.S. Lew, N. Sebe, C. Djeraba et R. Jain. Content-based multimedia information retrieval : State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1) :p. 1–19, fevrier 2006.

[Li95] S.Z. Li. *Markov Random Field Modeling in Image Analysis*. 1995.

[Liu07] Y. Liu et K. Ngi Ngan. Fast multiresolution motion estimation algorithms for wavelet-based scalable video coding. *Signal Processing : Image Communication*, 22 :p. 448–465, 2007.

[Liu08] D. Liu et T. Chen. DISCOV : A framework for discovering objects in video. *IEEE Transactions on Multimedia*, 10(2) :p. 200–208, Février 2008.

[LM06] O. Le Meur, P. Le Callet et D. Barba. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5) :p. 802–817, Mai 2006.

[Low99] D. Lowe. Object recognition from local scale-invariant features. Dans *Seventh International Conference on Computer Vision (ICCV)*, t. 2, p. 1150–1158. Kerkyra, Grece, 20-27 Septembre 1999.

[Low04] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :p. 91–110, Janvier 2004.

[Lui03] T.Y. Lui et E. Izquierdo. Scalable object-based image retrieval. Dans *International Conference on Image Processing (ICIP)*, t. 3, p. 501–504. Barcelone, Espagne, 12-17 septembre 2003.

[Mae] D. Maestroni, A. Sarti, M. Tagliasacchi et S. Tubaro. Fast in-band motion estimation with variable size block matching. Dans *ICIP04*.

[Mah07] A. Mahboubi, J. Benois-Pineau et D. Barba. Suivi et indexation des objets dans des séquences vidéo avec la mise-à-jour par confirmation rétrograde. Dans *CORESA'01, Dijon, France*. 2007.

[Mal98] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[Man99a] M.K. Mandal, T. Aboulnasr et S. Panchanathan. Fast wavelet histogram techniques for image indexing. *Computer Vision and Image Understanding*, 75(1/2) :p. 99–110, Juillet/Aout 1999.

[Man99b] M.K. Mandal, F. Idris et S. Panchanathan. A critical evaluation of image and video indexing techniques in the compressed domain. *Image and Vision Computing*, 17 :p. 513–529, Mai 1999.

[Man04] F. Manerba, J. Benois-Pineau et R. Leonardi. Extraction of foreground objects from an MPEG2 video stream in rough-indexing framework. Dans *Storage and Retrieval Methods and Applications for Multimedia 2004*, p. 50–60. SPIE, San Jose, CA, USA, 2004.

[Man05] F. Manerba. *Efficient Object Identification in image sequence for content indexing*. Thèse de doctorat, Université de Brescia, Italie et Université Bordeaux 1, Ecole Doctorale de Mathématiques et Informatique, France, 30 novembre 2005.

[Man08] F. Manerba, J. Benois-Pineau, R. Leonardi et B. Mansencal. Multiple moving object detection for fast video content description in compressed domain. *EURASIP Journal on Advances in Signal Processing*, 2008, Janvier 2008.

[Mar03] D. Marpe, V. George, H.L. Cycon et K.U. Barthel. Performance evaluation of Motion-JPEG2000 in comparison with H.264/AVC operated in pure intra coding mode. Dans *Wavelet Applications in Industrial Processing*, p. 129–137. SPIE, Providence, RI, USA, 2003.

[Mat02] J. Matas, O. Chum, M. Urban et T. Pajdla. Robust wide baseline stereo from Maximally Stable Extremal Regions. *Image and Vision Computing*, 22(10) :p. 761–767, Septembre 2002.

[Mey90] Y. Meyer. *Ondelettes et opérateurs*, t. 1 de *Actualités Mathématiques*. Herman, 1990.

[Mik04] K. Mikolajczyk et C. Schmid. Scale and affine invariant interest point detector. *International Journal of Computer Vision*, 60(1) :p. 63–86, 2004.

[Mis03] M. Misiti, Y. Misiti, G. Oppenheim et J.-M. Poggi. *Les ondelettes et leurs applications*. Traité IC2, série Traitement du Signal et de l'Image. Hermès Science Publications Lavoisier, 2003.

[Mog05] H. A. Moghaddam, T.K. Khajoie, A.H. Rouhi et M.S. Tarzjan. Wavelet correlogram : a new approach for image indexing and retrieval. *Pattern Recognition*, 38 :p. 2506–2518, Decembre 2005.

[Mor81] H. Moravec. Rover visual obstacle avoidance. Dans *International Joint Conference on Artificial Intelligence*, p. 785–790. Vancouver, Canada, 1981.

[Mot] Motion2D. Http ://www.irisa.fr/vista/Motion2D.

[MPE01] Committee MPEG7. Overview of the mpeg-7 standard (version 6.0), 2001. Report ISO/IEC JTC1/SC92/WG11 N4509.

[Nis06] D. Nister et H. Stewenius. Scalable recognition with a vocabulary tree. *IEEE Conference on Computer Vision and Pattern recognition*, 2 :p. 2161–2168, june 2006. http://vis.uky.edu/~stewe/ukbench/.

[Odo95] J.-M. Odobez et P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4) :p. 348–365, December 1995.

[Ohm04] J.-R. Ohm, M. van der Schaar et J. Woods. Interframe wavelet coding - motion picture reesentation for universal scalability. *Signal Processing : Image Communication*, 19(9) :p. 877–908, octobre 2004.

[Oos02] J. Oostveen, T. Kalker et J. Haitsma. *Recent Advances in Visual Information Systems*, t. 2314 de *Lecture Notes in Computer Science*, chap. Feature Extraction and Database Strategy for video fingerprinting. Springer, 2002.

[Pas96] G. Pass et R. Zabih. Histogram refinement for content-based image retrieval. Dans *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV)*, p. 96–102. Sarasota, FL, USA, 2-4 Decembre 1996.

[PF06] S. Philipp-Foliguet et L. Guigues. Evaluation de la segmentation d'images : état de l'art, nouveaux indices et comparaison. *Traitement du Signal*, 23 :p. 109–124, mars 2006.

[Pin] G.S. Pingali, A. Opalach, Y.D. Jean et I.B. Carlbom. Instantly indexed multimedia databases of real world event. *IEEE Trans. Multimedia*, 4(2) :p. 269–282.

[Pir09] P. Piro, S. Anthoine, E. Debreuve et M. Barlaud. Scalable spatio-temporal video indexing using sparse multiscale patches. Dans *7th International Workshop on Content-based Multimedia Indexing (CBMI)*, p. 95–100. Chania, Crète, 3-5 juin 2009.

[Que08] G. Quenot, J. Benois-Pineau, B. Mansencal, F. Precioso, D. Grisse, P. Lambert, B. Augereau, L. Goujon, D. Pellerin, M. Roubent et S. Ayache. Rushes summarisation by IRIM consortium : redundancy removal and multi-features fusion. Dans *2nd ACM TRECVid Video Summarization Workshop*, p. 80–84. Vancouver, British Columbia, Canada, 27-31 Octobre 2008.

[Ren09] W. Ren, S. Singh, M. Singh et Y.S. Zhu. State of the art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42 :p. 267–282, février 2009.

[Sal95] P. Salembier et J. Serra. Flat zones filtering, connected operators, and filters by reconstruction. *IEEE Transactions on Image Processing*, 4(8) :p. 1153–1160, 1995.

[Sal97] P. Salembier, F. Marqués, M. Pardàs, R. Morros, I. Corset, S. Jeannin, L. Bouchard, F. Meyer et B. Marcotegui. Segmentation-based video coding system allowing the manipulation of objects. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1) :p. 60–74, Février 1997.

[Sal00] P. Salembier et L. Garrido. Binary partition trees as an efficient representation for image processing, segmentation, and information retrieval. *IEEE Transactions on Image Processing*, 9(4) :p. 561–576, Avril 2000.

[Sav06] S. Sav, G.J.F. Jones, H. Lee, N.R. O'Connor et A.F. Smeaton. Interactive experiments in object-based retrieval. *Lecture Notes in Computer Science*, 4071 :p. 1–10, 2006.

[Sco79] D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66 :p. 605–610, 1979.

[Seb01] N. Sebe, Q. Tian, E. Loupias, M.S. Lew et T.S. Huang. Salient points for content based retrieval. Dans *CVPIR*, p. 401–410. 2001.

[Shi94] J. Shi et C. Tomasi. Good features to track. Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 593–600. Seattle, WA, Etats-Unis, 21-23 Juin 1994.

[Shi07] H. Shimazaki et S. Shinomoto. A method for selectiong the bin size of the time histogram. *Neural Computation*, 19(6) :p. 1503–1527, juin 2007.

[Sme00] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta et R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12) :p. 1349 – 1380, 2000.

[Smi81] T.F. Smith et M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147 :p. 195–197, 1981.

[Sti99] C. Stiller et J. Konrad. Estimating motion in images sequences. *IEEE Signal Processing Magazine*, 16(4) :p. 70–91, juillet 1999.

[Sto05] A. Stoica, M. Larabi et C. Fernandez-Maloigne. Visual quality enhancement for colour images in the framework of the JPEG2000 compression standard. *International Journal of Robotics and Automation*, 20(2) :p. 109–122, Janvier 2005.

[Str95] M. Stricker et M. Orengo. Similarity of color images. *Proceedings of SPIE, Storage and Retrieval for Image and Video Databases III*, 2420(?) :p. 381–392, Fevrier 1995.

[Stu26] H.A. Sturges. The choice of a class interval. *Journal of American Statistics Association*, 21 :p. 65–66, 1926.

[Swa91] M.J. Swain et D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1) :p. 11–32, Novembre 1991.

[Swe95] W. Sweldens. Wavelets and the lifting scheme : a 5 minutes tour. Dans *ICIAM GAMM 95 part II*, t. 76, p. 41–44. Hamburg, Allemagne, juillet 1995.

[Tau] D. Taubman. Kakadu ©unisearch Ltd. www.kakadusoftware.com.

[Tau00] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Transactions on Image Processing*, 9(7) :p. 1158–1170, juillet 2000.

[Tau02] D.S. Taubman et M.W Marcellin. *Jpeg2000 image compression fundamentals, standards and practice*. Springer, 2002.

[Tek95] A.M. Tekalp. *Digital Video Processing*. Prentice Hall, 1995.

[Tom91] C. Tomasi et T. Kanade. Detection and tracking of point features. Rap. tech. CMU-CS-91-132, International Journal of Computer Vision, 1991.

[TRE] TRECVIDEO. Http ://www-nlpir.nist.gov/projects/trecvid/.

[Tsa02] Y. Tsaig et A. Averbuch. Automatic segmentation of moving objects in video sequences : a region labelling approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(7) :p. 597–612, juillet 2002.

[TUM] TUM. Http ://www.ldv.ei.tum.de/Members/tobias/sequences/tmt/.

[Ugu05] B. UgurTöreyin, A. Enis Cetin, A. Aksay et M. Bilgay Akhan. Moving object detection in wavelet compressed video. *Signal Processing : Image Communication*, 20(3) :p. 255–264, mars 2005.

[Ura95] S. Urago, J. Zerubia et M. Berthod. A markovian model for contour grouping. *Pattern recognition*, 28(5) :p. 683–693, 1995.

[vel95] A. vellaikal et C.-C.J. Kuo. Content-based retrieval of color and multispectral images using joint spatial-spectral indexing. *Proceedings of SPIE, Digital Image Storage and Archiving Systems*, 2606(?) :p. 232–243, Octobre 1995.

[Vel02] R.C. Veltkamp et M. Tanase. Content-based image retrieval systems : a survey. Rap. tech. UU-CS-2000-34, version corrigée et étendue, GIVE Lab, Pays Bas, Octobre 2002.

[VQE] VQEG. Http ://www.its.bldrdoc.gov/vqeg/.

[Wan97] G. Wang, J.Z.and Wiederhold, O. Firschein et S.X. Wei. Wavelet-based image indexing techniques with partial sketch retrieval capability. *Proceedings of the Fouth Forum on Research and Technology Advances in Digital Libraries*, p. 13–24, Mai 1997.

[Wan98] J.Z. Wang, G. Wiederhold, O. Firschein et S.X. Wei. Content-based image indexing and searching using Daubechies' wavelets. *International journal on Digital Libraries*, 1(4) :p. 311–328, mars 1998.

[Wen99] X. Wen, T.D. Huffimire, H.H. Hu et A. Finkelstein. Wavelet-based video indexing and querying. *Multimedia Systems*, 7(5) :p. 350–358, Septembre 1999.

[Wu96] L. Wu, J. Benois-Pineau, Ph. Delagnes et D. Barba. Spatio-temporal segmentation of image sequences for object-oriented low bit-rate image coding. *Signal Processing : Image Communication*, 8(6) :p. 513–543, septembre 1996.

[Yua04] H. Yuan et X.-P. Zhang. Texture image retrieval based on a gaussian mixture model and similarity measure using a kullback divergence. Dans *IEEE International Conference on Multimedia and Expo*, t. 3, p. 1867–1870. juin 2004.

[Zen05] W. Zeng, J. Du, W. Gao et Q. Huang. Robust moving object segmentation on H.264/AVC compressed video using the block-based MRF model. *Real-Time Imaging*, 11 :p. 290–299, juin 2005.

[Zha96] Y.J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8) :p. 1335–1346, 1996.

Publications

International Conferences with Proceedings

[1] C. Morand, J. Benois-Pineau and J.Ph. Domenger. *HD Motion Estimation in a Wavelet Pyramid in JPEG2000 Context.* In Proceedings of the IEEE ICIP Workshop on Multimedia Indexing and Retrieval, San Diego, Californie, USA, pages 61–64, October 2008.

[2] C. Morand, J. Benois-Pineau and J.Ph. Domenger. *Scalable Indexing of HD Video.* In Proceedings of the International Workshop on Content-Based Multimedia Indexing, CBMI08, Londres, Royaume-Uni, pages 417–424, juin 2008.

[3] C. Morand, J. Benois-Pineau, J.Ph. Domenger and B. Mansencal. *Object-Based Indexing of Compressed Video Content : from SD to HD Video.* In Proceedings of the 14th International Conference on Image Analysis and Processing, VMDL07, Modena, Italy, pages 71–76, September 2007.

National Conferences with Proceedings

[4] C. Morand, J. Benois-Pineau and J.Ph. Domenger. *Extraction Spatio-Temporelle d'Objets dans la vidéo HD compressée par des ondelettes sous le paradigme de l'indexation primaire.* In Proceedings of CORESA'07, Montpellier, France, pages 252–256, November 2007.

Journal Papers

[5] Cl. Morand, J. Benois-Pineau, J.-Ph. Domenger, J. Zepeda, E. Kijak and Ch. Guillemot. *Scalable Object-based Video Retrieval in HD Video Databases.* Signal Processing : Image Communication, Elsevier. Accepted for Publication.

Book Chapters

[6] J. Benois-Pineau, S. Anthoine, C. Morand, E. Debreuve, J.-Ph. Domenger, W. Bel Haj and P. Piro *Scalable Indexing of HD Video.* Chapter of book High-quality visual experience : creation, processing and interactivity of high-resolution and high dimensional video signals. Ed M. Mrak, M. Grgic, M. Kunt, Springer Verlag. Accepted for Publication.

Technical Reports

[7] S. Anthoine, P. Piro, M. Barlaud, J. Benois-Pineau, C. Morand, C. Kaes, H. Nicolas, J.-Ph. Domenger. *Descripteurs scalables spatiaux et spatio-temporels.* livrable ANR-06-MDCA-010 ICOS-HD, July 2008.

[8] J. Benois-Pineau, C. Kaes, C. Morand, H. Nicolas, J.-Ph. Domenger, S. Anthoine, M. Barlaud, Ch. Guillemot. *Description de l'architecture initiale de codage vidéo.* livrable ANR-06-MDCA-010 ICOS-HD, December 2007.