Résumé en ligne des vidéos complexes, recherche interactive des images par le contenu, et aide au diagnostic médical basé sur l'analyse des images

Walid Barhoumi

RAPPORT DE SYNTHESE

présenté pour l'obtention de

HABILITATION UNIVERSITAIRE

Discipline : Informatique

par

Walid Barhoumi

Maître Assistant à l'ENICarthage
Equipe de Recherche : SIIVA — Laboratoire RIADI (ENSI)

Résumé en ligne des vidéos complexes, recherche interactive des images par le contenu, et aide au diagnostic médical basé sur l'analyse des images

---

Soutenu le 26 Juin 2015 devant le jury composé de :

Mohamed Mohsen Gammoudi	Professeur à l'ISAMM	Président
Nozha Boujemaa	Directeur de Recherche à l'INRIA	Rapporteur
Vincent Charvillat	Professeur à l'INP Toulouse	Rapporteur
Ezzeddine Zagrouba	Professeur à l'ISI	Membre
Amel Borgi	Maître de Conférences à l'ISI	Membre

Remerciements

Je remercie vivement Monsieur Mohamed Mohsen Gammoudi, Professeur à l'Institut Supérieur des Arts Multimédias de la Manouba, pour l'honneur qu'il me fait en acceptant de présider le jury de mon habilitation universitaire.

Je tiens à remercier tout particulièrement Madame Nozha Boujemaa, Directrice du Centre de Recherche INRIA Saclay-Ile-de France, et Monsieur Vincent Charvillat, Professeur à l'Institut National Polytechnique de Toulouse, de m'avoir fait le grand honneur d'accepter de rapporter sur cette habilitation universitaire.

Je suis très sensible à l'honneur que me fait Madame Amel Borgi, Maître de Conférences à l'Institut Supérieur d'Informatique, d'avoir accepté de faire partie de mon jury d'habilitation universitaire, manifestant ainsi son intérêt à mon travail.

Je voudrais exprimer toute ma reconnaissance à Monsieur Ezzeddine Zagrouba, Professeur à l'Institut Supérieur d'Informatique, d'avoir accepté de participer à mon jury, mais aussi pour sa rigueur scientifique, ses encouragements et la confiance qu'il m'a accordé.

Ce manuscrit relate des travaux réalisés au sein du l'équipe de recherche "Systèmes Intelligents en Imagerie et Vision Artificielle - SIIVA" du Laboratoire RIADI (ENSI). Je suis très reconnaissant à tous les membres de l'équipe SIIVA pour leur convivialité et leur enthousiasme. Les travaux présentés dans ce manuscrit bénéficient particulièrement du travail réalisé par des doctorants que j'ai eu le plaisir de co-encadrer : Slim Amri, Abir Gallas, Sami Dhahbi et Abir Baâzaoui, merci à vous pour le travail que vous avez accompli.

Je suis très reconnaissant à tous mes collègues de l'Ecole Nationale des Ingénieurs de Carthage pour leurs encouragements. Je remercie aussi amis et collègues qui par un conseil, un encouragement ou un sourire m'ont soutenu pour mener à terme mes travaux.

Mes remerciements vont également à ma famille pour leur soutien indéfectible. Une mention spéciale pour ma sœur Dorra pour la relecture détaillée de ce manuscrit.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Dans ce mémoire d'habilitation, j'ai essayé de résumer mes principales contributions au domaine de la recherche en traitement d'images et vision par ordinateur. Il présente une synthèse de mes travaux post-doctoraux, qui ont été menés depuis 2006 au sein de l'équipe de recherche "Systèmes Intelligents en Imagerie et Vision Artificielle" (SIIVA) du laboratoire RIADI. Ces travaux sont les résultats des activités d'encadrement des jeunes chercheurs de l'équipe SIIVA durant la préparation de leurs mémoires de mastère et de thèse de doctorat. En effet, suite à ma thèse de doctorat en informatique, qui a été soutenue en Janvier 2006, j'ai participé au montage de l'équipe de recherche SIIVA qui a permis notamment de structurer l'important potentiel des étudiants en mastère de recherche en génie logiciel de l'Institut Supérieur d'Informatique de Tunis (ISI) en collaboration avec l'Institut National des Sciences Appliquées et de Technologie de Tunis (INSAT). Ainsi, mes activités de recherche se sont élargies à diverses thématiques telles que l'indexation et la recherche de l'information multimédia et l'analyse des vidéos. En effet, les travaux de recherche que je mène concernent principalement trois axes de recherche.

Le premier axe s'articule autour de l'**analyse des vidéos** complexes. En particulier, les recherches s'intéressent aux outils automatiques pour la description compacte et représentative des contenus des vidéos. Afin de minimiser le fossé sémantique, la majorité des nouvelles méthodes combinent des critères bas niveau, donc non-sémantiques, avec la notion d'objet apportant forcément une notion de sémantique. Il existe, en fait, deux principales familles de méthodes fondées sur les objets. La première famille concerne les méthodes qui utilisent la notion d'objet pour extraire des images-clés représentatives du contenu de la vidéo. Dans ce cadre, nous avons proposé de sélectionner, à la volée, un nombre relativement restreint d'images-clés qui résument le contenu visuel saillant d'un plan vidéo. La technique proposée est basée sur la segmentation spatiale de chaque image afin de détecter les événements importants, de sorte que chaque image-clé présente un événement important dans le plan vidéo à savoir l'apparition et/ou la disparition des objets saillants de la scène. Les tests réalisés sur plusieurs vidéos standards ont démontré l'efficacité de cette technique, qui est capable de capter automatiquement le contenu sémantique d'un plan vidéo tout en évitant la redondance des images clés extraites, même lorsque la caméra retourne sur des parties de la scène déjà visitées auparavant (loop-closure). La deuxième famille regroupe les méthodes qui fournissent une décomposition fond/objets de la vidéo. En absence de connaissances préalables sur la scène étudiée et l'environnement d'acquisition, les méthodes de soustraction du fond fournissent le meilleur compromis entre performance et fiabilité. Il s'agit de construire une image mosaïque qui synthétise le contenu spatio-temporel de l'arrière-plan de la scène, avant d'appliquer une fonction de décision sur chaque image afin de décider en tout pixel si celui-ci appartient à l'arrière-plan ou à un objet mobile. Toutefois, dans le cas d'une caméra mobile, la modélisation du fond nécessite la compensation du mouve-

ment de la caméra. Nous avons proposé à cet égard une approche multi-primitive d'alignement d'images. L'originalité principale de cette approche réside dans l'utilisation de l'appariement des régions afin de pré-estimer le mouvement de la caméra et de limiter par la suite l'espace de recherche des homologues potentiels lors de l'appariement des points d'intérêt. Une évaluation objective de l'approche proposée a permis de prouver sa robustesse vis-à-vis du mouvement de la caméra, de la variation d'illumination, de l'état d'acquisition, du bruit et de la présence d'objets mobiles. En plus, vu que plusieurs pixels peuvent être projetés sur la même position, nous avons proposé une technique de mélange médian temporel dans laquelle l'intensité de chaque pixel sur la mosaïque est itérativement mise à jour par la valeur ayant une probabilité d'apparence maximale dans une distribution temporelle et spatiale. Ensuite, les masques des objets mobiles sont estimés en comparant chaque image alignée avec la partie correspondante dans le panorama du fond. Cependant, certains objets du fond peuvent être constamment en micromouvement et il est même possible que des objets soient ajoutés ou retirés de la scène, et la solution proposée doit réagir rapidement pour tenir compte de ces changements. Pour cela, nous avons proposé de raffiner les masques des objets mobiles par des segmentations spatiales, et le modèle de l'arrière-plan est par la suite mis à jour en considérant les modifications apportées aux différents masques de l'avant-plan. La qualité des résultats expérimentaux nous a permis de proposer de nombreuses applications en aval (suivi des objets mobiles, tatouage invisible des vidéos...). En l'occurrence, nous avons intégré notre approche de détection des objets mobiles dans une application de suivi non supervisée de multiples personnes, en présence intensive des effets d'occultation. Pour ce faire, nous avons commencé par séparer les différentes personnes mobiles, en se basant conjointement sur la forme des silhouettes et la segmentation en régions. Les différents blobs ainsi obtenus sont ensuite affectés à des pistes en utilisant un processus d'appariement intégrant à la fois un modèle d'apparence et un modèle de mouvement. La technique proposée est capable de suivre correctement plusieurs personnes mobiles dans des situations d'occultation complexe, même avec une seule caméra à faible résolution temporelle. Cette technique permet une précision meilleure, par rapport à des solutions standards, sans connaissance préalable du nombre de personnes suivies ni une initialisation de leurs positions. Néanmoins, en analysant les résultats produits par notre approche de mosaicing sur de longues vidéos, nous avons pu constater que la synthèse de tout le contenu de la vidéo en une seule mosaïque peut provoquer une taille énorme de cette mosaïque. Ceci est principalement dû au fait que la déformation du modèle de mouvement croît rapidement avec tout changement de l'angle de rotation ou du facteur d'échelle de la caméra, ce qui peut également affecter la qualité de la mosaïque générée. De là, nous avons opté à diviser automatiquement la vidéo en un ensemble de sous-séquences et représenter par la suite chacune par une seule mosaïque. En effet, nous avons proposé de résumer le contenu visuel d'une vidéo complexe en multiple mosaïques en optimisant en ligne le choix de l'image de référence afin de réduire la taille des mosaïques générées, tout en assurant une bonne qualité visuelle. Les limites de chaque mosaïque sont détectées à la volée en utilisant une estimation robuste des paramètres physiques de la caméra. En plus, l'image de référence de chaque mosaïque est choisie, d'une manière dynamique, au milieu de la vue et du zoom de la mosaïque, afin de rassembler sans distorsion le maximum d'images dans la même mosaïque. Cette approche permet de traiter en ligne une longue séquence vidéo, tout en empêchant l'accumulation des erreurs d'alignement et en maintenant des exigences minimales de mémoire. La réalisation en ligne de cette tâche est un problème ouvert qui n'a jamais été traité auparavant et qui pourra présenter une solution clé pour diverses applications multimédias (vidéo streaming, codage en ligne des vidéos...). En particulier, étant donné que

l'ensemble des images qui composent une vidéo n'est pas obligatoirement disponible en entier, notamment dans le contexte de la surveillance en ligne des cibles mobiles, nous avons adapté notre approche de détection des objets mobiles pour un traitement à la volée des personnes mobiles dans une séquence vidéo acquise avec une seule caméra mobile. Dans cette perspective, nous avons procédé à l'estimation en ligne des masques des personnes par un processus itératif de soustraction de fond. Les tests préliminaires ont montré que la méthode proposée permet de détecter à la volée plusieurs personnes mobiles dans différentes situations d'occlusion.

Le deuxième axe est celui de la **recherche de l'information visuelle**. Cet axe traite principalement l'indexation et la recherche des images par le contenu dans des bases généralistes du domaine du Web. Notre objectif est de réduire le fossé sémantique bloquant la quasi-totalité des solutions existantes et ceci en procédant au niveau région, voire objet, tout en donnant la possibilité d'apprentissage aux moteurs de recherche en faisant recours aux techniques de bouclage de pertinence. Ces techniques sont largement utilisées en indexation textuelle des images et pas encore suffisamment appliquées en indexation par le contenu. En effet, fondé sur l'hypothèse que n'importe quelle région pourrait être utile dans le procédé de recherche, nous avons considéré toutes les régions pour chaque image. De là, un graphe complet est défini relativement à chaque image, où chaque nœud représente une région floue grossièrement segmentée et il est caractérisée en termes de deux propriétés : les descripteurs de bas niveau de la région et son poids évaluant son importance visuelle au sein de l'image. Pour la description d'une région, nous avons utilisé l'information spectrale fournie par les transformations d'ondelettes. Nous avons adapté ces transformations pour le cas région tout en se limitant à l'utilisation de la sous-bande de haute fréquence du deuxième niveau des ondelettes, qui comporte la majorité de l'information utile de la région, afin de ne pas alourdir les tâches ultérieures (mise en correspondance, évaluation de la similarité...). En outre, puisque l'information spatiale est considérablement liée à la sémantique du contenu d'une image, cette information est incorporée dans la structure de graphe en caractérisant chaque arête entre deux régions par deux quintuplets illustrant les dispositions spatiales inter-régions dans les deux sens. Ce modèle de graphe, qui est entièrement basé sur la logique floue afin de mieux modéliser l'incertitude et l'ambigüité de l'étape de segmentation, incorpore autant d'information expressive et distinctive que possible. Par la suite, la comparaison de deux images revient à la recherche d'un isomorphisme entre deux graphes complets et pondérés, qui est un problème NP-complet. Pour cette raison, nous avons proposé une approche d'appariement de graphes basée sur des heuristiques garantissant un compromis entre la qualité et le temps de calcul. De l'autre côté, l'utilisateur a la possibilité de réaliser interactivement une ou plusieurs itérations de bouclage de pertinence positif et/ou négatif pour se rapprocher au mieux de son besoin. Nous avons suggéré à ce propos un mécanisme interactif simple, mais efficace, qui se base sur l'adaptation des poids des régions de l'image requête en fonction des rétroactions de l'utilisateur. Ces rétroactions essayent de rapprocher la requête-cible idéale, et ceci en donnant dans les itérations suivantes plus d'importance aux régions de l'image requête présentes en commun dans les exemples positifs tout en minimisant les poids de celles faisant partie des exemples négatifs. En plus, un élargissement de la bande des ondelettes, utilisées comme descripteurs de comparaison, est intégré dans la procédure de bouclage de pertinence, afin de prendre en considération plus de détails relatifs aux allures des objets. En effet, puisque seules les images qui gardent une liaison sémantique avec l'image requête seront considérées lors des itérations ultérieures du bouclage de pertinence, l'élargissement de la bande descriptive des régions lors du bouclage de pertinence permet de mieux caractériser les images tout en évitant

le risque de faire correspondre des images sémantiquement différentes mais qui contiennent des silhouettes similaires. Le bouclage de pertinence proposé est conçu selon la caractéristique de la représentation des images au niveau des régions et il est basé sur une adaptation empirique des poids des régions, pareillement aux techniques utilisées pour la recherche basée sur le texte. Il est conçu selon un modèle de représentation vectorielle qui permet d'obtenir des résultats encourageants même avec un nombre réduit d'exemples, ce qui n'est pas le cas pour les autres modèles de bouclage de pertinence (surtout celui basé sur l'apprentissage). Les expérimentations et l'étude comparative avec des approches existantes prouvent la pertinence des solutions proposées pour la recherche des images en termes d'apport sémantique offert par la modélisation des images par des graphes complets ainsi que par le bouclage de pertinence. Nous nous sommes aussi intéressés à la recherche interactive des images sur le Web en combinant conjointement le texte associé à l'image avec son contenu visuel. En particulier, nous avons considéré le cas des termes polysémiques qui peuvent apparaitre dans des textes associés à des images différentes. En effet, nous avons mis en œuvre une solution qui, après la saisie des mots clés par l'utilisateur, extrait les images annotées sur le Web par ces mots clés tout en procédant à une désambigüisation de requêtes textuelles polysémiques par des images représentatives des différents sens des mots en entrée. L'évaluation objective des résultats montrent l'efficacité de cette solution pour une recherche rapide des images avec des mots clés polysémiques. En outre, partant de l'hypothèse que la co-segmentation peut apporter des solutions efficaces pour la recherche et la classification des images par le contenu, nous avons commencé à explorer ce vaste champ de recherche. La plupart des techniques existantes modélisent la co-segmentation sous la forme d'un problème d'optimisation d'une fonction d'énergie, qui intègre un terme de données intrinsèques relatives aux images, un terme de lissage et un terme de correspondance qui pénalise la dissimilarité entre les images. Cette correspondance est souvent évaluée en comparant les histogrammes en absence de toute information spatiale, ce qui provoque des fausses détections et amplifie les effets du bruit. Pour cela, nous avons proposé d'utiliser l'entropie locale lors de la caractérisation d'une image par son histogramme, tout en faisant recours à une classification floue de cet entropie afin de réduire l'ambiguïté d'appartenance d'un pixel à un bin de l'histogramme. La qualité des résultats expérimentaux nous a encouragé à adapter la technique proposée de co-segmentation pour la détection des athlètes dans des vidéos sportives. Ces vidéos sont très complexes à cause de la complexité des arrière-plans, du mouvement sans contrainte des caméras et de la déformation non-rigide et rapides des athlètes. La solution proposée consiste à réduire la tâche de détection à une simple co-segmentation d'une paire d'images, afin d'en extraire les objets communs dans les avant-plans. En effet, la co-segmentation permet d'intégrer implicitement l'information temporelle pour la segmentation des athlètes dans des environnements sans contrainte et sans aucune intervention de l'utilisateur. Les premiers résultats obtenus montrent une amélioration significative de la méthode proposée par rapport aux techniques de soustraction du fond, qui sont souvent utilisées pour la segmentation des athlètes.

Le dernier axe est celui d'**imagerie médicale**. Les objectifs des travaux sur cet axe sont l'aide au diagnostic basé sur les images médicales et l'annotation et l'indexation automatiques des dossiers numériques des patients. La tendance actuellement observable en imagerie médicale est que les méthodes se mathématisent considérablement avec un objectif de généralisation des concepts pour une interprétation d'images de plus haut niveau. Cependant, les traitements, aussi génériques qu'ils soient, se spécialisent à un niveau ou un autre en fonction du type des images et des objectifs poursuivis par le diagnostic, et ceci est généralement nécessaire pour

obtenir des résultats plus fiables. En effet, trois classes d'architectures ont été utilisées dans la littérature des systèmes d'aide au diagnostic médical basé sur les images, à savoir celles basées sur la classification, celles basées sur le recalage et celles basées sur la recherche des images par le contenu. Toutefois, ces trois approches ont été souvent étudiées séparément malgré leurs points communs et leurs complémentarités. De là, nous avons proposé une architecture qui fusionne conjointement les trois approches. La fusion est basée sur la théorie de l'évidence, et ceci en considérant l'avis de chaque approche comme étant une source incertaine sur la malignité de l'organe étudiée. Cette architecture a prouvé en premier temps son efficacité dans le cadre du dépistage du mélanome (cancer de la peau), avec un taux de classification correcte d'environ 89% pour une base de 200 images. En outre, dans le même contexte d'aide au diagnostic du mélanome, nous avons proposé une méthode structurale pour la détection automatique du réseau de pigment dans les images dermatoscopiques. En effet, ces dernières années, il y a eu un intérêt croissant pour l'utilisation des caractéristiques de texture, notamment les réseaux de pigment, comme symptômes de malignité pour les lésions de la peau. Suite à l'extraction de la lésion de la peau à partir de l'image en entrée, la méthode proposée est basée sur un filtrage LoG d'une image de luminance bien sélectionnée afin de détecter les trous et les autres structures au sein de cette lésion. Par ailleurs, un processus de seuillage est introduit afin de filtrer uniquement les trous appartenant au réseau de pigment tout en excluant les autres structures rondes comme les points, les globules et les bulles d'huile. La principale contribution de la méthode proposée réside dans l'évaluation floue du degré d'appartenance d'un trou au réseau de pigment, ce qui permet de garder le maximum de candidats et de repousser la décision jusqu'à l'obtention de plus amples informations lors des étapes suivantes. Les trous retenus sont ensuite reliés, tout en vérifiant une contrainte spatiale, via un graphe représentant le réseau de pigment. Cette méthode a assuré une aire sous la courbe ROC de 0.821 pour détecter avec succès les réseaux de pigment avec un taux de classification correcte de 85% sur un ensemble de 122 images réelles. Afin d'appliquer l'architecture proposée dans le contexte du dépistage par mammographie du cancer des seins, nous avons commencé par étudier le problème général du recalage d'images médicales. Une revue de littérature focalisée sur les contraintes d'autonomie, de temps de calcul et de qualité de recalage nous a permis de conclure que seules les méthodes iconiques ne nécessitent pas une étape de segmentation et que le choix du modèle de déformation s'avère comme un compromis entre le temps de calcul et la qualité du recalage. Ceci nous a amené à opter pour l'approche de subdivision progressive tout en essayant d'y contribuer en optimisant aussi bien le temps du calcul que la qualité du recalage. En effet, nous avons adopté l'approche de subdivision progressive, non plus sur une image de taille fixe, mais plutôt sur une pyramide Gaussienne d'images. La validation de cette approche sur des mammographies des seins et sur des images IRM cérébrales a montré que l'utilisation d'une pyramide Gaussienne permet de diminuer énormément le temps de calcul, grâce à l'utilisation conjointe des hiérarchies des données et des déformations, sans pourtant dégrader la qualité du recalage. En effet, les transformations les plus simples sont estimées à partir des images de petites tailles, alors que les transformations les plus complexes sont estimées à partir de celles de grandes tailles. Nous avons aussi contribué à l'annotation sémantique des comptes rendus d'imagerie médicale de type DICOM-SR (Structured Reporting), en proposant une technique d'intégration des données multi-sources. Cette technique est basée sur l'utilisation des ontologies pour les nombreux avantages qu'elles proposent tels que leurs représentations formelles, leurs concepts qui sont moins ambigus que les termes, leurs vocabulaires partageables, leurs réutilisabilités ainsi que leurs évolutions. L'annotation proposée considère aussi les annotations

des anciens comptes rendus enregistrés dans la base de données ainsi que les informations de bas niveau des images des comptes rendus à annoter. Cette annotation a été efficacement validée dans un contexte de comptes rendus d'imagerie ostéoarticulaire. Par ailleurs, dans le cadre de la chirurgie assistée par ordinateur, nous nous sommes récemment intéressés à la reconstruction des modèles 3D à partir d'images 2D. Nous avons commencé par proposer une amélioration de la méthode de coloration des voxels, qui est une technique populaire de reconstruction d'un modèle 3D d'une surface volumétrique à partir d'un ensemble d'images calibrées. Elle consiste à affecter des couleurs aux voxels dans un volume tridimensionnel en garantissant la cohérence avec l'ensemble des images en entrée. Cependant, la qualité de la reconstruction est fortement dépendante d'une procédure de seuillage permettant de décider, pour chaque voxel, s'il est photo-consistant ou non. En effet, il est extrêmement difficile de définir les seuils appropriés, qui devraient être précis et stables pour tous les voxels. En plus de l'intégration de l'information géométrique en utilisant un seuillage par hystérésis qui prend en considération la connexité des voxels colorés, notre contribution réside dans le choix dynamique et entièrement automatique des seuils. En effet, nous avons défini, relativement à chaque voxel, un degré d'appartenance à la classe des voxels photo-consistants, en fonction de l'homogénéité de ses projections. Nous avons aussi modélisé les deux seuils du seuillage par hystérésis par des suites harmoniques adjacents qui évoluent en fonction du nombre des images sur lesquelles le voxel est projeté. Les résultats préliminaires prouvent la précision de la technique proposée qui permet l'incorporation de la cohérence spatiale lors de la reconstruction du volume, tout en évitant les voxels flottants et les trous.

Le reste de ce manuscrit est organisé comme suit : le premier chapitre synthétise les contributions apportées au résumé des vidéos basé conjointement sur la génération des multiples mosaïques et la détection des objets mobiles. Le deuxième chapitre aborde la recherche des images par le contenu. La contribution principale réside essentiellement dans l'intégration de la recherche à base des régions avec le bouclage de pertinence, dans l'objectif de s'approcher le plus possible de la perception humaine. Ensuite, nous présentons nos contributions quant au diagnostic médical basé sur l'analyse des images par le contenu. Le dernier chapitre est destiné à une synthèse des travaux en cours et des perspectives à suivre.

Chapitre 1

Résumé en ligne des vidéos complexes

1.1 Introduction

L'évolution rapide des caméscopes numériques, des réseaux de télécommunication et des moyens de stockage, explique le besoin urgent d'outils automatiques pour la description compacte et représentative des contenus des vidéos. Toutefois, la plupart des méthodes existantes pour la création des résumés vidéos nécessitent une forte interaction avec un opérateur d'annotation textuelle, ce qui pose des problèmes de subjectivité et de faisabilité. Par conséquent, les nouvelles méthodes basées sur le contenu combinent des critères bas niveau, donc non-sémantiques, avec la notion d'objet apportant forcément une notion de sémantique. Il existe en fait deux principales familles de méthodes fondées sur les objets. La première famille englobe les techniques qui utilisent la notion d'objet pour extraire des images-clés représentatives du contenu de la vidéo [24]. Toutefois, les dynamiques sous-jacentes ne seront pas correctement capturées et il est toujours difficile d'évaluer la pertinence des images choisies. La deuxième famille regroupe les méthodes qui fournissent une décomposition fond/objets sur l'ensemble des images. Puisque cette représentation apporte plus d'informations que les images-clés, nous avons focalisé notre intérêt sur ce type de résumé. Une synthèse bibliographique nous a permis de classifier les méthodes de séparation entre l'avant-plan et l'arrière-plan d'une vidéo en trois catégories : celles qui s'appuient sur une analyse du mouvement, celles qui se basent sur les différences d'images et enfin celles qui sont basées sur la soustraction du fond. Pour la première catégorie, la classification du mouvement est fondée sur une estimation du flot optique, qui est très sensible aux mouvements de forte amplitude. Ces méthodes sont très coûteuses et engendrent plusieurs objets sur-segmentés. En plus, elles échouent dans le cas des caméras effectuant des mouvements libres, surtout pour des fonds non-statiques et en présence d'objets mobiles rapides. Pour la seconde catégorie, il est très difficile d'extraire simultanément les objets rapides et les objets lents. En particulier, si l'objet mobile contient des zones homogènes, l'image des différences risque de contenir des valeurs faibles dans les zones de recouvrement, et l'objet n'est alors extrait que partiellement. En plus, ces méthodes sont très sensibles au bruit et rencontrent souvent des difficultés lors de l'identification des petits objets mobiles. Ainsi, en l'absence de connaissances préalables sur la scène étudiée et sur l'environnement d'acquisition, les méthodes de soustraction du fond sont les plus employées pour la séparation entre l'avant-plan et l'arrière-plan d'une vidéo complexe. Ces méthodes fournissent le meilleur compromis entre performance et fiabilité [142]. Il s'agit de construire un modèle de l'arrière-plan de la scène, puis d'appliquer une fonction de décision sur chaque image afin de décider en tout pixel si celui-

ci appartient à l'arrière-plan ou à un objet mobile. La modélisation de l'arrière-plan nécessite la construction d'une image mosaïque qui synthétise le contenu spatio-temporel du fond de la vidéo. Ceci revient principalement à calculer les transformations géométriques qui relient les images avant de les projeter sur le plan d'une seule image choisie comme référence. Cependant, puisque la déformation du modèle de mouvement croît rapidement avec l'angle de rotation de la caméra [95], la synthèse de tout le contenu de la vidéo en une seule mosaïque peut provoquer une taille énorme de cette mosaïque. Le changement du facteur d'échelle de la caméra peut également affecter la taille et la qualité de la mosaïque générée, puisque chaque image zoomée doit être redimensionnée pour être fusionnée dans la mosaïque. De là, nous avons opté à diviser automatiquement la vidéo en un ensemble de sous-séquences et représenter par la suite chacune par une seule mosaïque. Ceci permet de minimiser l'espace mémoire utilisé tout en optimisant la qualité de reconstruction [94] [97]. En effet, nous avons proposé de synthétiser en ligne le contenu visuel de l'arrière-plan d'une vidéo complexe en multiples mosaïques, de sorte que les limites et l'image de référence de chaque mosaïque sont détectées à la volée. La réalisation en ligne de cette tâche est un problème ouvert qui n'a jamais été traité auparavant et qui présentera une solution clé pour diverses applications multimédias [59]. Ensuite, les masques des objets mobiles sont estimés en comparant chaque image alignée avec la partie correspondante dans le panorama convenant. Cependant, certains objets du fond peuvent être constamment en micromouvement et il est même possible que des objets soient ajoutés ou retirés de la scène et la solution proposée doit réagir rapidement pour tenir compte de ces changements. Des portions d'objets en mouvement peuvent aussi partager les mêmes caractéristiques photométriques avec des objets du fond, ce qui complique la séparation entre les objets de l'arrière-plan avec ceux de l'avant-plan, surtout si ces derniers sont rapides et non-rigides. Aussi, outre les effets d'occultation des objets mobiles, les fluctuations causées par le processus d'acquisition se reflètent souvent par la présence d'objets sémantiquement non-significatifs (ombres, inter-réflexions...). Pour cela, nous avons proposé de raffiner les masques des objets mobiles par des segmentations spatiales, et le modèle de l'arrière-plan est par la suite mis à jour en considérant les modifications apportées aux différents masques de l'avant-plan. La qualité des résultats expérimentaux permet de proposer de nombreuses applications en aval [87] [130]. Dans notre cas, nous avons intégré notre solution dans une application de suivi non supervisée de multiples personnes.

1.2 Génération en ligne des multiples mosaïques

L'approche proposée permet de résumer en ligne le contenu visuel de l'arrière-plan d'une vidéo en multiples mosaïques. A chaque itération i, l'image reçue F_i est supposée être dans l'un de trois cas. Elle pourrait être projetée sans distorsions sur une mosaïque S_j parmi l'ensemble Ω des mosaïques déjà générées (***Cas1***). Sinon, nous tentons de mettre à jour l'image de référence d'une mosaïque existante afin qu'elle sera capable de rassembler sans distorsions l'image F_i et éventuellement les images qui suivent (***Cas2***). Par ailleurs, F_i représentera l'image de référence d'une nouvelle mosaïque $S_{Card(\Omega)+1}$ (***Cas3***). En effet, si la projection de F_i sur la mosaïque courante S_t, avec $t = Card(\Omega)$, risque de provoquer des distorsions, nous essayons de changer l'image de référence de S_t. Si le changement de l'image de référence n'est pas possible, nous cherchons une mosaïque S_j $(j < t)$ dans Ω où nous pouvons projeter F_i sans distorsions. Nous commençons par tester la mosaïque précédente S_{t-1}, sinon nous essayons avec S_{t-2}, et ainsi de suite. La mosaïque sélectionnée S_j sera mise à jour par un mélange du contenu nouvellement découvert. Si l'image F_i ne peut pas être projetée sur aucune des mosaïques construites

précédemment, même après un changement de référence, une nouvelle mosaïque, dont l'image de référence est F_i, doit être définie. Plus précisément, la première image F_1 est supposée être l'image de référence de la première mosaïque S_1 et représente ainsi son contenu visuel initial ($S_1 = F_1$). Ensuite, le mouvement de la caméra de chaque image F_i ($i \geq 2$), relativement à la vue de la caméra de l'image de référence de S_1, est estimé. Puisqu'une mosaïque préliminaire est toujours générée, nous avons opté pour une estimation image-mosaïque du mouvement afin d'éviter l'accumulation des erreurs de recalage [94]. Puis, les distorsions qui pourraient accompagner la projection de F_i sur S_1 sont estimées, en fonction des paramètres du mouvement de la caméra, afin de décider si F_i peut être fusionnée dans S_1, sans modifier son image de référence, ou non (la fonction binaire "*Affectation*"). Dans le cas où la fusion est possible, nous utilisons une technique de mélange médian temporel [5] (la fonction "*Melange*") afin d'écarter les objets en mouvement. Dans le cas contraire, nous essayons d'abord de changer l'image de référence de la mosaïque courante en choisissant sa dernière image F_l^1 comme nouvelle image de référence (la fonction binaire "*Maj_ Image_ Ref*"). Si ceci est possible, nous commençons par reprojeter S_1 sur la nouvelle image de référence (la fonction "*Changer_ Ref*") avant de projeter F_i sur la nouvelle mosaïque. Si le changement de l'image de référence n'est pas possible, une nouvelle mosaïque S_2 doit être construite, sachant que F_i est son image de référence. Si plus qu'une seule mosaïque est déjà générée ($t \geq 2$), une image reçue n'est supposée représenter la référence d'un nouvelle mosaïque que si elle ne peut pas être fusionnée ni à S_t, ni à S_{t-1}, ..., ni à S_1 ($\neg associe$), même en changeant l'image de référence (tout en donnant la priorité aux mosaïques "récemment" générées). Sachant qu'initialement $\Omega = \emptyset$, le schéma proposé peut être résumé, pour chaque itération i, comme suit :

Algorithme 1: Génération en ligne des multiples mosaïques

Entrées : F_i, Ω ;
Sorties : Ω ;

début
  $associe \leftarrow faux$;
  $j \leftarrow Card(\Omega)$; /* j est l'indice de la mosaïque courante. */

  tant que *($j \geq 1$ et $\neg associe$)* **faire**
    si *($Affectation(F_i, S_j)$)* /* **Cas1** */ **alors**
      $S_j \leftarrow Melange(S_j, F_i)$;
      $associe \leftarrow vrai$;
    sinon
      si *($Maj\_Image\_Ref(S_j)$)* /* **Cas2** */ **alors**
        $S_j \leftarrow Changer\_Ref(S_j, F_l^j)$;
        $S_j \leftarrow Melange(S_j, F_i)$;
        $associe \leftarrow vrai$;
      sinon
        $j \leftarrow j - 1$;

  si *($\neg associe$)* /* **Cas3** */ **alors**
    $S_{Card(\Omega)+1} \leftarrow F_i$;
    $\Omega \leftarrow \Omega \cup S_{Card(\Omega)+1}$;

1.2.1 Estimation du mouvement de la caméra

Cette phase permet d'estimer le mouvement de la caméra pour une image F_i relativement à la vue de la caméra dans l'image de référence d'une mosaïque S_j. Vu l'absence de toute connaissance sur la géométrie épipolaire (caméras simples sans tripodes et dont les paramètres intrinsèques et extrinsèques sont inconnus), la recherche des correspondants des primitives d'une image s'étale sur la totalité de l'autre image. Dans notre cas, ce mouvement est décrit avec le modèle géométrique projectif avec huit paramètres afin d'effectuer une transformation homographique entre les images. Comme la scène est bien approchée par un plan dans la plupart des situations du monde réel, ces huit paramètres permettent d'inclure tous les types de mouvements 3D rigides qui se produisent pour une caméra rotationnelle avec une distance focale variable [67]. Nous utilisons ici les coordonnées homogènes pour exprimer les points et la transformation projective peut être donc définie par (1.1) :

$$\begin{bmatrix} x' \\ \\ y' \end{bmatrix} = \begin{bmatrix} \dfrac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}} \\ \\ \dfrac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}} \end{bmatrix}, \tag{1.1}$$

où, $(x, y)^t$ sont les coordonnées d'un point dans le plan de l'image de référence de S_j et $(x^{'}, y^{'})^t$ sont celles du point correspondant dans F_i. La matrice $H_{i,j} = (h_{11}h_{12}h_{13}; h_{21}h_{22}h_{23}; h_{31}h_{32}h_{33})$ définit les paramètres du mouvement entre F_i et S_j et elle est habituellement normalisée avec $h_{33} = 1$. Pour estimer cette matrice, nous avons introduit une approche multi-primitive qui combine le niveau pixel avec celui région afin de garantir un bon compromis entre la complexité et la représentativité. Le premier module de cette approche consiste à extraire les régions et les points d'intérêt pour chaque image. Ensuite, une mise en correspondance des régions saillantes permet d'estimer un espace réduit pour la recherche des points d'intérêt homologues. L'estimation de cet espace revient à estimer préliminairement le mouvement de la caméra tout en tenant compte des corrélations entre les pixels voisins. Ensuite, un problème de minimisation est mis en place pour estimer l'homographie à partir des couples des points appariés.

1.2.1.1 Extraction des primitives

Ce module est composé de deux étapes qui peuvent être réalisées simultanément. La première étape consiste à segmenter en régions le couple d'images d'entrée et la deuxième étape est réservée pour l'extraction des points d'intérêt. Nous avons choisi d'utiliser la primitive région dont les attributs sont sémantiquement les plus riches et les plus stables, ce qui nous permet de pré-estimer le mouvement de la caméra (rotation et facteur d'échelle) et de limiter, par la suite, l'espace de la recherche des coordonnées des homologues de chaque point d'intérêt. Comme le but de la segmentation en régions dans notre cas n'est qu'une première estimation du mouvement de la caméra, un petit nombre de régions homologues suffit pour réaliser ce but. Pour ce faire, nous avons utilisé l'algorithme de ligne de partage des eaux (LPE) sur l'image gradient [160]. Cet algorithme a prouvé son efficacité pour obtenir des partitions préliminaires pour un large éventail d'images. En outre, afin de surmonter les effets de sur-segmentation, habituellement causée par LPE, un post-traitement fusionne itérativement les régions voisines les plus similaires. Pour cela, la structure de l'image est modélisée par un graphe d'adjacence des régions sur-segmentées, sachant qu'une mesure de similarité est associée à chaque couple

de régions adjacentes. Pour caractériser la similarité entre les régions, nous avons comparé leurs caractéristiques dans l'espace couleur LAB, qui quantifie les différences visuelles entre les couleurs [171] .De l'autre côté, nous avons utilisé le détecteur de Harris [70] pour extraire les points d'intérêt, vu qu'il est plus stable et plus robuste que les autres détecteurs, surtout en cas de rotation de la caméra autour de l'axe optique [57] (Fig. 1.1). En effet, puisque nous disposons d'une estimation préliminaire du facteur d'échelle grâce à l'appariement des régions, le détecteur de Harris est le plus approprié dans notre cas, même par rapport au détecteur SIFT (Scale Invariant Feature Transform) [107] qui peut générer un nombre énorme de points d'intérêt par image (de l'ordre de 1000), ce qui n'optimise pas le temps de traitement.

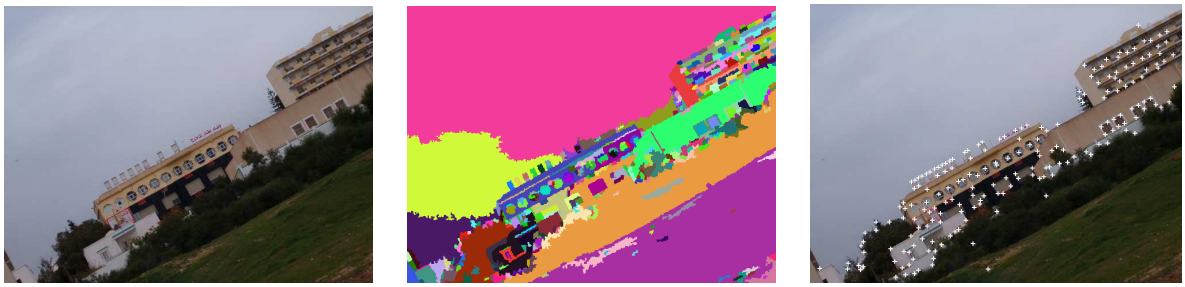


image originale segmentation en régions extraction des points d'intêrêt

FIGURE 1.1 – Extraction des primitives.

1.2.1.2 Appariement des primitives

Ce module commence par une étape d'appariement par prédiction/validation des régions dans l'objectif de produire une première estimation de la rotation autour de l'axe optique et du facteur d'échelle. Cette étape met en correspondance les deux cartes de régions produites par la segmentation, tout en écartant celles qui touchent les bords des images afin d'éviter les faux appariements. La prédiction est réalisée en mesurant, pour chaque paire de régions des deux images I_1 et I_2 [1], un score de corrélation afin de ne conserver que les couples des régions fortement corrélées. En effet, une fonction de corrélation Cor (1.2) évalue la similarité entre deux régions en se basant sur un ensemble $\mathcal{A}$ d'attributs normalisés de nature géométrique et photométrique [2]. Nous avons évité de pondérer les attributs vu qu'il est très difficile de quantifier l'importance de chaque attribut et sa stabilité quant au changement de l'angle de prise de vue et de luminosité et aux défauts du processus d'extraction des régions.

$$\mathrm{Cor}(R_1, R_2) = \frac{1}{Card(\mathcal{A})} \cdot \sum_{A \in \mathcal{A}} \frac{Min(A(R_1), A(R_2)}{Max(A(R_1), A(R_2))}. \tag{1.2}$$

Par la suite, parmi l'ensemble Λ de couples des régions fortement corrélées, nous conservons ceux qui vérifient la contrainte de position relative [23]. En effet, nous cherchons quatre couples qui vérifient une similarité spatiale entre les quadrilatères formés par leurs centres de gravité (Fig. 1.2). Ces régions accomplissent des scores de corrélation élevés et se déplacent dans une direction cohérente, ce qui permet de combiner la cohérence spatiale et celle photométrique pour définir le mouvement saillant de la caméra. Ceci se traduit par la vérification de l'équation

1. Dans notre cas, ces deux images représentent l'image courante F_i et une mosaïque S_j.
2. $\mathcal{A}$ = {Compacité, Rectangularité, Excentricité, Elongation, Sphéricité, Convexité, Moyenne de l'intensité}.

(1.3) et de la similarité des deux angles α_1 et α_2 ($\alpha_l = (\overrightarrow{G_l^1 G_l^2}, \overrightarrow{G_l^1 G_l^4})$, avec G_l^n est le centre de gravité de la région R_l^n de l'image I_l). L'ensemble retenu Λ_f permet l'estimation préliminaire du mouvement de la caméra, qui est modélisé par une transformation rigide (1.4) composée d'une rotation d'un angle α autour de l'axe optique, une translation L_x (*resp.* L_y) le long de l'axe des x (*resp.* des y) [52], et un facteur d'échelle f. Pour ce faire, nous commençons par considérer l'ensemble Λ_{me} des deux couples de régions (R_1^1, R_2^1) et (R_1^2, R_2^2) ($\subset \Lambda_f$) les plus corrélées. L'angle entre les vecteurs $\overrightarrow{G_1^1 G_1^2}$ et $\overrightarrow{G_2^1 G_2^2}$ représente une première estimation de α (Fig. 1.3) et le ratio des distances entre (G_1^1, G_1^2) et (G_2^1, G_2^2) constitue une première estimation de f (1.5).

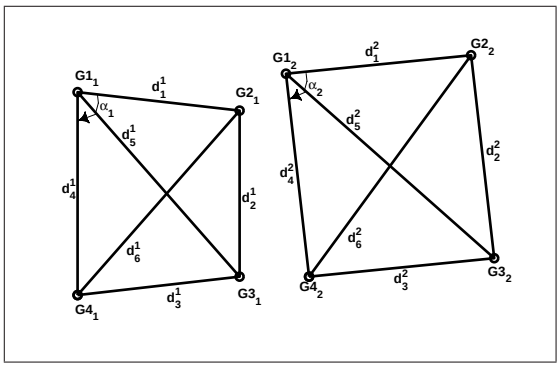


FIGURE 1.2 – Contrainte de la position relative.

$$\forall\ (j,k) \in \{1,\ldots,6\}^2, \frac{d_j^1}{d_j^2} \approx \frac{d_k^1}{d_k^2}. \tag{1.3}$$

$$\begin{bmatrix} x' \\ \\ y' \end{bmatrix} = \begin{bmatrix} f(x\cos\rho - y\sin\rho) + L_x \\ \\ f(x\sin\rho + y\cos\rho) + L_y \end{bmatrix}. \tag{1.4}$$

$$\alpha = \alpha_2 - \alpha_1 = \arctan(\frac{x_2^2 - x_2^1}{y_2^2 - y_2^1}) - \arctan(\frac{x_1^2 - x_1^1}{y_1^2 - y_1^1}) \text{ et} f = \frac{\|\overrightarrow{G_1^1 G_1^2}\|}{\|\overrightarrow{G_2^1 G_2^2}\|}. \tag{1.5}$$

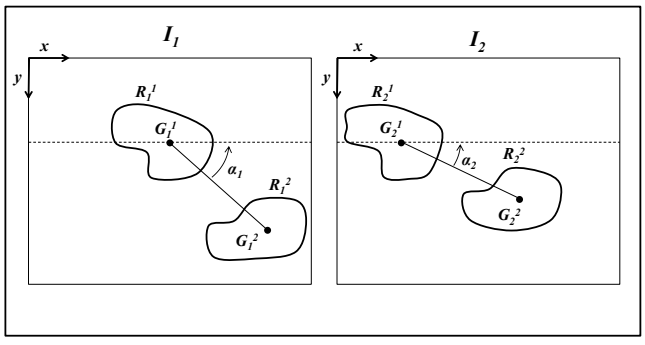


FIGURE 1.3 – Estimation de la rotation autour de l'axe optique.

Ensuite, nous appliquons la combinaison ξ de la rotation par l'angle α avec le facteur d'échelle f sur chaque pixel de l'image I_2, tout en appliquant une interpolation bi-cubique pour les points sans antécédents, afin d'annuler l'effet de ces mouvements [3]. La nouvelle image RI_2 sera utilisée au lieu de I_2 pour l'appariement des points d'intérêt afin de réduire l'espace de la recherche des homologues de chaque point. En effet, à partir de l'analyse des positions relatives des centres de gravité des régions homologues (R_1^j, R_2^j) dans Λ_f, nous pouvons estimer les dimensions L_x et L_y de la fenêtre de recherche des homologues des points d'intérêt. En effet, la valeur de L_x (*resp.* L_y) correspond à la longueur de l'intervalle de disparité en x (*resp.* en y) (1.6).

$$L_{direction} = \max_{(R_1^j, R_2^j) \in \Lambda_f} d^j_{direction} - \min_{(R_1^j, R_2^j) \in \Lambda_f} d^j_{direction}, \tag{1.6}$$

avec, d_x^j (*resp.* d_y^j) dénote la distance Euclidienne entre les coordonnées en x (*resp.* en y) des centres de gravité des régions associées. Pour un grand ensemble d'images de taille 640×480, une réduction de la fenêtre de recherche à la dimension moyenne de 5×5 a été achevée. Nous avons aussi évalué objectivement la répétabilité de l'estimation préliminaire du mouvement sous différentes transformations géométriques et photométriques. Pour ce faire, nous avons utilisé des séquences standards [114] qui illustrent plusieurs variations géométriques et photométriques, sachant que chaque image est accompagnée de l'homographie qui représente la vérité-terrain [4]. Nous avons défini une mesure de répétabilité $\Re$ (1.7) qui est basée sur le taux de chevauchement entre les régions homologues utilisées pour l'estimation du mouvement [5] (Fig. 1.4). En effet, nous avons considéré l'ensemble Λ_f des couples utilisés pour la définition de la fenêtre de la recherche (*resp.* l'ensemble Λ_{me} utilisé pour l'estimation préliminaire du mouvement) et nous avons appliqué l'homographie de référence H_{ref}, reliant les images I_{ref} et I_{test}, sur l'ensemble Λ_f (*resp.* Λ_{me}) des régions dans I_{ref}. Les résultats obtenus (Fig. 1.5) prouvent la robustesse de l'estimation préliminaire du mouvement de la caméra vis-à-vis de plusieurs transformations [167]. Notons que la chute du score de répétabilité, pour un changement d'échelle au-delà de 2.4, revient au fait que les frontières des régions deviennent lisses et le processus de segmentation devient moins précis. Pour un changement d'échelle de 1.35, le score de répétabilité obtenu avec Λ_f est légèrement meilleure que celui calculé avec Λ_{me}. Ceci est dû à la présence de quelques défauts de segmentation autour des frontières des régions de Λ_{me}. Ainsi, parmi tous les couples de régions de Λ_f, les deux couples de régions les plus corrélées (Λ_{me}) n'enregistrent pas toujours les meilleurs taux de ressemblance. En ce qui concerne les scores de répétabilité nuls pour des changements élevés de point de vue et d'échelle, ceci s'explique par l'absence de couples de régions appariées. En effet, les descripteurs utilisés ne sont pas invariants aux transformations projectives, et un changement significatif du point de vue et/ou d'échelle limite l'appariement des régions. Ceci peut s'expliquer par le fait qu'un tel changement conduit à une petite zone de chevauchement entre les images I_{ref} et I_{test} [114]. Dans ces conditions, les régions sont considérées comme des candidats inefficaces pour l'estimation préliminaire du mouvement, et l'approche proposée utilise uniquement l'appariement des points d'intérêt. Etant donné que la taille de la zone de chevauchement est extrêmement réduite dans ce cas, l'appariement est réalisé sans ambiguïté significative.

3. ξ est aussi appliquée sur les coordonnées des points d'intérêt et des centres de gravité des régions appariées.
4. Les séquences et les homographies sont disponibles sur `http://www.robots.ox.ac.uk/~km`.
5. Λ désigne l'ensemble Λ_f ou l'ensemble Λ_{me} en fonction de l'ensemble utilisé dans l'évaluation.

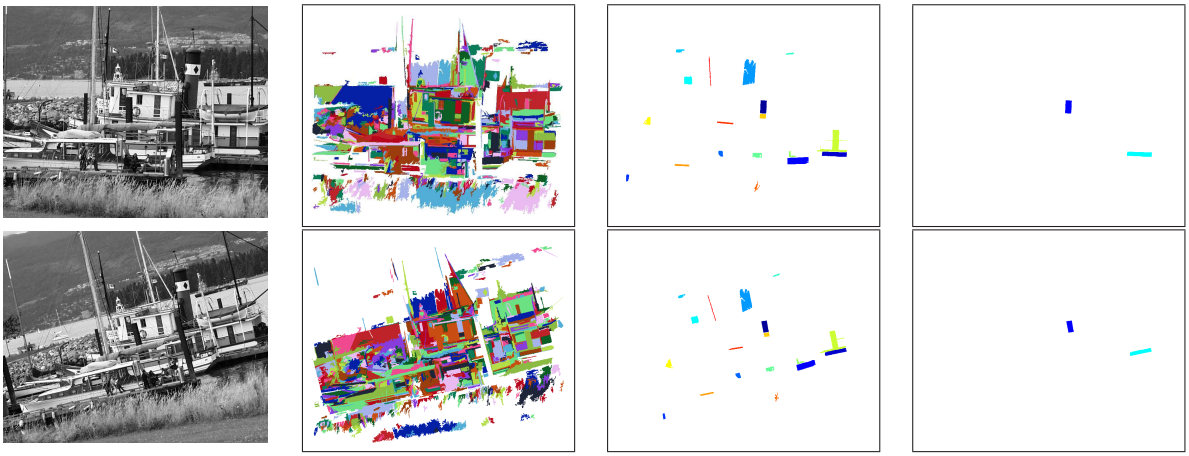
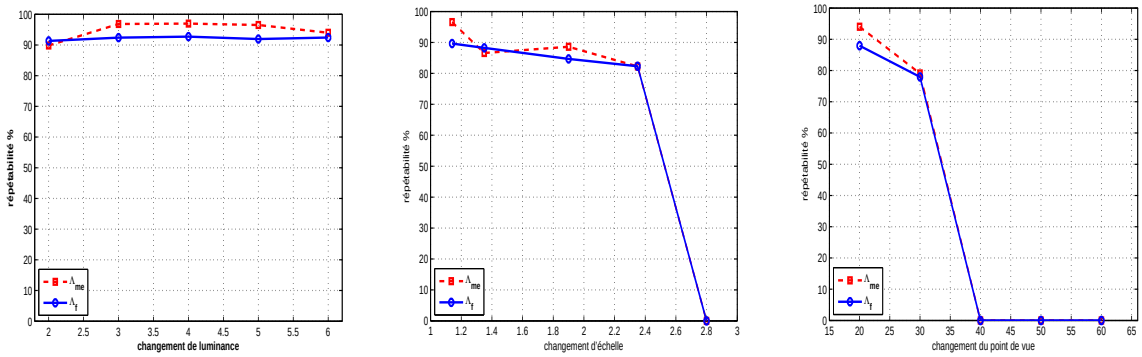


FIGURE 1.4 – Sélection des régions pour l'estimation du mouvement de la caméra.

$$\Re = \frac{1}{2.|\Lambda|} \sum_{(R_r,R_t)\in\Lambda} \left(\frac{Card(R_t \bigcap H^t_{ref}.R_r.H_{ref})}{Card(R_t)} + \frac{Card(R_t \bigcap H^t_{ref}.R_r.H_{ref})}{Card(H^t_{ref}.R_r.H_{ref})} \right). \tag{1.7}$$



(a) changement de luminance (b) changement du facteur d'échelle (c) changement du point de vue

FIGURE 1.5 – Evaluation objective de l'estimation préliminaire du mouvement de la caméra.

Après avoir estimé la fenêtre de recherche, sur l'image RI_2, des homologues de chaque point d'intérêt de I_1, tous les points de I_1 qui engendrent des fenêtres de recherche à l'extérieur de RI_2 sont des points sans homologues. Pour tous les autres points, nous calculons le score de corrélation ZNCC (Zero-mean Normalized Cross Correlation) (1.8) entre chacun de ces points et les points d'intérêts qui se trouvent à l'intérieur de la zone de recherche correspondante dans RI_2. Cette mesure est invariante aux changements linéaires locaux de l'intensité et fournit la meilleure performance d'appariement [52] [137]. Puis, seulement les hypothèses ayant des scores de corrélation élevés et qui ne violent pas la contrainte d'unicité sont retenues[6]. Ensuite,

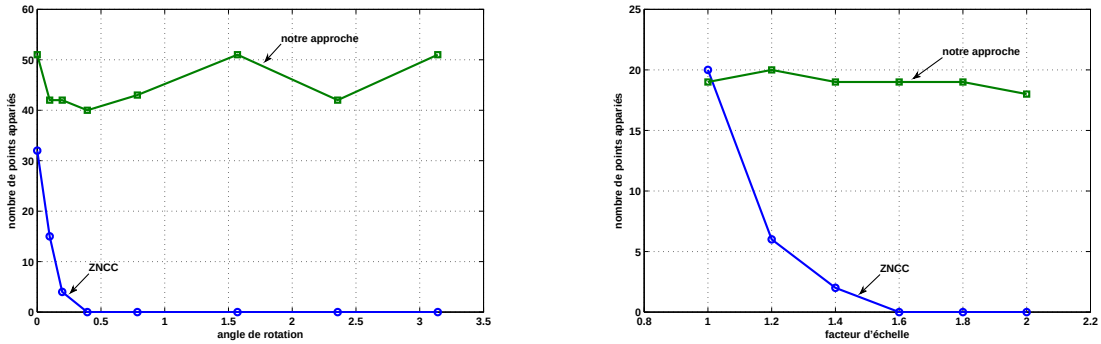
6. Ce cas n'est pas fréquent car les dimensions de la fenêtre de recherche sont suffisamment petites.

un processus itératif considère chaque couple (P_k^1, P_k^2) de points appariés et si l'un des huit voisins de P_k^2 réalise un score de corrélation meilleur que celui enregistré avec P_k^1, alors P_k^2 est remplacé par ce voisin. Enfin, nous raffinons l'ensemble des homologues par la technique RANSAC (RANdom SAmpling Consensus) de rejet des points isolés, qui a été de loin la méthode la plus adoptée pour le traitement des sources d'erreurs externes [136].

$$\begin{cases} \text{ZNCC}(x,y,x',y') = \dfrac{\sum_{i,j} Q^1_{i,j}.Q^2_{i,j}}{\sqrt{\sum_{i,j} Q^3_{i,j}}.\sqrt{\sum_{i,j} Q^4_{i,j}}}, \quad \text{avec} \\ Q^1_{i,j} = (I(x+i,y+j) - \overline{I(x,y)}),\ Q^2_{i,j} = (I'(x'+i,y'+j) - \overline{I'(x',y')}), \\ Q^3_{i,j} = (I(x+i,y+j) - \overline{I(x,y)})^2 \text{ et } Q^4_{i,j} = (I'(x'+i,y'+j) - \overline{I'(x',y')})^2, \end{cases} \tag{1.8}$$

où, $\overline{I_1(x,y)}$ (*resp.* $\overline{I_2(x',y')}$) représente la moyenne des niveaux de gris des pixels appartenant à une fenêtre centrée en (x,y) $(\in I_1)$ (*resp.* en (x',y') $(\in RI_2)$).

Pour montrer l'importance de l'appariement des régions pour l'estimation des fenêtres de recherches associées aux points d'intérêt, nous avons comparé les résultats de l'appariement de ces points par la méthode proposée et par l'unique utilisation du score ZNCC (Fig. 1.6). Nous avons déduit qu'à partir d'un angle de rotation de $\frac{\pi}{8}$, l'utilisation du score ZNCC uniquement, ne permet d'apparier aucun point d'intérêt, contrairement à notre approche pour laquelle le nombre de points appariés est faiblement lié à l'angle de rotation. De même, à partir d'un facteur d'échelle de 1.2, l'utilisation du score ZNCC uniquement ne permet pas d'apparier un nombre suffisant de points d'intérêt, contrairement à notre approche pour laquelle ce nombre est faiblement lié au facteur d'échelle. Par ailleurs, les statistiques de l'appariement montrent que les erreurs d'estimation de l'angle de rotation (≤ 0.023) et du facteur d'échelle (≤ 0.06) sont très faibles indépendamment de la transformation appliquée [167].



(a) cas d'une rotation autour de l'axe optique

(b) cas d'un facteur d'échelle

FIGURE 1.6 – Nombre des points appariés.

1.2.1.3 Estimation de l'homographie

Sous l'hypothèse de planarité de la scène, l'homographie $H_{1,2}$ reliant deux images I_1 et I_2 est une transformation linéaire en coordonnées homogènes (3.13). Ainsi, $H_{1,2}$ est la solution

d'un système à huit équations et huit inconnues pour quatre couples de points appariés non colinéaires. Toutefois, nos expérimentations ont montré que la qualité d'alignement en considérant uniquement quatre couples de points n'est pas satisfaisante (Fig. 1.8), vu qu'il peut y avoir des faux appariements avec des scores élevés de corrélation. Pour cela, un problème d'optimisation a été envisagé et il a été résolu en utilisant la méthode de factorisation QR [140] suivie par un algorithme de relaxation. Cet algorithme raffine l'homographie en éliminant itérativement les couples qui représentent des faux appariements. Ce processus s'arrête lorsqu'on atteint un niveau de précision sous-pixellique. Les expérimentations réalisées sur des séries d'images variées[7], ont prouvé l'efficacité de l'approche proposée pour l'estimation du mouvement de la caméra vis-à-vis de plusieurs types de mouvements complexes (rotation, changement d'échelle et changement de point de vue), de transformations photométriques (changement de luminosité et changement de contraste) et de la présence des objets en mouvement [167] (Fig. 1.7).

$$\begin{bmatrix} x' & y' & 1 & 0 & 0 & 0 & -x'x & -y'x \\ 0 & 0 & 0 & x' & y' & 1 & -x'y & -y'y \end{bmatrix} \cdot H_{1,2} = \begin{bmatrix} x \\ y \end{bmatrix}. \tag{1.9}$$

FIGURE 1.7 – Résultat de l'alignement de cinq images sur le plan de la troisième image.

Dans le but d'évaluer objectivement la qualité globale de l'alignement, nous avons utilisé l'erreur d'alignement E (1.10) qui mesure la moyenne de la valeur absolue de la différence des intensités sur l'espace de chevauchement O d'une paire d'images. Nous avons calculé cette erreur E pour onze paires afin de tracer trois courbes (Fig. 1.8) illustrant l'erreur d'alignement par utilisation de : 4 couples ayant les meilleurs scores de corrélation ($E1$), un nombre de couples appartenant à l'intervalle $]4, 12]$ ($E2$), et un nombre de couples supérieur à 12 ($E3$). Ces courbes

7. Les images ont été prises par des dispositifs simples (caméras sans tripodes et dont les paramètres intrinsèques et extrinsèques sont inconnus) et présentent des espaces de chevauchement entre 20% et 90%.

confirment l'insuffisance de l'unique utilisation des meilleurs couples et l'intérêt d'introduire une technique d'optimisation pour estimer l'homographie.

$$E = \frac{1}{Card(O)} \sum_{(x,y)\in O} |RI_2(x,y) - I_1(x,y)|. \tag{1.10}$$

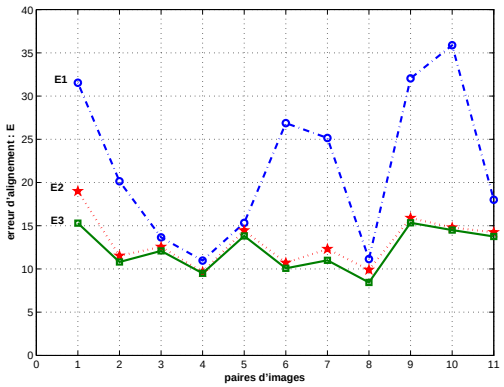


FIGURE 1.8 – Evaluation de l'erreur d'alignement.

1.2.2 Evaluation de la distorsion

Cette phase évalue la distorsion qui accompagne la projection d'une image F_i sur une mosaïque S_j. En effet, une fois le mouvement de la caméra entre F_i et S_j est estimé en termes d'homographie $H_{i,j}$, un algorithme linéaire d'ajustement de paquets est utilisé pour estimer les paramètres de la transformation physique entre ces images. Ces paramètres sont les angles ρ_x^i et ρ_y^i de rotation de la caméra autour de l'axe des x et de l'axe des y, ainsi que les distances focales f_i et f_j de l'image F_i et de l'image de référence F_{rj} de S_j. Ensuite, un algorithme de seuillage décide si l'image F_i peut être exclusivement affectée à S_j, et ceci soit en gardant son image de référence soit en la changeant, ou pas [20].

1.2.2.1 Estimation des paramètres physiques de la caméra

L'objectif de cette étape est de factoriser l'homographie $H_{i,j}$ reliant F_i à S_j en une séquence d'angles de rotation et de distances focales. En effet, il a été souvent confirmé que le modèle de mouvement projectif pour les caméras rotationnelles peut être efficacement considéré comme une concaténation d'opérations physiques élémentaires. De nombreuses méthodes linéaires et non linéaires [55] [32] ont étendu la méthode linéaire d'auto-calibration introduite dans [71] pour supporter le cas général des caméras rotationnelles avec des paramètres intrinsèques variables. Cependant, en plus de leurs cadres d'optimisation global qui atténuent des instabilités numériques, ces méthodes génèrent un cas particulier de la transformation dégénérée de l'image [92] lorsque la caméra effectue une rotation très faible. Dans ce cas, les solutions deviennent instables surtout pour des effets intenses de zoom et en présence de bruit. En effet, l'angle de rotation ne peut pas être estimé puisque les distances focales sont estimées en premier lieu et les angles de rotation sont ensuite calculés en se basant sur ces distances. Pour résoudre

ce problème (appelé "configuration quasi-dégénérée"), la méthode linéaire proposée dans [146] pour estimer la distance focale, a été étendue pour décomposer l'homographie en un produit des matrices des paramètres intrinsèques et extrinsèques de la caméra. Une estimation de la distance focale de la caméra est tout d'abord obtenue en profitant de la contrainte d'orthogonalité ainsi que de la contrainte du norme constante des lignes et des colonnes de la matrice de rotation. Pour éviter les transformations dégénérées, la distance focale de l'image de référence est définie par la valeur médiane des solutions obtenues pour toutes les paires d'images [94]. Les angles de rotation sont ensuite estimés, en se basant sur les propriétés trigonométriques des points centraux de chaque image. Néanmoins, cette solution ne peut être appliquée qu'en mode hors-ligne. Pour cela, nous avons adapté une méthode linéaire à faible coût [92] qui offre une solution simple d'auto-calibration pour la configuration spéciale d'un angle de rotation négligeable autour de l'axe des z (ce qui est généralement admis pour la plupart des cas). Ainsi, la transformation projective $H_{i,j}$, permettant de projeter F_i sur l'image de référence F_{rj} de S_j, peut être définie comme étant le produit de trois matrices $H_{i,j} = K_i R_i K_j^{-1}$, où K_i (*resp.* K_j) représente la matrice de calibration de la caméra pour F_i (*resp.* pour F_{rj}) et R_i désigne la rotation de la caméra entre F_i et F_{rj}. Sous l'hypothèse d'un angle de rotation négligeable autour de l'axe des z et considérant le fait que la première image de chaque mosaïque définit son image de référence (*i.e.* $\rho_x^1 = \rho_y^1 = 0$), l'homographie $H_{i,j}$ peut être défini par (1.11) :

$$H_{i,j} = \begin{pmatrix} \cos\rho_y^i & \sin\rho_x^i \sin\rho_y^i & -f_j \cos\rho_x^i \sin\rho_y^i \\ 0 & \cos\rho_x^i & f_j \sin\rho_x^i \\ \frac{1}{f_i}\sin\rho_y^i & -\frac{\sin\rho_x^i \cos\rho_y^i}{f_i} & \frac{f_j}{f_i}\cos\rho_x^i \cos\rho_y^i \end{pmatrix}. \tag{1.11}$$

avec, ρ_x^i (*resp.* ρ_y^i) est l'angle de rotation de la caméra autour de l'axe des x (*resp.* des y) entre F_i et F_{rj}, et f_i et f_j sont respectivement les distances focales de F_i et F_{rj}. Après quelques calculs algébriques pour déterminer les paramètres inconnus, les angles de rotation ρ_x^i et ρ_y^i peuvent être estimés avec précision (1.12) et les distances focales sont ensuite déduites à partir de ces angles de rotation[8] (1.13) (1.14). Cette méthode permet ainsi de résoudre le problème de la configuration quasi-dégénérée puisqu'elle commence par l'estimation des angles de rotation avant de déduire les distances focales. En plus, elle est bien adaptée pour un traitement en ligne et nos expérimentations ont montré que les paramètres récupérés sont très similaires à ceux qui sont estimés avec les méthodes non linéaires [59] [97].

$$\begin{bmatrix} \rho_x^i \\ \\ \rho_y^i \end{bmatrix} = \begin{bmatrix} \frac{h_{23}h_{33}}{|h_{23}h_{33}|} \arctan\sqrt{\left|\frac{h_{23}h_{32}}{h_{22}h_{33}}\right|} \\ \\ -\frac{h_{13}h_{33}}{|h_{13}h_{33}|} \arctan\sqrt{\left|\frac{h_{13}h_{31}}{h_{11}h_{33}}\right|} \end{bmatrix}. \tag{1.12}$$

$$\begin{aligned} f_j = \max(0, \frac{|\rho_y^i| - |\rho_x^i|}{|\,|\rho_y^i| - |\rho_x^i|\,|})\sqrt{|\frac{h_{13}h_{33}}{h_{11}h_{31}+h_{12}h_{32}}|} + \\ \max(0, \frac{|\rho_x^i| - |\rho_y^i|}{|\,|\rho_x^i| - |\rho_y^i|\,|})\sqrt{|\frac{h_{23}h_{33}}{h_{21}h_{31}+h_{22}h_{32}}|}. \end{aligned} \tag{1.13}$$

8. Si $h_{00}h_{20} + h_{01}h_{21} = 0$ (cas du zoom pur idéal), f_j est une constante non nulle choisie arbitrairement.

$$f_i = \frac{1}{2}\left(\sqrt{\frac{f_j^2(h_{11}^2 + h_{12}^2) + h_{13}^2}{f_j^2(h_{31}^2 + h_{32}^2) + h_{33}^2}} + \sqrt{\frac{f_j^2(h_{21}^2 + h_{22}^2) + h_{23}^2}{f_j^2(h_{31}^2 + h_{32}^2) + h_{33}^2}}\right). \tag{1.14}$$

1.2.2.2 Décision d'affectation d'une image à une mosaïque

Les distorsions se produisent dans la plupart des cas lorsque la caméra tourne énormément par rapport à l'image de référence. Les effets du zooming de la caméra peuvent également affecter considérablement la transformation. De là, nous avons proposé une règle de décision qui combine un coût ζ_r^i de la distorsion causée par la rotation avec un coût ζ_z^i de la distorsion du zooming pour décider si F_i devrait être attribuée à S_j. En effet, afin d'éviter la configuration quasi-dégénérée, toute mosaïque ne doit pas couvrir théoriquement 90 degrés, ou plus, de la rotation de la caméra sur n'importe quelle direction. Toutefois, puisque les distorsions géométriques augmentent rapidement lorsque la caméra tourne loin de l'image de référence, l'angle de vision est beaucoup plus petit dans la pratique. D'ailleurs, la première distorsion significative de la rotation est détectée dans l'image 246 pour la séquence "Stefan", bien que des angles de rotation plus élevés soient enregistrés pour plusieurs images précédentes (Fig. 1.9). Ainsi, l'angle de rotation seul n'est pas suffisant pour détecter la distorsion de rotation qui dépend aussi de la distance focale f_j de l'image de référence de S_j ainsi que de la distance focale f_i de F_i. En effet, pour un angle de rotation ρ, plus le système de coordonnées de la mosaïque est loin de la caméra, plus la distorsion de rotation est large, qui est aussi fortement influencée par tout changement de f_i (Fig. 1.10). Pour cela, nous avons défini le coût ζ_{rx}^i (*resp.* ζ_{ry}^i) de la distorsion causée par la rotation autour de l'axe des x (*resp.* des y) comme le produit de la distance focale f_j par la tangente de l'angle Ψ qui est la somme de l'angle de rotation ρ_x^i (*resp.* ρ_y^i) et la moitié du champ de vision FOV_x (*resp.* FOV_y) de F_i (Fig. 1.11). Le coût ζ_r^i (1.15) ne doit pas dépasser un seuil de rotation γ_r, qui est défini dynamiquement en fonction de la distance focale de F_i. En effet, pour chaque image F_i de taille M×N, le seuil γ_r est le produit de f_i par $\tan(\beta)$, où β ($= 30$ degrés) est l'angle de vue du modèle du système visuel humain [82]. En outre, nous avons défini le coût du zooming ζ_z^i comme le rapport entre f_i et f_j, qui doit être moyen ($\in [\gamma_f^{inf}, \gamma_f^{sup}]$) afin d'éviter les distorsions causées par le zoom-avant de la caméra ($\zeta_z^i < \gamma_f^{inf}$) ainsi que celles générées lorsque la caméra fait un zoom-arrière ($\zeta_z^i > \gamma_f^{sup}$).

$$\begin{aligned} &\zeta_r^i = \max(\zeta_{rx}^i, \zeta_{ry}^i) = \max(f_j \tan(\rho_x^i + \tfrac{FOV_x}{2}), f_j \tan(\rho_y^i + \tfrac{FOV_y}{2})), \\ &\text{avec,} \quad FOV_x = 2\arctan(\tfrac{M}{2f_i}) \quad \text{et} \quad FOV_y = 2\arctan(\tfrac{N}{2f_i}). \end{aligned} \tag{1.15}$$

Ainsi, une image F_i n'est autorisée à être affectée à une mosaïque S_j, sans changer son image de référence, que si la projection de F_i sur S_j ne cause pas des effets forts du facteur d'échelle et qu'elle n'est pas accompagnée d'un coût élevé de la rotation (***Cas1***). Cependant, lorsque le seuil de rotation est atteint dans un sens sans qu'il existe dans la mosaïque des images dans le sens opposé (la mosaïque n'est pas symétrique), l'image de référence de cette mosaïque doit être mise à jour. De même, si le seuil de zoom-avant (*resp.* zoom-arrière) est atteint et il n'existe pas dans la mosaïque des images zoomées en arrière (*resp.* en avant) par rapport à l'image de référence, cette dernière doit être modifiée. Dans ces deux cas, il est inappréciable de changer la mosaïque tant que la projection est toujours possible en changeant l'image de référence. En effet, le choix de la dernière image F_l^j (dans la plupart des cas, $F_l^j = F_{i-1}$) de la mosaïque S_j comme nouvelle image de référence de S_j garantit que celle-ci soit au centre de la vue et au

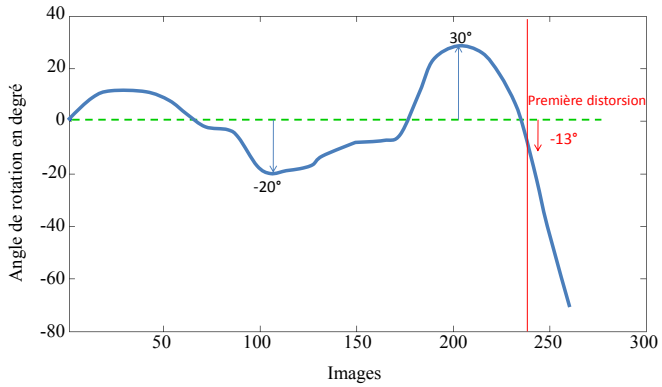
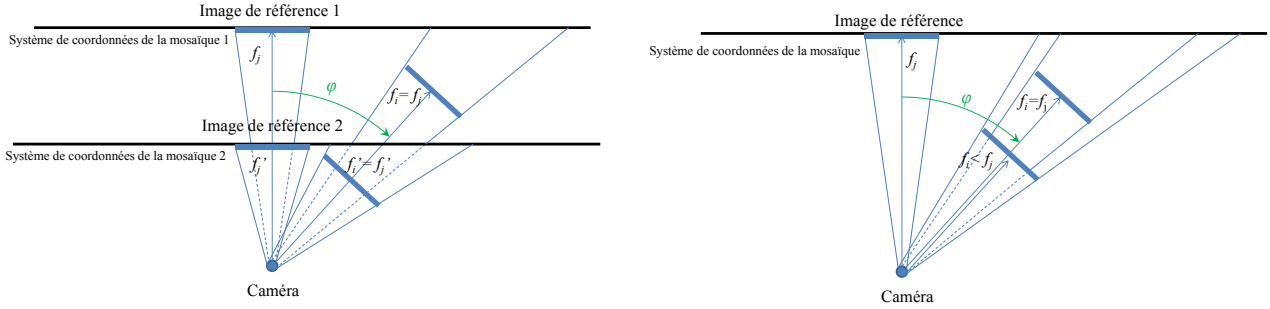


FIGURE 1.9 – Variation de l'angle de rotation ρ_y^i pour la séquence "Stefan".



(a) (b)

FIGURE 1.10 – Influence des distances focales, f_j (a) et f_i (b), sur la distorsion de la rotation.

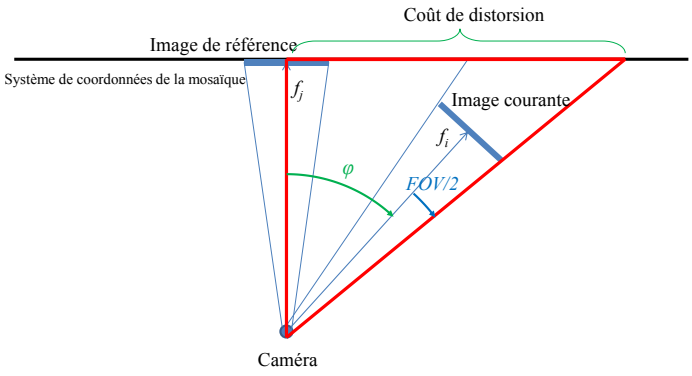


FIGURE 1.11 – Evaluation du coût de la distorsion causée par la rotation.

centre du zoom de la séquence (***Cas2***). Sinon, si le seuil de rotation est atteint dans un sens et il existe dans la mosaïque des images dans le sens opposé (la mosaïque courante est quasiment symétrique), il faut commencer une nouvelle mosaïque S_{t+1} dont l'image de référence est F_i (***Cas3***) (Fig. 1.12). En effet, si nous choisissons F_l^j comme image de référence, la partie visitée par la caméra dans le sens opposé peut causer des distorsions vu que le coût de rotation dans cette partie est supérieur au seuil. De même, si le seuil du zoom-avant (*resp.* zoom-arrière) est atteint et il existe dans la mosaïque des images zoomées en arrière (*resp.* en avant), une nouvelle mosaïque devra être générée puisque la partie zoomée en arrière (*resp.* en avant) ne peut plus être projetée dans le plan de référence de F_l^j [17]. Ce schéma peut être synthétisé comme suit :

si $(\zeta_{rx}^i < \gamma_r$ ET $\zeta_{ry}^i < \gamma_r$ ET $\zeta_z^i \in [\gamma_f^{inf}, \gamma_f^{sup}])$ **alors** ***Cas1***
sinon
si $[\zeta_{ry}^i > \gamma_r$ ET $(\nexists F_l \in S_j / \ \rho_y^l \cdot \rho_y^i < 0)]$ OU $[\zeta_{rx}^i > \gamma_r$ ET $(\nexists F_l \in S_j / \ \rho_x^l \cdot \rho_x^i < 0)]$ OU $[\zeta_z^i > \gamma_f^{sup}$ ET $(\nexists F_l \in S_j / \ \zeta_z^l \in [\gamma_f^{inf}, 1])]$ OU $[\zeta_z^i < \gamma_f^{inf}$ ET $(\nexists F_l \in S_j / \ \zeta_z^l \in [1, \gamma_f^{sup}])]$ **alors** ***Cas2***
sinon ***Cas3***

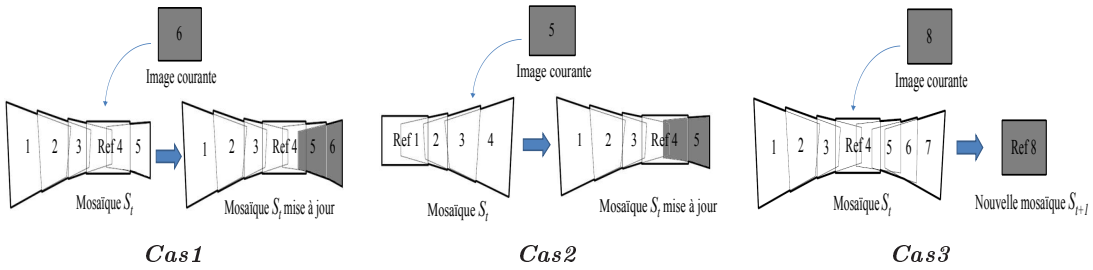


FIGURE 1.12 – Illustration des trois cas possibles de l'approche proposée.

Enfin, étant donné que plusieurs pixels peuvent être projetés sur la même position, un processus de mélange doit être utilisé pour déterminer la valeur de chaque pixel dans l'image panoramique. A cet égard, nous avons proposé une technique de mélange médian temporel dans laquelle l'intensité de chaque pixel sur la mosaïque est itérativement mise à jour par la valeur avec la probabilité d'apparence maximale dans une distribution temporelle et spatiale. Cette technique permet de supprimer récursivement les objets mobiles au fur et à mesure que les images arrivent (Fig. 1.13). Mais, dans le cas des vidéos très complexes, quelques parties des objets mobiles, en particulier ceux qui sont très rapides et qui sont occultés par d'autres objets, apparaissent partiellement comme traces des fantômes dans les panoramas. Ceci est dû au fait qu'une partie du fond n'est que partiellement visible dans la séquence. Dans ce cas, l'image produite par le mélange médian est considérée comme une première approximation de l'arrière-plan, qui sera raffinée, plus tard, conjointement avec l'avant-plan (*c.f* Section 1.3). L'approche proposée de génération en ligne des multiples mosaïques a été validée sur plusieurs vidéos complexes [9]. En l'occurrence, comme le mouvement de la caméra est relativement réduit pour la séquence "Moutain", qui est composée de 100 images de taille 352×192, toute la séquence a été synthétisée en une seule mosaïque de haute qualité visuelle (Fig. 1.14.a), sans aucun changement de l'image de référence. Cependant, pour la séquence "Tabletennis", qui est composée de 89 images de taille 352×288, l'approche proposée évite la croissance importante de la taille de la mosaïque et divise cette séquence en deux partitions sans distorsions visibles (Fig. 1.14.b). Les images de cette séquence sont continuellement zoomées en arrière, à partir d'un gros plan sur la main du joueur jusqu'à une large vue du joueur complet. Comme la vue couverte dans chaque image est plus grande que celle de l'image précédente, la taille totale de la mosaïque augmente sans cesse et il est préférable de diviser cette séquence en plusieurs mosaïques, bien

9. Les seuils γ_f^{inf} et γ_f^{sup} utilisés pour le partitionnement multi-mosaïque ont été paramétrés à 0.5 et 2, successivement.

que la plupart du contenu visuel de chaque image ait été déjà visible dans les images précédentes. En particulier, la mise à jour dynamique de l'image de référence permet de réduire le nombre de mosaïques nécessaires (de 3 à 2), tout en diminuant les tailles des mosaïques produites d'un facteur de 0.56. Le seuil du zoom-arrière est atteint la première fois à l'image 54, et l'image 53 devient ainsi l'image de référence à la place de l'image 1. Quand le seuil du zoom-arrière est atteint une deuxième fois, une deuxième mosaïque a été créée à partir de l'image 79. Les résultats obtenus prouvent que l'approche proposée apporte des améliorations par rapport à la littérature en termes de distorsions et du nombre de mosaïques (Tab. 1.1).

$1-20$ $1-50$ $1-90$

FIGURE 1.13 – Evolution du mélange médian pour la séquence "Stefan".

(a) S_1 : 1-100, ref : 1 (b) S_1 : 1-78, ref : 53 S_2 : 79-89, ref : 79

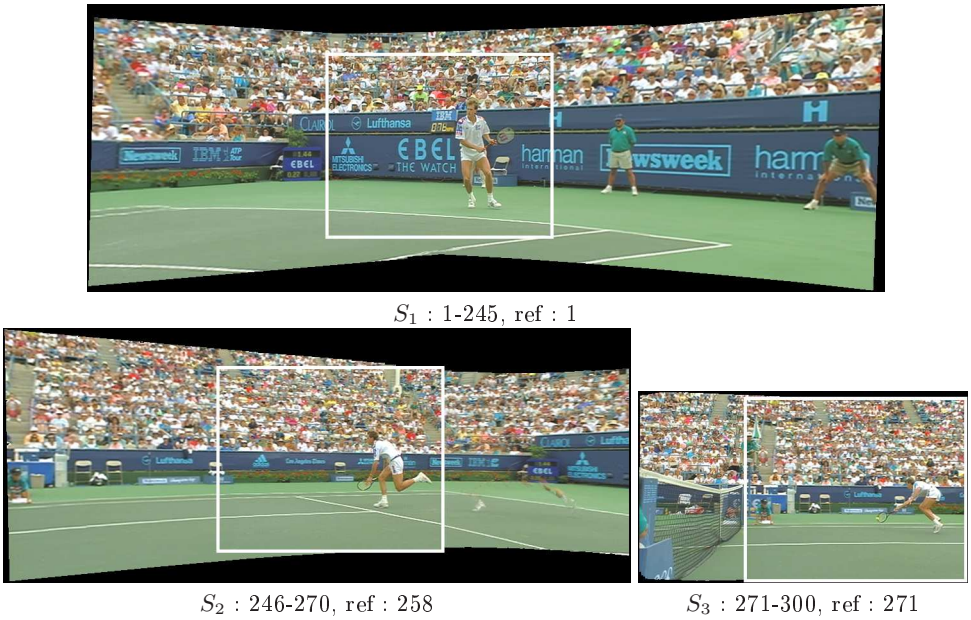
FIGURE 1.14 – Génération des mosaïques pour les séquences "Mountain" (a) et "Tabletennis" (b).

Méthode	Partitions (image de référence)
Farin et al. [59]	1-51 (1), 52-77 (52), 78-89 (78)
Kuo et Chen [95]	1-51 (9), 52-77 (57), 78-89 (83)
Approche proposée	1-78 (53), 79-89 (79)

TABLE 1.1 – Comparaison des résultats pour la séquence "Tabletennis".

Pour la séquence "Stefan", qui est composée de 300 images de taille 346×280, le mouvement de la caméra est très étendu et la majorité des techniques n'arrivent pas à produire une seule mosaïque pour les 300 images. L'approche proposée permet de générer trois mosaïques de haute qualité visuelle (Fig. 1.15), sachant que la mise à jour de l'image de référence permet de réduire le nombre de mosaïques (de 4 à 3) tout en diminuant leurs tailles d'un facteur de 0.82. Le seuil de rotation autour de l'axe des y est atteint la première fois à l'image 246 pendant la rotation de la caméra vers la gauche. Puisque la caméra a déjà visité la partie droite, une deuxième mosaïque est créée à partir de l'image 246. Le seuil de rotation est atteint la première fois dans la deuxième mosaïque à l'image 258, qui devient l'image de référence de la deuxième mosaïque au lieu de l'image 246. Le seuil de rotation est atteint une deuxième

fois dans le deuxième mosaïque et une troisième mosaïque est créée. Par ailleurs, notre approche a produit trois mosaïques pour la séquence "Rail", qui est composée de 140 images de taille 344×284. Vu l'absence d'objets mobiles dans cette séquence, nous l'avons utilisé pour une évaluation objective de la qualité visuelle des mosaïques. Nous avons défini le rapport $PSNR$ pour chaque image par la différence entre l'image originale et sa rétroprojection à partir de la mosaïque correspondante (Fig. 1.16). Bien que le mouvement de la caméra soit vaste pour cette séquence, il est clair que l'approche proposée atteint des valeurs de $PSNR$ assez prometteuses [10].



S_1 : 1-245, ref : 1

S_2 : 246-270, ref : 258 S_3 : 271-300, ref : 271

FIGURE 1.15 – Génération des mosaïques pour la séquence "Stefan".

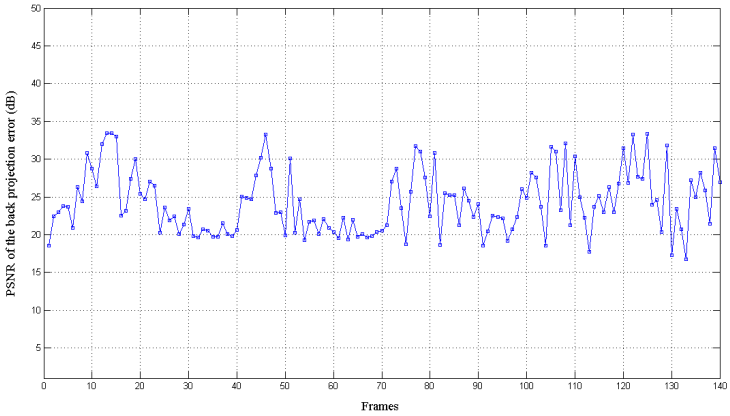


FIGURE 1.16 – Courbe PSNR de la séquence "Rail".

10. Une valeur de $PSNR$ aux alentours de $30dB$ garantit en général une qualité visuelle très acceptable [125].

1.3 Détection des objets mobiles

Les techniques de soustraction du fond sont les plus employées pour la détection d'objets mobiles dans des vidéos complexes. Il s'agit de bâtir un modèle de l'arrière-plan avant d'appliquer une fonction sur chaque image afin de décider en tout pixel si celui-ci appartient à l'arrière-plan ou à l'avant-plan. Parmi ces techniques, on distingue celles qui établaient une classification arrière-plan/avant-plan en se basant sur le niveau du pixel, de celles qui le font au niveau de la région. L'avantage de celles basées sur le niveau pixel est qu'elles ne nécessitent pas une détection des primitives 2D [117]. Toutefois, elles ne prennent pas en compte le degré de corrélation entre les pixels voisins [40] et, par conséquent, elles sont plus sensibles au bruit, aux changements soudains dans la scène et à la présence d'une grande variabilité d'objets mobiles non-rigides [142]. De là, nous avons proposé de combiner le niveau pixel et le niveau région afin de détecter les objets mobiles dans des scènes complexes (arrière-plan non stationnaire + objets mobiles non-rigides) filmées par des caméras librement mobiles. L'approche proposée commence par initialiser le modèle de l'avant-plan par une simple soustraction du fond. Ensuite, une étape de raffinement permet d'identifier, d'une façon coopérative, les objets de l'arrière-plan et ceux de l'avant-plan (tout en discriminant entre les objets mobiles et les ombres), afin de mettre à jour aussi bien le modèle de l'avant-plan que celui de l'arrière-plan.

1.3.1 Initialisation de l'avant-plan

Une fois le modèle initial de l'arrière-plan est produit, une carte de différence est définie pour chaque image afin de localiser grossièrement les objets en mouvement. Dans le cas d'une caméra fixe et de données sans bruit, la carte de différence est tout simplement la différence absolue entre l'image F_i et le fond panoramique S_j. Comme nous visons le cas général d'une caméra en mouvement libre, nous avons commencé d'abord avec un processus de stabilisation pour compenser le mouvement de la caméra entre F_i et S_j. Ensuite, une procédure de réduction de la résolution spatiale des images permet de minimiser le bruit intrinsèque de capture ainsi que les petites régions non significatives qui appartiennent aux traces des objets mobiles. En effet, les objets de l'avant-plan ont tendance à être un peu bruyants à cause des effets d'interaction entre ces objets et le fond (tels que le chevauchement des objets mobiles et les micromouvements des objets de l'arrière-plan). Ainsi, au lieu de soustraire directement l'image alignée RF_i $(= H_{i,j} \cdot F_i)$ de la partie correspondante S_j^i dans S_j, nous avons commencé par redimensionner les deux images à $\frac{1}{r}$ de leurs tailles originales. Puis, l'image de différence D_i entre RF_i et S_j^i est définie dans l'espace RVB (1.16). Cette image est redimensionnée de nouveau à la taille originale de l'image d'entrée F^i en utilisant une interpolation bi-cubique. Ce processus de redimensionnement réduit considérablement les effets du bruit (Fig. 1.17) sans exclure les petits objets significatifs de l'avant-plan (tels que la balle de tennis dans la séquence "Stefan").

$$D_i(x,y) = \max_{c \in \{R,V,B\}} \left| RF_i(x,y,c) - S_j^i(x,y,c) \right|. \tag{1.16}$$

Le seuillage de l'image de différence D_i permet d'obtenir une localisation grossière des silhouettes des objets mobiles (Fig. 1.17). Ce seuillage produit un masque binaire M_i, où 1 indique l'appartenance du pixel à un objet en mouvement. Afin de prendre en compte le degré de corrélation entre les pixels voisins, nous avons utilisé une technique de seuillage par hystérésis [138] dans laquelle chaque pixel est affecté à l'avant-plan ou à l'arrière-plan en fonction de deux seuils T_l et T_h $(T_l < T_h)$. Si la valeur d'un pixel est inférieure à T_l (*resp.* supérieure à T_h), le pixel est affecté à l'arrière-plan (*resp.* à l'avant-plan). Si la valeur du pixel est comprise entre T_l et T_h, le pixel n'est affecté à l'avant-plan que s'il existe un chemin le reliant à un pixel déjà affecté à l'avant-plan (1.17). Néanmoins, la précision de la modélisation de l'avant-plan dépend en partie des seuils fixés empiriquement, et un processus de raffinement supplémentaire est encore nécessaire. En effet, le modèle de l'avant-plan n'est pas encore exact et les limites des objets mobiles débordent souvent leurs limites réelles. En particulier, le bruit

provoque inévitablement quelques petites régions non significatives de l'avant-plan, qui ne correspondent pas à de véritables objets en mouvement.

$$[M_i(x,y)=1] \Longleftrightarrow [D_i(x,y) \geq T_h] \,\text{OU}\, [D_i(x,y) \geq T_l \text{ ET } \exists (x',y') \in C \text{avec} D_i(x',y') \geq T_h], \quad (1.17)$$

où, C est une composante connexe englobant le pixel (x,y) et ceux déjà assignés à 1 dans D_i.

(a) image après compensation du mouvement de la caméra (b) partie correspondante dans l'image du fond (c) masque de l'avant-plan sans redimensionnement (d) masque de l'avant-plan avec redimensionnement

FIGURE 1.17 – Modélisation initiale de l'avant-plan.

1.3.2 Raffinement de l'avant-plan et de l'arrière-plan

Le raffinement conjointe de l'avant-plan et de l'arrière-plan est basé sur la sémantique des régions afin d'améliorer l'exactitude de la séparation entre l'arrière-plan et l'avant-plan. Notons que ce raffinement peut être appliqué itérativement plusieurs fois, si nécessaire, pour les vidéos très complexes [5].

1.3.2.1 Raffinement de l'avant-plan

Pour raffiner au mieux les silhouettes des objets détectés, nous avons utilisé plus d'informations que celles présentes dans l'image de différence. En effet, cette image illustre une fusion des informations de l'image courante et de l'image du fond, ce qui impose le risque de détecter de faux objets mobiles. Comme la segmentation en régions retrouve généralement la forme des objets avec précision, nous avons fusionné les résultats de la détection par soustraction du fond avec ceux fournis par la segmentation spatiale afin de réduire les fausses détections et d'augmenter la précision de la localisation. En effet, le masque binaire M_i, relatif à une image alignée RF_i, est raffiné en utilisant la carte des régions de l'image F_i déjà définie lors de l'étape d'alignement. Afin de compenser le mouvement de la caméra, nous avons commencé par appliquer l'homographie correspondante $H_{i,j}$ sur la carte des régions de l'image F_i, tout en utilisant l'interpolation par plus proche voisin pour traiter les pixels sans valeurs. Ensuite, chaque région R_i^k de la segmentation alignée Seg_i est examinée pour décider si elle appartient à un objet de l'avant-plan ou non (1.18). Si une grande partie de R_i^k est couverte par le masque des objets mobiles dans M_i, la région R_i^k est entièrement déclarée comme une partie d'un objet mobile; sinon toute la région R_i^k est déclarée comme une partie du fond (Fig. 1.18).

$$\left[R_i^k \text{ est une région de l'avant-plant}\right] \Longleftrightarrow \left[\frac{Card(\{(x,y) \in R_i^k / M_i(x,y)=1\})}{Card(R_i^k)} \approx 1 \right]. \quad (1.18)$$

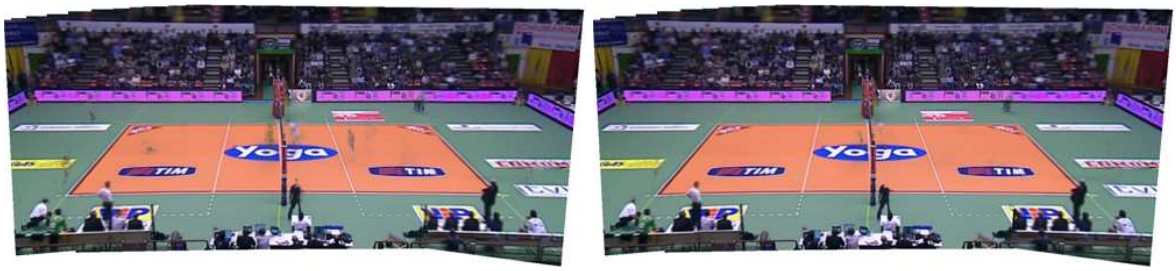
1.3.2.2 Raffinement de l'arrière-plan

Certains parties des objets en mouvements, en particulier ceux qui sont en mouvements rapides et/ou qui sont occultés par d'autres objets, apparaissent partiellement comme traces des fantômes dans

(a) masque initial de l'avant-plan (b) carte des régions (c) masque raffiné de l'avant-plan

FIGURE 1.18 – Raffinement de l'avant-plan.

l'image panoramique produite par la procédure de mosaicing. En effet, comme plusieurs objets peuvent être cachés par d'autres objets en mouvement, il est très habituel qu'une partie du fond n'est jamais, ou partiellement, visible dans la séquence traitée. De là, le modèle de l'arrière-plan doit être mis à jour en considérant les modifications apportées aux différents masques de l'avant-plan. Les pixels étant exclus de (*resp.* intégrés à) un objet de l'avant-plan doivent être attribués au (*resp.* enlevés du) modèle final de l'arrière-plan. Pour cela, le fond est construit de nouveau tout en considérant les "nouveaux" pixels de l'arrière-plan et en écartant ceux qui ont été exclus. Ceci permet de convertir d'une manière efficace des régions non significatives de l'avant-plan vers l'arrière-plan, tout en déplaçant quelques pixels du fond à l'avant-plan (notamment les petits trous) (Fig. 1.19). En raisonnant au niveau région, nous imposons que les pixels connectés et qui suivent le même mouvement seront considérés comme un seul objet. Ainsi, seules les régions, et non pas les pixels, qui différent de l'arrière-plan sont considérées comme des objets mobiles. En effet, les régions de l'avant-plan et celles de l'arrière-plan sont mutuellement mises à jour afin de rejoindre une partie de l'arrière-plan au modèle de l'avant-plan, et vice versa. Cela permet d'améliorer les silhouettes des objets mobiles, tout en éliminant les effets du bruit, sans avoir besoin des opérateurs morphologiques et des filtres de tailles [43]. Néanmoins, les mosaïques obtenues sont souvent caractérisées, surtout autours des objets mobiles, par la présence des effets d'ombrage qui seront éliminés lors de l'étape suivante.



(a) modèle initial de l'arrière-plan (b) modèle final de l'arrière-plan

FIGURE 1.19 – Raffinemet de l'arrière-plan.

1.3.2.3 Traitement de l'ombrage

Le problème des ombres portées est un problème très important liée à l'analyse des vidéos [131]. Pour résoudre ce problème, nous nous sommes basés sur le fait qu'une ombre portée ne modifie par la chrominance de la région passant à l'ombre mais elle diminue, tout simplement, l'intensité lumineuse renvoyée par cette surface. Notre solution consiste à comparer la région supposée être à l'ombre avec celle de l'image du fond, afin de déterminer s'il s'agit d'un objet en mouvement ou simplement d'une ombre portée (1.19). Ainsi, parmi les régions de l'avant-plan, les ombres sont identifiées en fonction de leur apparence à l'égard du modèle de l'arrière-plan dans l'espace couleur LAB, qui sépare explicitement

la luminosité des dimensions de la couleur [171]. Une région R_i^k de l'avant-plan représente un effet d'ombrage si elle a pratiquement les mêmes composantes couleur (a^*, b^*) dans le modèle de l'arrière-plan S_j^i et dans l'image correspondante RF_i. Cependant, elle est beaucoup plus sombre, étant donné la composante de luminosité (L^*) $(\geq \delta_l)$ dans le modèle de l'arrière-plan, avec une certaine limite $(\leq \delta_h)$ pour ne pas confondre les régions mobiles de nature matte avec des effets d'ombrage. Ce traitement permet de réduire les faux positifs de la détection des silhouettes (Fig. 1.20). Le seul inconvénient de ce traitement est qu'il ne fonctionne correctement que sur des zones non achromatiques et il est ainsi difficile de décider si un objet gris sombre apparaissant sur un fond gris clair est réellement un objet ou une ombre portée [3].

$$\left[R_i^k \text{ est un effet d'ombrage}\right] \Longleftrightarrow \left[\begin{array}{c} \sqrt{\sum\limits_{c\in\{a^*,b^*\}} \sum\limits_{(x,y)\in R_i^k} (S_j^i(x,y,c) - RF_i(x,y,c))^2} \approx 0 \\ \text{ET} \\ \delta_l \leq \dfrac{\sum\limits_{(x,y)\in R_i^k} (S_j^i(x,y,L^*) - RF_i(x,y,L^*))}{Card(R_i^k)} \leq \delta_h \end{array}\right]. \tag{1.19}$$

(a) détection des ombres (b) silhouettes finales des objets mobiles

FIGURE 1.20 – Traitement de l'ombrage.

Les résultats obtenus pour plusieurs séquences complexes prouvent l'efficacité de notre approche de création des résumés vidéos [4]. En l'occurrence, la séquence "imansmall", composée de 703 images de taille 60×120, illustre la façon avec laquelle nous traitons le cas des objets mobiles qui s'arrêtent durant un "long" laps de temps sans toute activité intéressante (et deviennent ainsi une partie de l'arrière-plan), qui est une situation très critique. Alors que des fantômes sont souvent détectés dans les zones d'arrêt de ces objets par les techniques classiques de soustraction du fond [47], la solution proposée résout ce problème substantiellement (Fig. 1.21). En effet, l'intégration de l'information région permet de considérer une connaissance sémantique de haut niveau sur les objets en mouvement et de minimiser par conséquence les erreurs de détection. Pour la séquence "volley-ball", composée de 52 images de taille 680×425, l'avant et l'arrière plans sont très texturés et les intervalles couleurs des différents objets (mobiles et stationnaires) chevauchant fortement. Notons aussi le mouvement libre de la caméra pour cette séquence et la nature compliquée des interactions entre les traces spatio-temporelles des joueurs. Les résultats produits prouvent que la solution proposée est bien adaptée aux changements soudains et graduels aussi bien dans l'arrière-plan que dans l'avant-plan. En plus, elle réagit correctement aux effets d'occultation, que ce soit partiel ou complet (il y a plus que 73 événements d'occultation sur la séquence entière). En effet, bien que le modèle initial de l'avant-plan n'est pas assez correcte, la phase du raffinement permet de détecter précisément tous les objets mobiles qui sont hautement articulés, tout en abandonnant les effets d'ombrage (Fig. 1.19, Fig. 1.20). En outre, nous avons évalué objectivement notre approche de détection des objets mobiles. Pour ce faire, nous avons généré manuellement la vérité-terrain d'un échantillon de la séquence de référence "Stefan" (Fig. 1.22), dans laquelle la caméra effectue

un large mouvement panoramique horizontal, un mouvement panoramique vertical et un zoom arrière. Cette séquence est caractérisée aussi par l'articulé du mouvement du joueur, le micromouvement du fond (supporteurs) et la présence du flou de mouvement. Nous avons comparé le pourcentage de classification correcte PCC (1.20) [56] pour les images de l'échantillon, en utilisant respectivement l'approche proposée et une approche standard par soustraction du fond [43]. Il est clair que la solution proposée détecte précisément les silhouettes des objets mobiles (PCC moyen = 0.987) (Fig. 1.23). En plus, bien que le nombre d'objets en mouvement soit faible, les valeurs de la métrique PCC n'ont pas de grandes variations (écart type = 0.004), contrairement aux mesures enregistrées en utilisant l'approche classique sujet de la comparaison [43] (écart type = 0.029). En particulier, malgré leurs petites tailles, la raquette et même la balle ont été détectées correctement dans chaque image de la séquence, quoiqu'elles suivent des mouvements complètement différents de celui du corps de joueur.

$$PCC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{1.20}$$

avec, TP (*resp.* TN) est le nombre de pixels de l'avant-plan (*resp.* de l'arrière-plan) correctement détectés et FP (*resp.* FN) est le nombre de pixels appartenant à l'arrière-plan (*resp.* à l'avant-plan) dans la vérité-terrain et qui sont détectés en tant que pixels de l'avant-plan (*resp.* de l'arrière-plan).

(a) arrière-plan (b) objets mobiles : im122 im370 im691

FIGURE 1.21 – Résumé de la séquence "imansmall".

(a) vérité-terrain : im3 im60 (b) objets détectés : im3 im60

FIGURE 1.22 – Evaluation objective de l'approche proposée sur la séquence "Stefan".

1.4 Application en suivi de multiples personnes

La séparation entre l'avant-plan et l'arrière-plan d'une vidéo constitue une fin en soi pour certaines applications, notamment en compression [170]. Cependant, d'autres applications nécessitent l'extraction des informations de plus haut niveau telles que l'interprétation sémantique des mouvements des objets. Dans ce cadre, nous avons proposé une méthode non supervisée de suivi de multiples personnes, en présence intensive des effets d'occultation, dans une vidéo acquise avec une seule caméra. Pour qu'une méthode de suivi soit complètement automatique, il faut non seulement suivre la ou les cibles mais

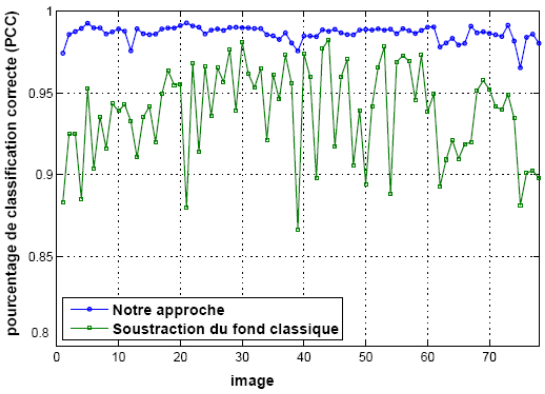


FIGURE 1.23 – Comparaison objective entre l'approche proposée et [43].

aussi les initialiser automatiquement tout en gérant leurs éventuels arrêts et sorties du champ de la caméra. De nombreuses méthodes de suivi ont été proposées, et elles peuvent être regroupées en trois principales classes [154] : suivi de noyaux par détection séquentielle, suivi de contours par segmentation dynamique et suivi par appariement des blobs. Puisqu'on a déjà détecté les blobs qui contiennent les personnes mobiles, nous avons opté à un suivi par appariement de blobs [49]. La première phase de la méthode proposée consiste à séparer les différentes personnes mobiles détectées. En effet, une séparation verticale permet de localiser grossièrement les différentes personnes, qui sont ensuite raffinées d'avantage par la segmentation en régions. La deuxième phase consiste à affecter les blobs détectés à des pistes en utilisant un processus d'appariement basé conjointement sur un modèle d'apparence et un modèle de mouvement.

1.4.1 Détection des personnes

Afin de localiser et séparer les personnes pendant l'occultation ou la juxtaposition, nous avons proposé une méthode basée conjointement sur la forme des silhouettes et la segmentation en régions. Après l'extraction des composants connexes de la carte binaire de l'avant-plan, la première étape consiste à segmenter verticalement les différentes personnes de chaque composante en analysant la forme de sa projection verticale [69]. En effet, chaque blob *Blob* est assumé contenir soit une personne indépendante, soit plusieurs personnes qui s'occultent partiellement [11]. Plus précisément, l'analyse du contour $\mathcal{C}$, de la partie supérieure de *Blob*, et du contour $\mathcal{C}'$, de la partie inférieure de *Blob*, permet de décider si ce blob contient des personnes multiples ou pas. D'ailleurs, sous l'hypothèse que les personnes dans la scène sont dans des attitudes approximativement droites, ce qui est valide dans la plupart des situations pratiques [122], le nombre de personnes correspond en général au nombre des sommets dominants dans la courbe $\mathcal{C}$ (1.21).

$$\forall x' \in [\min_{(x,y)\in Blob} x, \max_{(x,y)\in Blob} x],\ \mathcal{C}(x') = \max(\{y/(x',y) \in Blob\}). \tag{1.21}$$

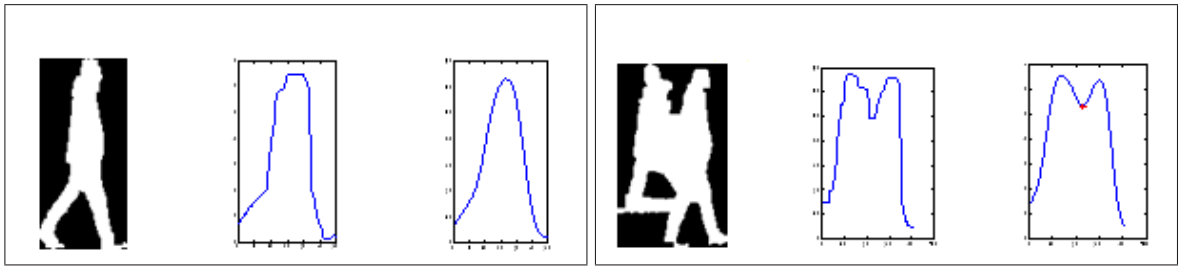
Ainsi, après une suite d'opérations de lissage dans le but d'enlever les extrema locaux de $\mathcal{C}$, nous devons avoir une courbe n-modale dans le cas où le blob est composé de n personnes, et les $n-1$ minimums locaux de la courbe $\mathcal{C}$ définissent les séparateurs des personnes dans le blob en question (Fig. 1.24). Toutefois, l'hypothèse de la correspondance du nombre de pics dominants dans la courbe $\mathcal{C}$ au nombre de personnes, n'est pas correcte dans certains cas. En effet, le fait que la courbe $\mathcal{C}$ est multimodale ne signifie pas nécessairement que le blob *Blob* est composé de plusieurs personnes, vu que

11. Le cas d'occlusion complète est traité implicitement pendant la phase d'association des blobs lors du suivi.

plusieurs pics peuvent appartenir à la même personne mobile, notamment pour les cas d'athlètes très articulés et les personnes détenant des objets dans leurs mains (*e.g.* la raquette dans Fig. 1.25). Par conséquent, il est indispensable de déterminer si le pic dominant est une personne cachée ou une partie de la personne actuelle. Pour ce faire, la courbe $\mathcal{C}'$ de la partie inférieure, représentant les positions verticales minimales en fonction de positions horizontales, est également déterminée. Si la moyenne des pixels constituant la courbe $\mathcal{C}'$ est proche de zéro, alors le pic correspond à un objet caché. Sinon, il s'agit d'une partie de la personne mobile actuelle, qui ne devrait pas être séparée (Fig. 1.25). Ainsi, l'analyse conjointe des deux courbes $\mathcal{C}$ et $\mathcal{C}'$, permet de décider si le blob contient plusieurs personnes [12]. Ensuite, dans le cas de plusieurs personnes occultées, le blob $Blob$ devrait être divisé en n personnes $Blob_i$ ($i \in \{1, \ldots, n\}$) selon l'équation suivante (1.22) :

$$Blob_i = \{(x, y) \in Blob / Min_{i-1} < x \leq Min_i\}, \tag{1.22}$$

avec, Min_i est le i-ème minimum local.



Cas d'une seule personne | Cas de deux personnes partiellement occulteés

FIGURE 1.24 – Détection de l'occlutation : (colonne 1) $Blob$, (colonne 2) la courbe $\mathcal{C}$, (colonne 3) la courbe $\mathcal{C}$ après lissage.

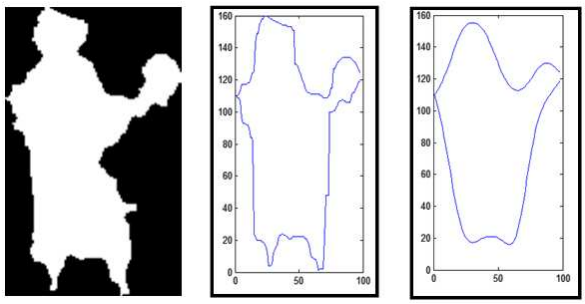


FIGURE 1.25 – Aperçu d'une seule personne tenant un objet : (colonne 1) $Blob$, (colonne 2) le contour du blob englobant les courbes $\mathcal{C}$ et $\mathcal{C}'$, (colonne 3) le contour après lissage.

Après avoir séparé verticalement chaque blob $Blob$ en un ensemble de n personnes $\{Blob_1, \ldots, Blob_n\}$, l'étape suivante consiste à raffiner la silhouette de chaque personne $Blob_i$ [7]. En effet, comme les personnes sont habillées de différentes façons conduisant généralement à un ensemble de régions de différentes couleurs dont la majorité sont alignées verticalement [66], nous avons utilisé l'ensemble des régions $\{R_1, \ldots, R_m\}$ qui composent $Blob$ afin de raffiner les silhouettes des personnes. Pour ce faire, nous avons défini pour chaque région R_j un degré d'appartenance $\mu_{Blob_i}(R_j)$ à chaque blob $Blob_i$ (1.23), afin de distinguer les régions appartenant à une seule personne de celles partagées par plusieurs personnes.

$$\mu_{Blob_i}(R_j) = \frac{Card(\{(x,y) \in R_j / Min_{i-1} < x \leq Min_i\})}{Card(R_j)}. \tag{1.23}$$

Si une région R_j a un recouvrement maximum avec un blob $Blob_i$, alors elle devrait lui être affectée exclusivement. Sinon, la région appartient à la zone de chevauchement entre plusieurs personnes et elle devrait être tout simplement verticalement séparée. Une telle région apparait autour de la ligne verticale relative au minimum local de la courbe $\mathcal{C}$, principalement lorsque les personnes qui s'occultent sont habillées avec les mêmes couleurs. Il paraît ainsi plus exact de garder pour ces régions la séparation verticale initiale. Notons que l'ensemble des régions définissant $Blob_i$ ne sont pas souvent connexes et ceci est principalement dû à la présence de quelques petites régions dans les zones de chevauchement entre les personnes (Fig. 1.26). Pour cela, un post-traitement fusionne chacune de ces petites régions avec la plus grande composante connexe adjacente afin d'assurer la connexité de chaque personne $Blob_i$. Cet algorithme de raffinement de la séparation des personnes peut être résumé comme suit :

Algorithme 2: Raffinement de la séparation des personnes

Entrées : $(Min_0, \ldots, Min_n)$, $\{R_1, \ldots, R_m\}$;
Sorties : $\{Blob_1, \ldots, Blob_n\}$;

début
  $\forall\, i \in \{1, \ldots, n\}$, $Blob_i \leftarrow \emptyset$;
  pour chaque *région* R_j **faire**
    si $\exists\ Blob_i$ / $\mu_{Blob_i}(R_j) \simeq 1$ **alors**
      $Blob_i \leftarrow Blob_i \cup R_j$;
    sinon
      $\forall\, i \in \{1, \ldots, n\}$, $Blob_i \leftarrow Blob_i \cup \{(x,y) \in R_j / Min_{i-1} < x \leq Min_i\}$;

(a) (b) (c) (d) (e) (f)

FIGURE 1.26 – Séparation des personnes : (a) image originale, (b) *Blob*, (c) séparation verticale, (d) segmentation en régions, (e) raffinement de la séparation, (f) post-traitement.

Sachant que l'ensemble des images qui composent une vidéo n'est pas obligatoirement disponible en entier, notamment dans le contexte de la surveillance en ligne des cibles mobiles, nous avons adapté notre approche de détection des personnes mobiles pour un traitement à la volée. En effet, nous avons procédé à l'estimation en ligne des masques des personnes, dans chaque image reçue, par un processus itératif de soustraction de fond. Les tests préliminaires ont montré que la solution proposée permet de détecter correctement plusieurs personnes à la volée dans différentes situations d'occlusion. L'évaluation objective sur la séquence "Stefan" a permis d'enregistrer un pourcentage de classification correcte *PCC* de 0.71, bien qu'elle ne nécessite aucune connaissance a priori. Notons que malgré que la méthode proposée est entièrement en ligne, elle est aussi précise que les méthodes qui nécessitent la disponibilité de tout la séquence étudiée pour pouvoir détecter les objets mobiles [12].

1.4.2 Appariement des blobs

L'objectif de cette phase est d'intégrer un modèle d'apparence avec un modèle de mouvement afin de mettre en correspondance les blobs détectés dans les différentes images. Le modèle du mouvement prédit les positions et les tailles des personnes suivies, ce qui réduit la zone de recherche des correspondants de chaque personne dans les images suivantes. Etant donné que la petite taille des personnes cibles complique considérablement la construction des modèles statistiques, nous avons utilisé le filtre de Kalman [84] tout en y intégrant la position et la taille dans l'état d'un blob suivi. Le but est d'utiliser les mesures observées au fil du temps afin de prédire ces mesures, ainsi que leurs incertitudes, dans les prochaines images, même en présence de bruit. Comme le filtre de Kalman permet la prise en compte de l'erreur de modélisation, nous avons choisi le modèle de mouvement à vitesse constante, en admettant qu'une éventuelle accélération observée serait intégrée au bruit de mesure. L'état d'un blob $Blob_i$ est représenté ainsi par un vecteur $X = (x, y, \dot{x}, \dot{y}, s, \dot{s})^t$, où (x, y) et $(\dot{x}, \dot{y})$ sont successivement les coordonnées et les vitesses en x et en y du centre de gravité du blob, s la taille du blob et $\dot{s}$ la vitesse de s. De là, la relation entre l'état d'un objet à un instant t et celui du même objet à l'instant $t-1$ est défini par (1.24) :

$$\begin{bmatrix} x_t \\ y_t \\ \dot{x}_t \\ \dot{y}_t \\ s_t \\ \dot{s}_t \end{bmatrix} = \begin{bmatrix} x_{t-1} + \dot{x}_{t-1} + w_1 \\ y_{t-1} + \dot{y}_{t-1} + w_2 \\ \dot{x}_{t-1} + w_3 \\ \dot{y}_{t-1} + w_4 \\ s_{t-1} + \dot{s}_{t-1} + w_5 \\ \dot{s}_{t-1} + w_6 \end{bmatrix}, \tag{1.24}$$

où $(w_1, \ldots, w_6)^t$ est le vecteur bruit du système. Ce bruit est modélisé par une Gaussienne à moyenne nulle et une matrice de covariance indépendante. Cette matrice illustre les changements possibles dans le processus de prédiction qui n'ont pas été déjà comptabilisés dans l'état de transition. Une fois les positions et les tailles sont prévues, un modèle d'apparence permet de distinguer les personnes lorsqu'elles sont fusionnées. Pour ce faire, nous avons utilisé l'histogramme couleur dans l'espace LAB tout en quantifiant davantage les composantes chromatiques que la composante achromatique [42]. En plus de l'invariance aux rotations et aux changements d'échelle [115], le succès des méthodes de suivi par histogrammes revient à leur faible complexité associée à une bonne robustesse vis-à-vis du bruit. Ainsi, étant donné l'emplacement (x^i, y^i) et la taille s^i de chaque blob collecté C_i $(\in \Lambda)$ dans l'image courante, le filtre de Kalman prédit la position (x_p^j, y_p^j) et la taille s_p^j de chaque blob prédéfini D_j $(\in \Lambda')$ dans cette image. Ces valeurs sont ensuite combinées avec les histogrammes des couleurs afin d'évaluer la similarité globale $\mathtt{S}$ (1.25) entre chaque couple de blobs (C_I, D_j) dans $\Lambda \times \Lambda'$ [12]. Enfin, le problème d'association des blobs est équivalent à un problème d'affectation linéaire plusieurs-à-plusieurs (many-to-many), où l'association choisie est celle qui minimise un coût d'affectation sur l'ensemble des appariements entre les différents blobs. Ceci revient à trouver parmi l'ensemble Λ des blobs perçus C_i, ceux qui correspondent à des blobs déjà connus D_j $(\in \Lambda')$. Cette association doit vérifier la contrainte d'unicité, tout en considérant qu'un blob peut apparaître ou disparaître. En effet, si un blob perçu ne peut être associé à aucun objet connu, alors il s'agit d'un blob nouvellement apparu. Dans le cas inverse, si un blob connu n'est associé à aucun des blobs perçus, alors cet objet a disparu (il peut être masqué par un autre objet, hors de la portée des capteurs...). L'association des blobs de l'image courante avec ceux de l'image précédente est réalisée en premier lieu moyennant l'algorithme "Shortest Augmenting Path" (SAP) [83]. Ensuite, les blobs non-assignés dans l'image courante peuvent être soit associés à ceux qui étaient perdus dans les images précédentes, soit considérés comme des nouveaux objets.

$$\mathtt{S}(C_i, D_j) = \left(\frac{\min(s^i, s_p^j)}{\max(s^i, s_p^j)}\right)\left(\frac{1}{2}\left[\frac{\min(x^i, x_p^j)}{\max(x^i, x_p^j)} + \frac{\min(y^i, y_p^j)}{\max(y^i, y_p^j)}\right]\right)\left(\frac{\sum_{k=1}^{M} \min(h_{C_i}(k), h_{D_j}(k))}{\min(\sum_{k=1}^{M} h_{C_i}(k), \sum_{k=1}^{M} h_{D_j}(k))}\right), \tag{1.25}$$

12. Nous avons utilisé l'intersection des histogrammes pour comparer les histogrammes des couleurs [30].

où, $h_B(k)$ est la k-ème composante de l'histogramme h_B associée à un blob B.

Afin de valider notre méthode de suivi de multiples personnes, nous l'avons testé sur plusieurs séquences standards. Toutes ces séquences ont été acquises avec une seule caméra mobile et les personnes suivies sont habillées avec des couleurs similaires. En plus, nous avons procédé à un échantillonnage des images de chaque séquence afin d'évaluer la validité de cette méthode pour des vidéos de faible résolution temporelle ($\sim$ 7 images par seconde). Fig. 1.27 illustre les résultats produits pour la séquence "PETS09.S2L1" dans laquelle trois personnes traversent indépendamment une scène extérieure avant qu'elles interagissent. En particulier, personne3 a été occultée par personne2, partiellement puis totalement, entre les images 44 et 47, avant d'être obstruée par personne1 entre les images 49 et 51. La solution proposée détecte et reconnaît correctement le mouvement des différentes personnes même quand elles suivent des directions distinctes. Par ailleurs, les deux personnes qui apparaissent plus tard dans la scène ont été identifiées comme une seule personne (personne4), vu qu'elles illustrent une occlusion totale pendant tout le reste de la séquence. Un échantillon de cette séquence standard a également été utilisé pour une évaluation objective de la méthode proposée. Nous avons généré manuellement la vérité-terrain (Fig. 1.28), tout en prédisant la silhouette complète de chaque personne dans les cas d'occlusions (*i.e.* un pixel peut appartenir à deux personnes quand elles s'occultent), et nous avons mesuré la distance Euclidienne entre chaque blob de la vérité-terrain et le blob détecté par notre méthode (Tab. 1.2). La moyenne de la distance pour les quatre personnes suivies est d'environ 3.12, sachant que la taille moyenne de ces personnes est d'environ 4804 pixels. Cette valeur est très encourageante en prenant en considération les nombreux événements d'occlusion complexe qui surviennent dans cette séquence, et que, contrairement à la vérité-terrain, chaque pixel est associé exclusivement à une seule personne par notre méthode.

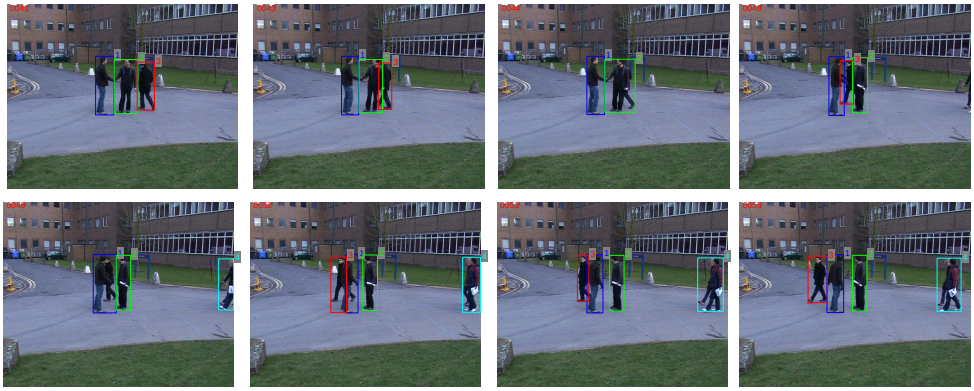


FIGURE 1.27 – Echantillon des résultats sur la séquence "PETS09.S2L1" (entre l'image 41 et l'image 55).

(a) blobs détectées (b) vérité-terrain : personne1 personne2 personne3 personne4

FIGURE 1.28 – Evaluation objective de la méthode proposée pour le suivi des personnes.

Par ailleurs, et à des fins de comparaison, nous avons appliqué la méthode proposée sur les séquences utilisées dans [162], où les auteurs ont comparé leur méthode ("BATracker») par rapport à trois autres

image	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55
personne1	1.40	0.85	1.13	4.07	1.38	1.33	0.23	1.61	9.84	4.68	5.73	5.07	0.51	0.72	1.84
personne2	4.91	8.06	8.68	7.11	5.54	1.36	7.09	0.96	3.53	1.39	1.05	0.97	1.78	0.51	1.06
personne3	0.79	4.05	1.85					9.07				5.88	2.20	1.43	0.09
personne4									1.42	0.76	3.25	1.37	4.33	3.00	6.02

TABLE 1.2 – Evaluation objective de la méthode proposée sur la séquence "PETS09.S2L1".

méthodes ("Meanshift" [44], "Template Matching" (TM) [132] et "MILTracker" [14]). Bien qu'elle ne nécessite aucune connaissance a priori, notre méthode surpasse clairement, pour ces séquences, l'ensemble des méthodes comparées (Fig. 1.29). En particulier, elle permet de suivre correctement toutes les personnes même si une occlusion totale se produit depuis le début de la séquence [6]. En effet, en raison de la mauvaise mise à jour de l'information au cours des effets d'occlusion complexe, les méthodes "Meanshift", "Template Matching" et "MILTracker" ne sont pas en mesure de suivre correctement toutes les cibles le long de la durée des occlusions. La méthode "BATracker" suit avec succès les différentes personnes, mais sans grande précision comparativement à notre solution (*c.f.* personne3). En l'occurrence, la distance Euclidienne enregistrée pour personne3 dans l'image 124, en utilisant la vérité-terrain produite dans [162], est de 2.23 avec notre méthode et de 19.72 avec "BATracker".

(a) MILTracker (b) Meanshift (c) TM (d) BATracker (e) notre méthode

FIGURE 1.29 – Echantillon des résultats produits sur la séquence "Seq2" de [162].

1.5 Conclusion

Nous avons synthétisé le long de ce chapitre nos travaux de recherche sur le résumé des vidéos complexes basé conjointement sur la génération des multiples mosaïques et la détection des objets mobiles. L'objectif est de séparer l'avant-plan d'une vidéo de son arrière-plan, qui est une tâche très compliquée en raison du mouvement aléatoire de la caméra et de la présence de plusieurs objets mobiles non-rigides. Pour cela, nous avons commencé par introduire une approche multi-primitive d'alignement d'images, qui permet de compenser d'une manière entièrement non-supervisée le mouvement de la caméra entre toute image de la séquence et le plan d'une seule image choisie comme référence. L'originalité principale de

cette approche réside dans l'utilisation de l'appariement des régions afin de pré-estimer le mouvement de la caméra (rotation et facteur d'échelle) et de limiter par la suite l'espace de recherche des homologues potentiels lors de l'appariement des points d'intérêt. Une évaluation objective de cette approche a permis de prouver sa robustesse vis-à-vis du mouvement de la caméra, de la variation d'illumination, du bruit d'acquisition et de la présence d'objets mobiles. Notons que même sous l'hypothèse de planarité de la scène, l'approche proposée est partiellement invariante au changement du point de vue. Toutefois, dans le cas de large changement du point de vue, il serait impossible d'aligner les images sans une reconstruction 3D dense. Ensuite, nous avons proposé de résumer en ligne le contenu visuel de l'arrière-plan en multiples mosaïques. Les limites de chaque mosaïque, ainsi que son image de référence, sont détectées à la volé en utilisant une estimation robuste des paramètres de la caméra. Pour ce faire, les angles de rotation de la caméra et la distance focale, pour chaque image reçue, sont estimés à partir de l'homographie image-mosaïque. Ces paramètres feront ensuite l'entrée d'une procédure de décision d'affectation qui permet de confirmer si l'image peut être projetée et mélangée sans distorsions sur l'une des mosaïques qui ont été déjà construites jusqu'à cet instant. L'approche proposée est capable d'associer des images de différentes positions temporelles à une même mosaïque, même si la caméra retourne sur une partie précédemment visitée de l'arrière-plan. De nombreuses expérimentations réalisées sur des vidéos de scènes d'intérieur et d'extérieur enregistrées avec des caméras mobiles, démontrent l'efficacité l'approche proposée. En effet, cette approche résume en ligne le contenu visuelle d'une longue séquence vidéo tout en empêchant l'accumulation des erreurs d'alignement et en maintenant des exigences minimales en termes d'espace mémoire.

Une fois les mosaïques sont estimées, nous avons proposé un schéma de détection des objets en mouvement complexe (rapidité, déformation, occultation, ombrage...). Après avoir aligné chaque image de la séquence vidéo sur le plan de référence de la mosaïque correspondante, l'idée de base du schéma proposé est la construction d'un modèle précis de l'avant-plan à partir d'une approximation préliminaire de ce modèle. Ainsi, des masques grossiers représentant les objets en mouvement sont estimés en comparant chaque image alignée avec la partie correspondante dans l'image du panorama. Ces masques sont ensuite raffinés par les segmentations spatiales déjà définies lors de la phase d'alignement. Ceci permet de convertir d'une manière efficace des régions non significatives de l'avant-plan vers l'arrière-plan, et vice versa. Cette approche, qui utilise conjointement les informations pixel et région afin de surmonter les inconvénients des techniques basées sur la soustraction du fond, est capable d'extraire automatiquement avec une grande précision les objets mobiles dans chaque image de la vidéo, sans que l'arrière-plan, qui peut être non-uniforme, ne soit calculé plusieurs fois. La contribution de la solution proposée est importante surtout que la plupart des défis sont comptabilisés d'une manière gracieuse. Pour le meilleur de nos connaissances, notre solution est beaucoup moins contraignante que la majorité des solutions proposées [62]. En effet, nous ne disposons que de l'hypothèse de planarité de la scène, qui est vérifiée dans la plupart des séquences vidéos, tout en assumant que les modèles des objets de l'avant-plan et de l'arrière-plan ainsi que leurs éventuels mouvements sont inconnus. Néanmoins, l'approche proposée échoue dans le cas où les objets mobiles se déplacent dans des directions totalement différentes (tels que les déplacements en zigzag), puisque nous avons supposé que les objets en mouvement saillante se déplacent dans une direction cohérente pendant une période du temps. L'approche proposée pour la détection des objets mobiles a été validée en suivi de multiples personnes mobiles en présence d'occultations. Nous avons commencé par séparer les différentes personnes mobiles détectées, en se basant conjointement sur la forme des silhouettes et la segmentation en régions. Ensuite, les différents blobs obtenus sont affectés à des pistes en utilisant un processus d'appariement intégrant à la fois un modèle d'apparence et un modèle de mouvement. La méthode proposée est capable de suivre correctement, avec une seule caméra et même à faible résolution temporelle, plusieurs personnes mobiles dans des situations d'occlusion complexe, sans connaissance préalable sur le nombre de personnes suivies ni une initialisation de leurs positions. En effet, une étude comparative a confirmé que cette méthode de suivi de multiples personnes est caractérisée par une précision meilleure par rapport à des solutions standards. Toutefois, cette méthode risque d'échouer dans le cas de scènes de foule d'une ampleur significative, comme les aéroports et les centres commerciaux, à cause du flux massif de personnes dont les mouvements sont souvent aléatoires et caractérisés par plusieurs changements brusques de directions.

Chapitre 2

Recherche des images par le contenu

2.1 Introduction

La quantité immense des images qui ne cesse d'évoluer, surtout avec l'apparition d'une panoplie d'appareils d'acquisition simples et à coût raisonnable, exige le besoin de proposer des systèmes automatiques d'indexation et de recherche des images par le contenu visuel (CBIR, pour Content-Based Image Retrieval). Historiquement, ces systèmes ne présentaient pas la première solution pour rechercher des images. Le début était avec les systèmes de recherche à base de descripteurs textuels (TBIR, pour Text-Based Image Retrieval), où le système recherche les images dont l'annotation textuelle est la plus similaire à celle donnée par l'utilisateur. Cette indexation représente une tâche subjective, longue et répétitive pour le gestionnaire des bases [35]. Ainsi, dans le but de réduire le fossé entre ce que l'utilisateur désire avoir et ce que le système lui propose comme résultats, l'idée des CBIR s'avère plus prometteuse. En effet, plusieurs approches basées sur la description de bas niveau du contenu de l'image ont été proposées afin de garantir plus d'objectivité et d'imiter au mieux le système perceptuel humain. Cependant, elles sont toujours peu performantes en ce qui concerne la recherche des images dans des bases généralistes. Ceci est essentiellement dû au fossé sémantique entre l'information qu'on peut extraire des données visuelles (de bas niveau) et l'interprétation (de haut niveau) que les mêmes données ont pour un utilisateur dans une situation donnée. Pour remédier à ce problème, deux solutions ont été présentées : la recherche des images à base du contenu des régions (RBIR, pour Region-Based Image Retrieval) et le bouclage de pertinence (BP). La première solution vise à éviter l'aspect global de la description du contenu de l'image en se basant sur des descripteurs locaux. Cette solution présente deux variantes : l'une basée sur la décomposition de l'image en blocs et l'autre basée sur la segmentation en régions. Toutefois, la première variante provoque toujours le problème de fossé sémantique, vu que les blocs ne coïncident pas forcément avec des objets réels. Concernant le RBIR basé sur la segmentation, on commence par l'extraction des objets d'intérêt qui guideront par la suite le processus d'indexation et celui de recherche. Pour la deuxième solution, l'utilisateur a la possibilité d'exprimer plus précisément son besoin en raffinant sa requête par bouclage de pertinence. Ceci permet de s'approcher plus de l'intention de l'utilisateur afin de diminuer le taux du fossé sémantique. L'idée d'optimiser les résultats de recherche des systèmes CBIR en utilisant le BP existe déjà, mais l'incorporation du BP en RBIR est beaucoup moins répandue. Ceci est principalement dû au fait que le nombre des régions varie souvent d'une image à une autre, ce qui complique la représentation des différentes images dans un espace uniforme de description [10]. Dans ce contexte, nous avons proposé de combiner conjointement les descripteurs régions, voire objets, et les informations fournies par l'utilisateur interactivement. En effet, nous visons à améliorer les résultats de la recherche en s'approchant au maximum de l'intention de l'utilisateur tout en se basant sur des descripteurs physiques de haut niveau (régions) de l'image. Le premier niveau d'interaction avec l'utilisateur correspond à la présentation de l'image requête[1]. Cette image est ensuite segmentée en régions par une technique floue afin d'obtenir des objets grossièrement

1. Nous avons évité les régions requêtes afin de simplifier l'interface et les interactions avec l'utilisateur [79].

définis. L'ensemble des régions obtenues sera modélisé par la suite via un graphe orienté représentant la signature de l'image où chaque nœud est étiqueté avec les caractéristiques de bas niveau, à base de descripteurs de sous-bandes d'ondelettes, d'une région particulière et il est pondéré par une valeur floue qui illustre l'importance visuelle de cette région au sein de l'image. En outre, l'information spatiale est incorporée dans la structure du graphe en caractérisant chaque arête entre deux régions par deux quintuplets illustrant les relations spatiales entre ces régions. Notre modèle de graphe est totalement basé sur la logique floue afin de mieux modéliser l'incertitude et l'ambiguïté de l'étape de segmentation [80]. Puis, le graphe, extrait en-ligne de l'image requête, est comparé avec les graphes, extraits hors-ligne, relatifs aux images composants la base en utilisant une technique de complexité efficace. En effet, la mise en correspondance des images revient à l'appariement des graphes correspondants, qui est un problème difficile. Pour cette raison, nous avons proposé une approche d'appariement basée sur des heuristiques garantissant un compromis entre la qualité et le temps de calcul. Par ailleurs, en cas d'insatisfaction de l'utilisateur des résultats affichés, il peut effectuer des itérations de BP afin de se rapprocher au mieux de son besoin. En effet, nous avons introduit un mécanisme interactif qui se base sur l'adaptation des poids de régions de l'image requête en fonction des rétroactions de l'utilisateur. Ce bouclage de pertinence est conçu selon la caractéristique de la représentation à base de régions décrites par des sous-bandes d'ondelettes. Il est basé sur une adaptation empirique des poids des régions, pareillement aux techniques utilisées pour la recherche basée sur le texte [58]. L'étude comparative avec des approches similaires a prouvé la robustesse de l'approche proposée pour l'indexation et la recherche des images en termes d'apport sémantique offert par la modélisation riche des images par des graphes complets ainsi que par le bouclage de pertinence. Enfin, partant de l'hypothèse que la co-segmentation peut contribuer à la recherche des images par le contenu, notamment pendant les itérations de bouclage de pertinence, nous avons commencé à explorer cet axe de recherche. La plupart des techniques existantes modélisent la co-segmentation sous la forme d'un problème d'optimisation qui cherche à minimiser une fonction d'énergie. Cette fonction considère un terme de correspondance qui pénalise la dissimilarité entre les images à co-segmenter. Pour évaluer cette correspondance, les techniques existantes comparent les histogrammes en absence de toute information de cohérence spatiale. Pour cela, nous avons proposé d'intégrer l'information spatiale afin d'éviter les fausses détections et les effets du bruit. En effet, en plus de l'utilisation de l'entropie locale lors de la caractérisation d'une image par son histogramme, notre principale contribution réside dans la classification floue de l'entropie locale, ce qui permet de réduire l'ambiguïté d'appartenance d'un pixel à un bin de l'histogramme. Les résultats préliminaires ont prouvé l'efficacité de la technique proposée pour la co-segmentation des images. Cette technique a été adaptée pour la segmentation des athlètes dans une grande variété de poses. L'idée est de minimiser la complexité de cette tâche de segmentation, et ceci en la réduisant à une simple co-segmentation de paires d'images, afin d'en extraire dans chacune les objets communs dans des environnements sans contrainte et sans aucune intervention de l'utilisateur. La solution proposée a été appliquée sur diverses vidéos d'activités sportives et les résultats préliminaires sont très encourageants.

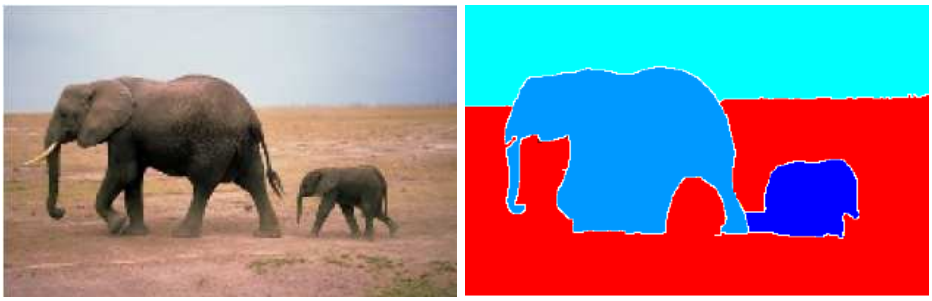
2.2 Définition de la signature d'une image

Nous avons utilisé la notion de graphe afin de modéliser le contenu visuel des objets d'une image ainsi que les relations spatiales entre ces objets. La modélisation proposée, qui est entièrement basée sur la logique floue, incorpore autant d'informations expressives et distinctives que possible. Fondé sur l'hypothèse que n'importe quelle région pourrait être utile dans le procédé de recherche [161], toutes les régions de chaque image sont considérées. De là, un graphe complet est défini relativement à chaque image, où chaque nœud représente une région floue grossièrement segmentée et il est mesuré en termes de deux propriétés : les descripteurs de bas niveau de la région et son poids évaluant son importance visuelle au sein de l'image. En outre, les relations spatiales entre les régions sont également considérées, au sein du graphe signature d'une image, en termes de cinq mesures floues. Ainsi, la structure de graphe utilisée récapitule aussi maximum que possible la teneur visuelle d'une image avec des descripteurs entièrement flous. De cette façon, plus d'informations peuvent être fournies pour comparer des images. Formellement, le graphe d'une image I, qui est un graphe relationnel attribué, est défini par un quadruplet $G =$

$(\Re, \mathcal{A}, \Im_{\Re}, \Im_{\mathcal{A}})$, où $\Re$ est un ensemble fini des nœuds-régions, $\mathcal{A}$ ($\subseteq \Re \times \Re$) est un ensemble fini des arêtes entre les nœuds, $\Im_{\Re}$ est la fonction caractérisant les attributs des nœuds et $\Im_{\mathcal{A}}$ est celle définissant les attributs des arêtes. En effet, chaque nœud dans le graphe caractérise une région floue de l'image et chaque arête représente les relations spatiales entre deux régions qui ne sont pas forcément adjacentes. Les attributs relatifs aux nœuds définissent le contenu visuel de la région correspondante, en termes d'un ensemble de taille réduite de descripteurs de bas niveau à base de sous-bandes d'ondelettes, ainsi que son importance sémantique au sein de l'image. Les attributs d'une arête caractérisent, en termes de descripteurs flous, les relations spatiales entre les régions reliées par cette arête.

2.2.1 Extraction et quantification floue des régions

Pour partitionner une image I en un ensemble de régions floues, nous avons utilisé une technique floue de segmentation grossière à base de l'algorithme de ligne de partage des eaux (LPE). En plus, un post-traitement d'accroissement des petites régions adjacentes permet de surmonter les effets de sur-segmentation accompagnant souvent le LPE [121]. La raison principale derrière le choix du LPE est que cette technique peut être appliquée efficacement sur n'importe quel type d'image, ce qui la présente comme un choix adéquat pour des traitements sur des bases d'images généralistes. L'algorithme utilisé permet d'extraire à partir de chaque image un nombre réduit de régions floues qui représentent grossièrement les objets sémantiques composant l'image (Fig. 2.1), ce qui permet d'optimiser par la suite la qualité et le temps de calcul de la mise en correspondance et de la recherche. En effet, ces régions floues sont les unités de base des traitements dans nos approches de RBIR et de BP. En particulier, lors de la quantification d'une région, chaque pixel de la région sera pondéré avec son degré d'appartenance à cette région. Ceci permet de favoriser les pixels au centre de la région, par rapport à ceux sur les bords, lors de l'indexation d'une image, ce qui assure par la suite la robustesse de la procédure de recherche contre les éventuelles erreurs de segmentation.



(a) image originale (b) image segmentée

FIGURE 2.1 – Segmentation en régions d'une image.

Ensuite, chaque région est décrite en termes de son apparence visuelle et de son importance sémantique au sein de l'image. D'une part, l'apparence visuelle d'une région est rapprochée par des primitives de bas niveau. Pour ce faire, la majorité des systèmes RBIR sont basés sur la caractérisation des régions par des descripteurs couleurs [81]. Toutefois, en plus du problème de choix du bon espace couleur (entre RGB et Lab essentiellement) dans le cas des bases généralistes, il est souvent nécessaire de renforcer par d'autres descripteurs ce qui impose la réduction de l'espace de description [79]. En effet, les tests réalisés nous ont permis de déduire que chaque espace présente des points forts dans un certain contexte (*e.g.* le benchmark utilisé) et pas dans l'autre [64]. Par exemple, les espaces HSV et HSL ne sont pas uniformes donc l'utilisation de la distance Euclidienne perd sa signification pour ces deux espaces. D'autres espaces couleurs tels que Lab, Luv, LCH et Ltθ, qui sont des dérivations uniformes des espaces HSV et HSL, nécessitent une transformation à partir de l'espace RGB. De ce fait, nous risquons de surcharger la procédure d'indexation et de recherche dans le cas où nous devons calculer ces transformations ou nous utilisons une combinaison d'espaces couleur afin de combler toutes les défaillances. De là, nous

avons choisi d'unifier les données et les descripteurs des régions tout en prenant en compte l'information spectrale multi-résolution représentée par les ondelettes [86]. Cependant, les bandes d'ondelettes portent généralement sur l'image entière, et pour chaque niveau de fréquence, une compression spécifique par un ensemble de filtres est effectuée sur toute l'image. Ainsi, et afin d'adapter les ondelettes au contexte des régions, les techniques existantes sont basées sur la représentation de chaque région R par son rectangle circonscrit RC_R [64], ce qui permet de maintenir la représentation matricielle. Toutefois, le problème majeur de cette représentation réside dans le traitement des rectangles qui se chevauchent. Dans notre cas, nous avons pondéré chaque pixel (x, y) par son degré d'appartenance $\mu_R(x, y)$ à la région R à laquelle il est affecté (2.1). De là, les pixels des autres régions et qui sont encadrés par le rectangle circonscrit RC_R auraient un degré d'appartenance nul à la région R ce qui permet de les ignorer lors de la transformation en ondelettes sur R. La quantification d'une région est ainsi focalisée uniquement sur l'objet d'intérêt. En plus de l'intégration de l'information texture, les ondelettes permettent de garder l'allure générale des couleurs des objets d'une manière grossière, sans tenir compte des petits changements visibles sur un nombre limité de pixels. De plus, les données fréquentielles qui lui sont relatives sont au même temps distinctives [73]. Ainsi, il est possible de ne pas utiliser toute l'information spectrale mais de se contenter seulement d'une partie d'elle, ce qui permet en particulier de réduire le nombre des descripteurs. Dans notre cas, nous nous sommes limités à la sous-bande de haute résolution de niveau 2. Cette restriction est justifiée par le fait que cette bande comporte généralement la globalité de l'information utile de la région comme le montre les résultats obtenus lors de la recherche. D'ailleurs, nous avons trouvé presque les mêmes résultats de recherche avec l'information entière des ondelettes mais en augmentant le temps de calcul par 85.31% [63]. En outre, l'utilisation des bandes de basse fréquence, qui décrivent l'allure des objets détectés, peut provoquer de fausses correspondances, notamment pour les objets naturels. Ceci est principalement dû à l'instabilité de la forme pour ces objets surtout dans le cas de larges bases d'images[2]. Cependant, lors du bouclage de pertinence, les exemples non-pertinents seront éliminés et seules les images qui gardent éventuellement une liaison sémantique avec l'image requête seront considérées lors des recherches ultérieures. De là, l'élargissement de la bande descriptive des régions lors du bouclage de pertinence permet de mieux caractériser les images cibles tout en évitant le risque de faire correspondre des images sémantiquement différentes mais qui contiennent des objets avec des silhouettes similaires (*c.f.* Section 2.5). Ainsi, l'apparence visuelle $\Im_{\Re}(R)$ d'une région R ($\in \Re$) est décrite par une matrice spectrale creuse (au niveau des pixels appartenant aux autres régions englobées par RC_R) M^R_{HH}, qui illustre l'information spectrale fournie par la sous-bande HH de niveau 2 des ondelettes appliquées sur les composantes couleurs de R (2.1).

$$\forall R \in \Re,\ \forall (x,y) \in RC_R,\quad M^R_{HH}(x,y) = \mu_R(x,y).[H_x * [H_y * [H_x * [H_y * I(x,y)]_{\downarrow 2,1}]_{\downarrow 1,2}]_{\downarrow 2,1}]_{\downarrow 1,2}, \quad (2.1)$$

où, $*$ indique l'opérateur de convolution, $H = (H_x, H_y)^t$ est un filtre Canny passe-haut et '$\downarrow 2,1$' (*resp.* '$\downarrow 1,2$') désigne l'échantillonnage vers le bas le long des lignes (*resp.* des colonnes).

D'autre part, l'importance visuelle de chaque région R dans une image I (de dimension M×N) est évaluée par son poids w_R (2.2) qui est la moyenne de deux facteurs flous. Un premier facteur d'emplacement spatial w^e_R reflète le processus de la perception visuelle humaine qui accorde plus d'importance aux régions localisées au centre d'attention. Ce degré de centralisation s'exprime sous la forme de la distance entre le centre G_I de l'image I et le centre de gravité G_R de la région R, et ceci tout en pondérant chaque pixel de R par son degré d'appartenance. En outre, chaque région R est pondérée selon son pourcentage de surface w^s_R, illustrant son importance dans l'image correspondante relativement au système visuel humain [111]. Les poids de l'ensemble des régions d'une image ont été normalisés afin que leur somme soit bien égale à 1 (Fig. 2.3). Ces poids peuvent être ensuite mis à jour d'une manière adaptative pendant la phase de bouclage de pertinence en fonction des suggestions de l'utilisateur. Notons que les différentes valeurs utilisées pour la quantification des régions (descripteurs + poids) sont toutes floues, ce qui permet de minimiser la propagation de l'erreur et ceci en modélisant toujours l'incertitude d'affectation d'un pixel à une région unique.

2. Pour cette raison, nous avons évité les attributs de forme surtout que notre segmentation est grossière.

$$\forall R \in \Re, \quad w_R = \frac{1}{2}.\left(\underbrace{\frac{1}{N.M}.\sum_{(x,y)\in R} \mu_R(x,y).d^{-1}(\begin{pmatrix} x \\ y \end{pmatrix}, G_I)}_{w_R^e} + \underbrace{\frac{1}{N.M}.\sum_{(x,y)\in R} \mu_R(x,y)}_{w_R^s} \right). \tag{2.2}$$

2.2.2 Relations spatiales entre les régions

Chaque arête entre deux nœud-régions est marquée par deux quintuplets quantifiant les relations spatiales entre ces deux régions. En effet, la position relative d'une région par rapport à une autre est évaluée en termes d'une valeur illustrant le degré d'inclusion (relation topologique) et de quatre valeurs qui estiment les relations directionnelles : "au-dessus de", "en-dessous de", "à gauche de" et "à droite de". Notons que, pareille à la quantification des régions, nous avons aussi utilisé la logique floue pour définir les relations spatiales entre les régions afin de surmonter l'imprécision typique de la représentation des descripteurs. En particulier, la logique floue permet un certain degré de variation des valeurs des descripteurs, ce qui améliore la robustesse et l'efficacité de l'indexation et de la recherche. Ceci s'explique par le fait que cette logique traite les cas d'incertitude et offre une marge d'imprécision ce qui réduit l'effet des erreurs de la segmentation. Ainsi, chaque arête (R_1, R_2) $(\in \mathcal{A})$ connectant deux régions R_1 et R_2, non forcement adjacentes, est labellisée par deux quintuplets, $\Im_{\mathcal{A}}(R_1, R_2)$ et $\Im_{\mathcal{A}}(R_2, R_1)$, décrivant les degrés de positionnement et d'inclusion de R_1 par rapport à R_2, et vice versa. En effet, le quintuplet décrivant les relations spatiales entre R_1 et R_2 est défini par quatre mesures reflétant la position floue de R_1 relativement à R_2 dans l'une des directions suivantes : "à gauche de" $\Im_{\mathcal{A}}(R_1 \leftarrow R_2)$ (2.3) (Fig. 2.2), "à droite de" $\Im_{\mathcal{A}}(R_1 \rightarrow R_2)$ (2.4), "en-dessous de" $\Im_{\mathcal{A}}(R_1 \downarrow R_2)$ (2.5) et "au-dessus de" $\Im_{\mathcal{A}}(R_1 \uparrow R_2)$ (2.6) ; et par une valeur floue $\Im_{\mathcal{A}}(R_1 \subset R_2)$ (2.7) indiquant le degré d'inclusion ("à l'intérieur de") de R_1 dans R_2.

$$\Im_{\mathcal{A}}(R_1 \leftarrow R_2) = \frac{1}{Card(R_1)}.Card(\{(x,y) \in R_1 / x \in [\min_{(x',y')\in R_1} x', \min_{(x',y')\in R_2} x']\}). \tag{2.3}$$

$$\Im_{\mathcal{A}}(R_1 \rightarrow R_2) = \frac{1}{Card(R_1)}.Card(\{(x,y) \in R_1 / x \in [\max_{(x',y')\in R_2} x', \max_{(x',y')\in R_1} x']\}). \tag{2.4}$$

$$\Im_{\mathcal{A}}(R_1 \downarrow R_2) = \frac{1}{Card(R_1)}.Card(\{(x,y) \in R_1 / y \in [\max_{(x',y')\in R_2} y', \max_{(x',y')\in R_1} y']\}). \tag{2.5}$$

$$\Im_{\mathcal{A}}(R_1 \uparrow R_2) = \frac{1}{Card(R_1)}.Card(\{(x,y) \in R_1 / y \in [\min_{(x',y')\in R_2} y', \min_{(x',y')\in R_1} y']\}). \tag{2.6}$$

$$\Im_{\mathcal{A}}(R_1 \subset R_2) = \frac{1}{Card(R_1)}.Card(\{(x,y) \in R_1 \cap [\min_{(x',y')\in R_2} x', \max_{(x',y')\in R_2} x'] \times [\min_{(x',y')\in R_2} y', \max_{(x',y')\in R_2} y']\}). \tag{2.7}$$

Ces relations spatiales diffèrent de celles d'Allen [36] qui sont déterministes (une région est soit à droite, à gauche, au-dessus ou en-dessous d'une autre région). Notre façon d'évaluer les relations spatiales entre deux régions vient de palier au problème souvent rencontré lors de tout processus d'appariement des régions basé sur les graphes d'adjacences et les positions relatives des régions. En effet, dans la plupart des cas, les relations d'Allen ne permettent pas de définir proprement les relations spatiales

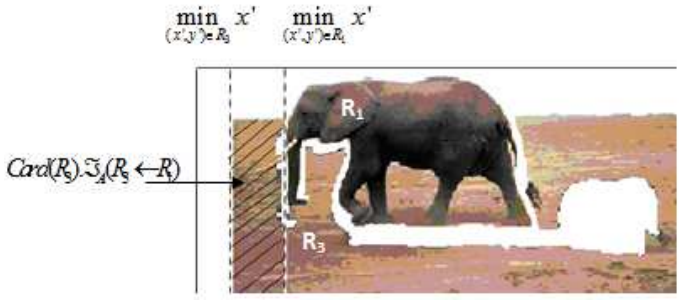


FIGURE 2.2 – Evaluation floue de la position relative "à gauche de" de la région R_3 par rapport à la région R_1.

entre deux régions, vu qu'elles le font d'une manière stricte ce qui n'est pas totalement vrai. Cependant, nos relations spatiales permettent de produire une certaine probabilité qui évalue la position, même partielle, d'une région par rapport à une autre [168]. Ainsi, nous obtenons la signature complète d'une image sous forme d'une combinaison entre la signature de chaque région et l'information de position relative de chaque région par rapport à toutes les autres régions (Fig. 2.3). Cette signature est modélisée via une matrice carrée où les lignes et les colonnes sont les régions R_i $(1 \leq i \leq Card(\Re))$ de l'image dans un ordre décroissant des poids w_i. Les informations relatives à la signature de l'image sont présentées au niveau des cases de cette matrice où celles en diagonale comportent les descripteurs bas niveau $M_{HH}^{R_i}$ des régions alors que les autres cases contiennent les descripteurs des arêtes (Tab. 2.1).

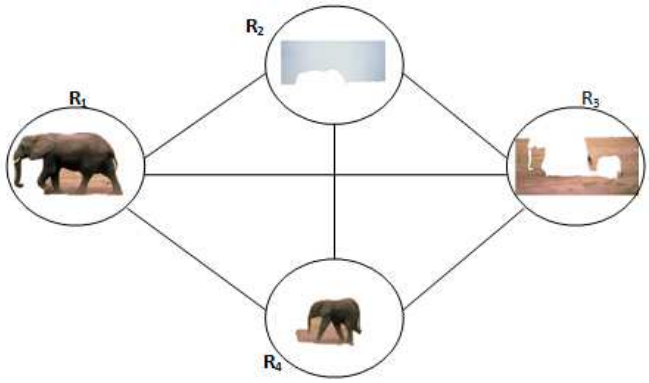


FIGURE 2.3 – Graphe signature de l'image utilisée dans Fig. 2.1 ($w_1 = 0.42$, $w_2 = 0.15$, $w_3 = 0.3$ et $w_4 = 0.13$).

	R_1	R_3	R_2	R_4
R_1		(0.13,0.15,0.46,0.02,1)	(0.19,0.19,0,0.64,0)	(1,0,0.08,0,0)
R_3	(0.04,0.06,0,0.34,0)		(0.3,0.15,0,1,0)	(0.46,0.13,0.12,0.014,0)
R_2	(0.2,0.12,1,0,0)	(0.01,0.01,1,0,0)		(0.32,0.15,1,0,0)
R_4	(0,1,0.18,0.12,0)	(0.72,0.11,0.13,0.58,1)	(0.65,0.31,0,1,0)	

TABLE 2.1 – Quintuplets des relations spatiales entre les régions de l'image utilisée dans Fig. 2.1.

2.3 Evaluation de la similarité d'images

Plusieurs techniques de mise en correspondance de deux ensembles de régions ont été proposées pour comparer deux images dans le contexte du RBIR. Les deux mesures les plus communément employées sont l'IRM (Integrated Region Matching) [100] et l'EMD (Earth Mover's Distance) [128]. Toutefois, ces techniques sont très complexes et elles fonctionnent dans les deux sens (*i.e.* de l'image requête I_{req} vers l'image testée I_{test}, et vice versa). Cependant, il est plus logique de baser les appariements sur les régions de l'image requête, puisque c'est elle que l'utilisateur a choisi pour lancer sa recherche donc les images qu'il souhaite retrouver sont celles qui lui ressemblent le plus. Ainsi, dans le but d'illustrer plus sémantiquement la similarité d'une image par rapport à I_{req}, nous avons proposé une approche rapide de mise en correspondance des graphes correspondants. Cette approche est basée sur une comparaison simultanée des propriétés et de la disposition spatiale des régions floues. Un fois les deux graphes sont appariés, une mesure de similarité permet d'évaluer le degré de similarité de l'image testée par rapport à celle en requête.

2.3.1 Mise en correspondance des graphes

La mise en correspondance de graphes est un problème très complexe, surtout que dans notre cas nous utilisons des graphes complets. Ceci revient à apparier les nœuds des graphes afin d'identifier leurs points communs. Cet appariement peut se faire à travers la recherche d'une relation d'isomorphisme de graphes ou de sous-graphes dans l'objectif de prouver l'existence d'une relation d'équivalence ou d'inclusion entre les deux graphes. L'isomorphisme de graphes consiste à décider si deux graphes sont structurellement identiques. L'appariement dans ce cas est bijectif et il est considéré comme étant un problème de complexité ouverte [143]. L'isomorphisme de sous-graphes consiste à trouver un graphe qui est inclus dans l'autre via une fonction injective. La complexité d'isomorphisme de sous-graphes est NP-complet. Toutefois, vu que deux objets "similaires" ne sont pas nécessairement "identiques", plusieurs techniques de comparaison de graphes à tolérance d'erreurs, telles que la recherche du plus grand sous-graphe commun et la distance d'édition de graphes, ont été proposées [143]. Dans notre cas, la mise en correspondance entre une image I_{test} et l'image requête I_{req} est formalisée comme étant un problème de recherche du plus grand sous-graphe induit commun, ce qui consiste à trouver un sous-graphe de G_{test} qui est isomorphe à un sous-graphe de G_{req}. En effet, la recherche du plus grand sous-graphe commun dans le cadre d'un appariement univoque peut être induit (ayant le plus grand nombre de sommets communs) ou partiel (ayant le plus grand nombre d'arcs communs). Dans les deux cas, le problème d'appariement est considéré comme NP-difficile. Pour ce type d'appariement, des contraintes, dures ou souples, doivent être imposées lors de la mise en correspondance. Pour un graphe étiqueté, une contrainte dure indique que les étiquettes des nœuds ou des arcs doivent être identiques, alors qu'une contrainte souple consiste à réaliser une correspondance qui maximise la similarité entre les éléments comparés. Les contraintes dans notre cas portent sur les arcs et elles sont dures. Pour ce faire, nous avons proposé une approche rapide de mise en correspondance qui consiste à traiter les régions de l'image requête I_{req} par ordre décroissant de poids. En effet, les régions de I_{req} ayant les plus grands poids désignent généralement les objets d'intérêt que l'utilisateur cherche puisque elles se localisent au centre et occupent la plus grande partie de l'image. Ceci ajoute une dimension sémantique et assure une comparaison plus logique entre les différentes images. Ainsi, nous commençons dans une première étape par correspondre chaque région de l'image requête à un ensemble de régions candidates dans I_{test}, qui respectent un seuil de similarité bien précis en termes de descripteurs de bas niveau (les sous-bandes d'ondelettes). Cette affectation n'est pas définitive car elle ne prend en considération que l'apparence visuelle des régions. Ensuite, nous passons à la phase de comparaison des arêtes, qui nous permet soit d'affirmer les choix de la première phase soit de découvrir de nouvelles correspondances (Fig. 2.4). Ceci s'explique par le fait que le degré de similarité d'une région par rapport à une autre est une combinaison du degré de similarité de leurs apparences visuelles ($\Im_{\Re}$) et d'une deuxième valeur de similarité de leurs dispositions spatiales ($\Im_{\mathcal{A}}$). Ainsi, les appariements initiaux à base de descripteurs de régions sont employés comme évidence pour un modèle de régression, basé sur la programmation dynamique, qui estime la correspondance visuelle et topologique à travers l'image entière. En effet, afin

d'aboutir à des résultats pertinents dans un temps de calcul acceptable, nous avons adapté l'algorithme classique de la programmation dynamique, tout en intégrant la notion des poids des régions, avec une complexité polynomiale égale à $\theta(n^4)$ ($n = Card(\Re_{req})$). Notons que dans le cas où deux régions ont deux poids égaux, le fait de commencer par l'une ou l'autre n'affecte pas le résultat final de la mise en correspondance, en particulier pour les régions avec les plus grands poids puisque c'est elles qui guideront la phase de mise en correspondance. L'appariement proposé est univoque (*i.e.* chaque sommet de G_{req} peut être apparié au maximum à un sommet de G_{test}), ce qui permet de couvrir toutes les possibilités de différentes cardinalités des nœuds des graphes entre G_{req} et G_{test}. Ainsi, ayant en entrée les graphes G_{req} et G_{test}, la mise en correspondance de ces deux graphes cherche à définir pour chaque région R_i de l'image requête I_{req} (ces régions sont triées selon leurs poids w_i), la région R_i^* de l'image test I_{test} qui peut lui correspondre, comme le résume l'algorithme suivant :

Algorithme 3: Mise en correspondance entre deux graphes d'images

Entrées : G_{req}, G_{test} ;
Sorties : $\{(R_i, R_i^*) \in \Re_{req} \times \Re_{test},\ 1 \leq i \leq Card(\Re_{req})\}$;

début
  répéter
    $i \leftarrow 1$; /* indice de la région de I_{req} en cours de traitement, tels que $w_1 \geq w_2 \geq \ldots \geq w_n$ */
    $Cand_i \leftarrow \{R'_j \in \Re_{test} / d(R_i, R'_j) \simeq 0\}$; /* liste des candidats de G_{test} correspondants à R_i */
    si $i = 1$ **alors**
      $R_i^* \leftarrow \underset{R'_j \in Cand_i}{\text{argmin}} \|\Im_{\Re}(R_i) - \Im_{\Re}(R'_j)\|$;
    sinon
      $R_i^* \leftarrow \underset{R'_j \in Cand_i}{\text{argmin}} \sum_{k=1}^{i-1} (\|\Im_{\mathcal{A}}(R_i, R_k) - \Im_{\mathcal{A}}(R'_j, R_k^*)\| + \|\Im_{\mathcal{A}}(R_k, R_i) - \Im_{\mathcal{A}}(R_k^*, R'_j)\|)$;
    $i \leftarrow i + 1$;
  jusqu'à $i > n$;

2.3.2 Mesure de similarité inter-images

Une fois le sous-graphe de G_{test} isomorphe à G_{req}, ou à un sous-graphe de G_{req}, est déterminé, et ceci en modélisant simultanément les propriétés des régions dans un premier ordre et les relations spatiales entre les régions dans un second ordre [54], l'étape suivante cherche à évaluer le degré de similarité entre l'image testée et celle en requête. Ceci revient à estimer à quel point le graphe G_{test} est sémantiquement semblable à G_{req}. Pour ce faire, nous avons proposé une mesure globale de la similarité $Sim(I_{test} \% I_{req})$, de I_{test} par rapport à I_{req}, qui est basée sur une comparaison simultanée des propriétés visuelles des régions floues et des relations spatiales entre ces régions. La comparaison est guidée par les régions les plus importantes dans l'image requête, sachant que chaque région de cette image est associée au maximum à une région de l'image testée. En effet, la mesure proposée est définie par l'inverse de la distance moyenne pondérée, par les poids des régions de l'image requête, aussi bien entre les régions appariées qu'entre les arcs appariés (2.8). Cet arrangement équilibré entre la mesure de similarité des nœuds et celle des arêtes permet de profiter autant que possible du contenu sémantique des images comparées. Notons que la mesure Sim^{-1}, qui évalue la dissimilarité entre deux images, préserve la caractéristique suivante de la métrique distance : "la distance entre les mêmes images est égale à 0" [22]. Toutefois, vu qu'elle est fortement guidée par les poids des régions de l'image requête, cette mesure de dissimilarité ne vérifie pas la propriété de symétrie.

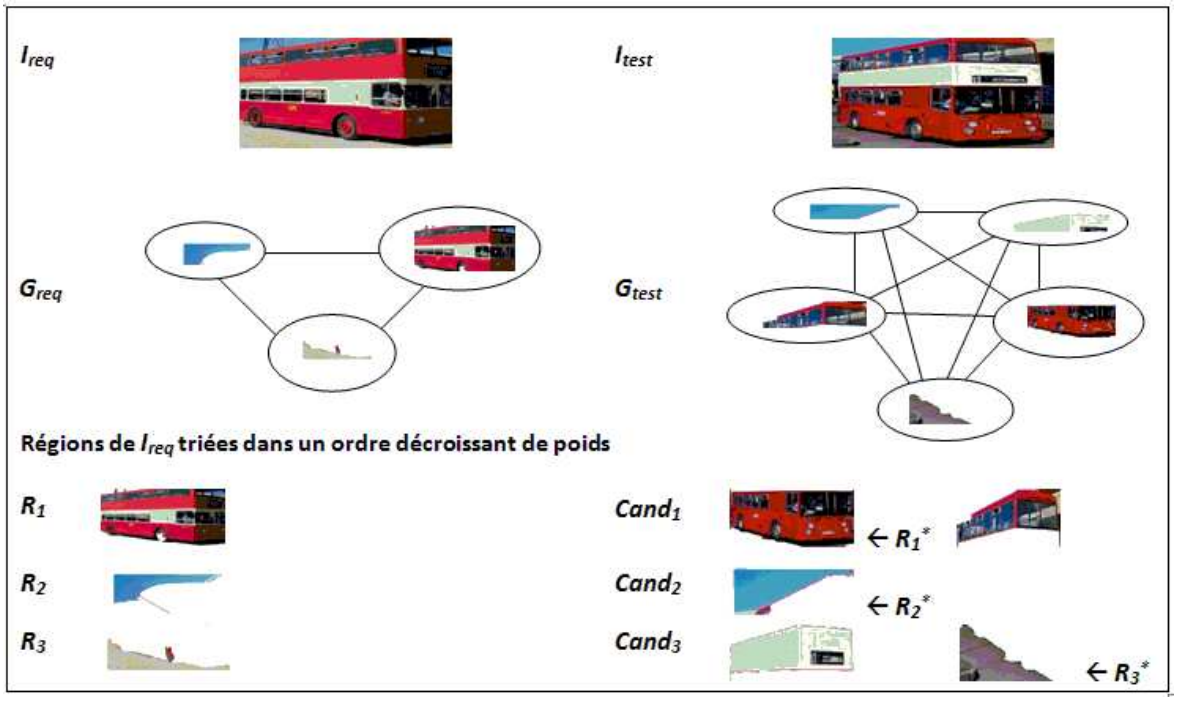


FIGURE 2.4 – Exemple d'application de l'algorithme de mise en correspondance des graphes.

$$Sim(I_{test}\%I_{req}) = \left[\frac{1}{n}.\sum_{i=1}^{n} 1_{\overline{\Re}}(R_i).w_{R_i}.\|\Im_{\Re}(R_i) - \Im_{\Re}(R_i^*)\| + \frac{2}{n.(n-1)}.\right.$$
$$\left.\sum_{i=1}^{n-1}\sum_{k=i+1}^{n} 1_{\overline{\Re}}(R_i).w_{R_i}.(\|\Im_{\mathcal{A}}(R_i,R_k) - \Im_{\mathcal{A}}(R_i^*,R_k^*)\| + \|\Im_{\mathcal{A}}(R_k,R_i) - \Im_{\mathcal{A}}(R_k^*,R_i^*)\|)\right]^{-1}. \quad (2.8)$$

où, $1_{\overline{\Re}}$ est la fonction caractéristique (dite aussi fonction indicatrice) de l'ensemble $\overline{\Re}$ ($\subset \Re_{req}$) des régions de I_{req} qui ont des correspondants dans I_{test} (*i.e.* $\overline{\Re} = \{R_i \in \Re_{req}/Cand_i \neq \emptyset\}$).

Afin de valider l'approche proposée pour la recherche des images à base de contenu des régions, nous l'avons appliquée sur des bases d'images généralistes. Le choix de la base d'images " Wang ", utilisée dans le système SIMPLIcity [153], s'explique par le fait qu'elle représente un standard des bases généralistes pour évaluer les systèmes de recherche des images [147]. Cette base hétérogène, composée de 1000 images divisées équitablement en 10 classes étiquetées ("Afrique", "Plage", "Bâtiment", "Bus", "Dinosaure", "Eléphant", "Fleur", "Cheval", "Montagne" et "Nourriture"), présente plusieurs niveaux de difficulté vu son hétérogénéité même au sein d'une seule classe [99]. Pour évaluer objectivement notre approche, nous avons produit les courbes de Rappel/Précision qui illustrent la métrique la plus communément employée pour mesurer l'efficacité d'un système de recherche d'information. Le rappel (2.9) mesure la capacité de récupération d'un système pour présenter toutes les images d'une même classe, alors que la précision (2.10) reflète sa capacité d'afficher uniquement les images pertinentes. Pour montrer l'efficacité de notre approche basée sur des descripteurs bas niveau d'ondelettes, le premier ensemble de courbes indique le taux de rappel en fonction de la précision des recherches effectuées avec la sous-bande HH_2 seulement et avec toute l'information d'ondelettes (Fig. 2.5). Il est clair que l'amélioration est trop faible aux dépens d'un temps de calcul très élevé.

$$Rappel = \frac{Card(\{Images\ Pertinentes\} \cap \{Images\ Retrouvées\})}{Card(\{Images\ Pertinentes\})}. \tag{2.9}$$

$$Précision = \frac{Card(\{Images\ Pertinentes\} \cap \{Images\ Retrouvées\})}{Card(\{Images\ Retrouvées\})}. \tag{2.10}$$

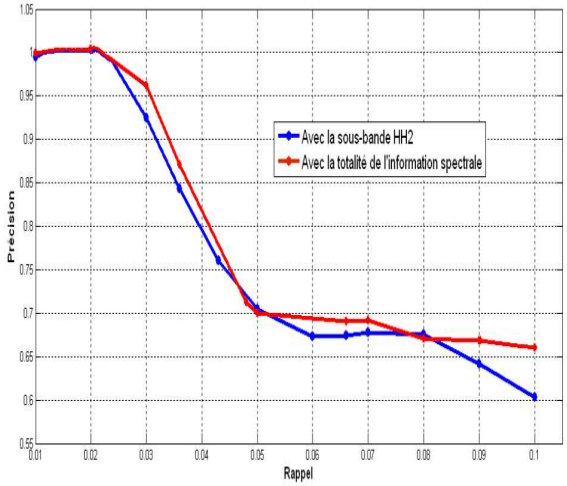


FIGURE 2.5 – Comparaison entre les résultats de la recherche avec la sous-bande HH_2 uniquement et ceux avec toutes les bandes du spectre.

Le deuxième ensemble de courbes Rappel/Précision illustre les résultats produits pour les différentes classes de la base "Wang" (Fig. 2.6). Ces résultats sont fortement sensibles au degré d'hétérogénéité de chaque classe. Prenons l'exemple de la classe " Dinosaure". Cette classe est la plus homogène pour un système perceptuel humain par rapport aux autres classes, vu que l'objet d'intérêt est toujours centralisé avec une taille importante et le fond est quasiment uniforme pour les 100 images (Fig. 2.7). Ceci explique le taux élevé de précision ainsi que du rappel. De plus, cette classe ne se compose pas de photographies mais plutôt de dessins. De là, nous pouvons conclure que l'approche proposée permet implicitement la distinction entre ces deux types d'images (photographie *vs.* dessin) puisque les résultats affichés ne comportent que des images de cette classe (Rappel/Précision = 1). De cette manière, il sera possible d'obtenir une classification raisonnable et préliminaire de l'ensemble des images de notre base sans avoir besoin d'utiliser un algorithme de classification dédié à cette tâche [63]. Pour la classe "Eléphant", en dépit de la multitude d'arrière-plans qui apparaissent dans les images de cette classe, les résultats de la recherche sont très encourageants. Ceci est principalement dû aux heuristiques utilisées lors de la procédure d'appariement des graphes et qui sont basées sur les poids des régions et sur l'utilisation de l'information spatiale inter-régions pour corriger les correspondances spectrales des régions. De l'autre côté, la dégradation des résultats pour la classe "Plage" s'explique par le fait que les principaux objets (eau, ciel, terre) composant les images de cette classe sont souvent présents dans la quasi-totalité des images de la collection de "Wang" (Fig. 2.7), qui est essentiellement une base d'images naturelles. Pour les classes "Afrique", "Bâtiment" et "Nourriture", la chute du taux de Rappel/Précision est due à l'augmentation du niveau d'hétérogénéité au sein de ces classes et de la différence élevée des apparences dans certains cas entre les mêmes objets de la même classe (Fig. 2.7), ce qui pose un problème d'appartenance à la classe dans le cas où nous ne considérons pas l'information apportée par la vérité-terrain. En plus, les confusions au niveau de certains résultats retournés pour ces classes sont dues à la forte ressemblance des objets pour des images sémantiquement différentes. Par exemple, pour la classe "Bâtiment", l'image requête contient souvent un objet d'intérêt (bâtiment)

mais également d'autres objets aux alentours (gazon, ciel...). De là, en retour à cette requête, quelques images sont retrouvées parce qu'elles comportent presque les mêmes composantes visuelles que l'image requête, bien qu'elles ne fassent pas partie de la classe "Bâtiment" (Fig. 2.7).

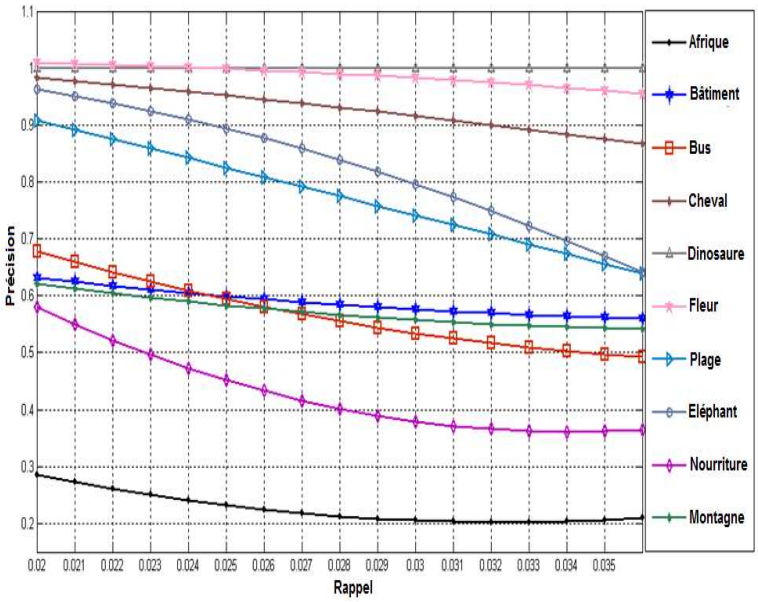


FIGURE 2.6 – Les courbes Rappel/Précision qui illustrent les résultats de la recherche pour les 10 classes de la base "Wang".

En outre, dans le but de comparer notre contribution par rapport aux travaux les plus pertinents dans le même contexte de RBIR avec des descripteurs d'ondelettes, nous avons eu recours à deux systèmes : le système "Anaktisi" [164] et le système "SIMPLIcity" [153]. Le choix de ce dernier système s'explique par le fait qu'il est l'un des systèmes les plus référencés dans la littérature du CBIR [79]. En outre, il utilise le concept de graphes comme structure pour la modélisation des images et lors de l'appariement. Par rapport à "SIMPLIcity", la courbe de Rappel/Précision de notre solution indique une amélioration considérable de la qualité des résultats (Fig. 2.8), bien que nous utilisons un vecteur descriptif réduit par rapport à "SIMPLicity", qui prend en considération les informations de texture, de couleur et d'ondelettes. En effet, notre courbe Rappel/Précision reste quasiment constante par rapport à celle de "SIMPLIcity" qui décroît remarquablement lorsque le nombre des résultats affichés augmente. Ceci peut être expliqué par la perte de l'information sémantique à cause de l'utilisation d'une description par ondelettes sur des blocs de l'image et non pas sur des objets d'intérêt. Toutefois, vu que "SIMPLIcity" utilise l'espace couleur Lab, ses résultats sont légèrement meilleurs pour les classes "Cheval" et "Eléphant", où les objets de l'avant-plan se distinguent nettement de l'arrière-plan qui est souvent quasi-uniforme [65]. Pour le système "Anaktisi", qui est basé sur une combinaison de plusieurs composantes couleurs avec des descripteurs de texture au niveau des blocs, il est clair que les résultats fournis par notre approche sont meilleurs, notamment lorsque le nombre des images retrouvées croît (Fig. 2.8). Notons que l'utilisation de descripteurs combinés par "Anaktisi" permet de réduire, par rapport à "SIMPLIcity", le manque de la représentation sémantique des blocs. Ce qui caractérise l'approche proposée par rapport à celles proposées par [153] et [164] est la modélisation au niveau des régions et l'application des ondelettes sur ces régions et non pas sur des blocs qui ne correspondent pas forcément à des objets réels[3]. En plus, notre solution est entièrement basée sur la notion de la logique floue ce

3. Même en le comparant à [73], qui utilise des descripteurs d'ondelettes, de texture et de couleur sur les régions, l'approche proposée produit des résultats meilleurs [63].

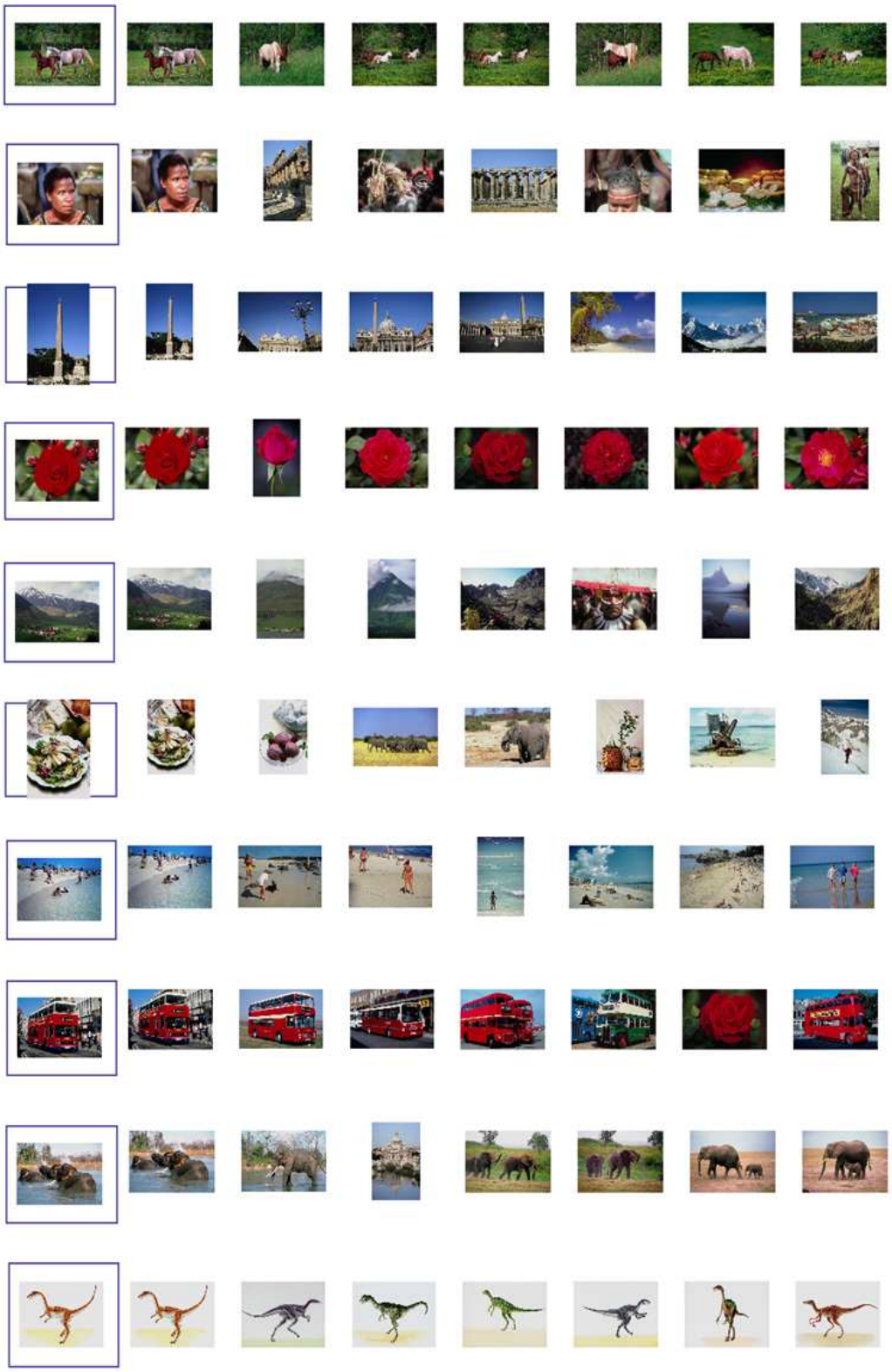


FIGURE 2.7 – Résultats de la recherche pour un échantillon de 10 images requêtes illustrant les 10 classes de la base "Wang" : la première image est celle en requête suivie par les sept premières images retrouvées.

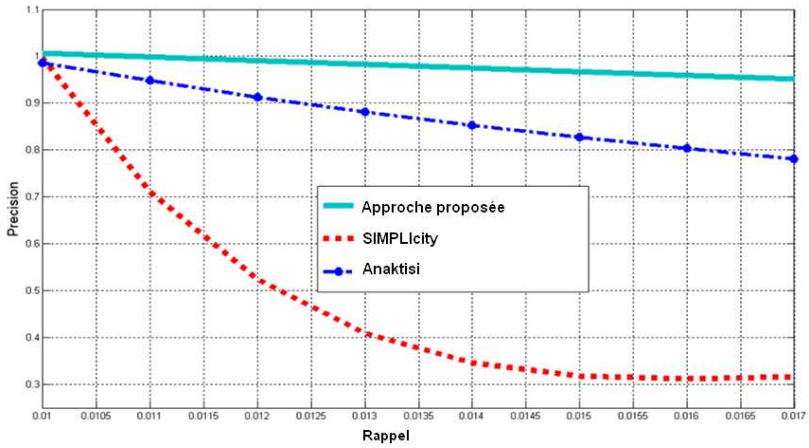


FIGURE 2.8 – Courbes Rappel/Précision illustrant la comparaison entre l'approche proposée et les systèmes "SIMPLIcity" et "Anaktisi".

qui lui permet de modéliser l'incertitude et d'éviter les erreurs potentielles d'appartenance exacte aux régions. En particulier, contrairement aux systèmes comparés, qui utilisent l'information spectrale avec le résultat de la segmentation afin de corriger ses éventuelles erreurs, c'est au niveau de la segmentation même que nous essayons de corriger les résultats produits. En l'occurrence, alors que SIMPLIcity utilise l'information spectrale calculée sur des blocs de tailles réduites de l'image dans une étape séparée de celle de la segmentation (aucune liaison entre les régions obtenues lors de la segmentation et les blocs définis par la suite), notre approche cherche à lier et à affecter l'information spectrale par le résultat de la segmentation floue. Par ailleurs, grâce à la considération de l'information spatiale, il est toujours possible pour notre approche de procéder à des corrections supplémentaires après l'appariement du contenu visuel. Ainsi, nous considérons notre approche comme plus simple que ces approches tout en étant plus performante vu les résultats produits sur les mêmes bases d'images.

2.4 Bouclage de pertinence

Dans le but de raffiner les résultats fournis après une première récupération d'images, le bouclage de pertinence (BP) profite des feedbacks de l'utilisateur afin d'ajouter de la sémantique à la procédure de recherche des images lors des prochaines itérations [156]. En effet, sous l'hypothèse que le jugement de l'utilisateur est plus pertinent que celui du système, le BP permet à l'utilisateur de raffiner sa requête en fournissant au système des images exemples et/ou des images contre-exemples de ce qu'il souhaite obtenir. Le système prend en compte cette sélection et procède à un apprentissage actif pour retourner des résultats plus pertinents. Par la suite, l'utilisateur peut re-sélectionner d'autres exemples et réitérer le même processus, sachant que la qualité d'une méthode de bouclage de pertinence est définie par le nombre d'itérations et d'interactions permettant d'aboutir à la satisfaction de l'utilisateur. Une synthèse bibliographique nous a permis de classifier les méthodes de bouclage de pertinence en recherche d'images en trois catégories : par modèle d'apprentissage, par modèle probabiliste et par modèle de représentation vectorielle [63]. Pour la première classe, l'une des méthodes les plus utilisées est l'apprentissage par la méthode des séparateurs à vaste marge (SVM) [155]. Bien que cette méthode offre une flexibilité au niveau des données traitées et une rapidité d'apprentissage et de test, elle ne fournit généralement pas des résultats pertinents à cause du nombre limité des échantillons (exemples positifs et négatifs) sur lesquels elle opère. Nous citons également la méthode d'apprentissage par amplification (boosting) [79], qui consiste à considérer l'ensemble des descripteurs utilisés dans la collection traitée tout en accordant un classifieur à taux d'erreur élevé pour chaque descripteur. Par la suite, le classifieur qui minimise le taux d'erreur est choisi avec une certaine pondération. Le problème de boosting réside

dans le nombre des descripteurs, qui doit être très élevé pour garantir de bons résultats. En plus, cette méthode fonctionne mieux sur les descripteurs de l'image qui ne sont pas fortement corrélés avec la perception humaine [81]. La tendance actuelle consiste à réaliser l'apprentissage par des algorithmes génétiques [48] et par des algorithmes de colonies [169]. Ceci peut être vu comme effet de mode mais les résultats obtenus dans certains cas argumentent ce choix [33] [34]. Pour le modèle probabiliste, les itérations de bouclage de pertinence sont des probabilités conditionnelles, où il faut à chaque phase tenir compte des actions faites pendant les phases précédentes. L'information acquise antérieurement ne permet pas seulement d'éliminer les fausses correspondances (surtout pour le BP négatif) mais aussi de définir de nouvelles métriques de similarité plus souples et plus convenables aux attentes de l'utilisateur. Dans [88], les auteurs ont proposé un modèle Bayésien qui soutient les classes d'images qui assignent une probabilité élevée aux images définies comme exemples positifs et pénalise les classes qui accordent une probabilité élevée d'appartenance aux exemples négatifs. Dans [145], les auteurs ont présenté une classe Bayésienne pour les exemples positifs dans l'objectif d'estimer la distribution Gaussienne qui représente la classe des images recherchées, alors que les exemples négatifs sont employés pour modifier le rang des candidats recherchés. De même, un algorithme glouton d'approximation est défini dans [120], dont l'idée de base est d'utiliser la totalité de la collection d'images en accordant à chaque sous-ensemble de cette collection une étiquette. Par la suite, les probabilités calculées portent sur les informations données par ces étiquettes. Par rapport à la classe du modèle d'apprentissage, le nombre réduit des exemples présentés par l'utilisateur ne pose plus un problème pour le modèle probabiliste. Toutefois, le problème des techniques de bouclage de pertinence par les modèles probabilistes réside dans la définition exacte du modèle adapté (fonction à appliquer). En plus, la requête présentée au début de la recherche affectera forcément les probabilités calculées ultérieurement puisque c'est elle qui définit la probabilité initiale. Pour la classe du modèle vectoriel, le principe est de se focaliser sur la représentation de l'image sous la forme d'un vecteur descriptif. En effet, l'idée de l'apprentissage est toujours présente même avec ce modèle, sauf qu'il n'y a pas concrètement des classifieurs ou des machines d'apprentissage dédiées mais plutôt des mises à jour successives des données relatives à la représentation de la collection d'images. Deux variantes principales existent dans ce modèle : la première se base sur la pondération des descripteurs et la deuxième sur la pondération des images elles-mêmes. Pour la première variante, les systèmes de CBIR proposent l'intervention de l'utilisateur pour mettre le point sur les descripteurs utilisés lors de la recherche. En effet, chaque descripteur est lié à une pondération de façon que la recherche soit plus influencée par le descripteur le plus pondéré. Après une première recherche, un nouvel ordre de pondération des descripteurs utilisés est proposé en fonction des feedbacks de l'utilisateur. Cette technique est aussi appelée "feature re-weighting" au niveau de laquelle on applique le mouvement du point-requête. Ceci consiste en un déplacement du point-requête vers la zone des exemples positifs et un éloignement de la zone des exemples négatifs dans l'espace de descripteurs des images [90]. La deuxième variante consiste à accorder un score à chaque image reflétant l'intérêt qu'elle représente pour l'utilisateur. Initialement, ce score est identique pour toutes les images de la collection. Par la suite, après la première recherche, la valeur de ce score est mise à jour pour traduire le degré de similarité avec l'image requête ou exprimer l'ordre d'apparence lors de la recherche. La définition du niveau de pertinence par l'utilisateur affecte ce score pendant les recherches ultérieures. Une image sélectionnée comme pertinente peut regagner un poids additionnel favorisant son degré de similarité avec d'autres images et permettant ainsi de retrouver de nouvelles correspondances, alors que les images notées comme non-pertinentes seront écartées des recherches postérieures via une minimisation de score ou même une annulation [88]. La limite principale de ces deux variantes est le risque d'aboutir à des optimums locaux. En effet, la projection des images sur les axes vectoriels des descripteurs utilisés peut rassembler des images semblables selon l'apparence et non pas selon la sémantique. Ainsi, il sera difficile de retrouver d'autres images ayant la même sémantique lors de la phase de bouclage de pertinence. Pour cela, [78] proposent une nouvelle structuration de la collection d'images sous la forme d'un arbre. Chaque nœud de cet arbre correspond à un cluster d'images similaires, et à chaque niveau de l'arbre, le nœud du niveau supérieur se divise en un ensemble de nœuds représentant des clusters plus fins jusqu'à arriver aux feuilles qui sont des images uniques.

2.4.1 Bouclage de pertinence par mise à jour des poids des régions de l'image requête

Vu que la signature de l'image dans notre cas est modélisée par une matrice descriptive avec une pondération relative à chaque région, nous avons adopté un bouclage de pertinence par modèle de représentation vectorielle avec pondération d'images, et plus précisément des régions de l'image requête. Ainsi, les mises à jour successives des pondérations, telles qu'elles sont présentées dans le modèle vectoriel, permettent de cerner mieux d'une itération à une autre l'aspect sémantique de la requête selon les recommandations de l'utilisateur. Notons que l'utilisation d'un modèle de représentation vectorielle permet d'obtenir des résultats encourageants même avec un nombre réduit d'exemples négatifs et/ou positifs ce qui n'est pas le cas pour les autres modèles de bouclage de pertinence (surtout celui par apprentissage). En effet, nous avons proposé une technique de bouclage de pertinence qui se base sur la description des régions composant une image par les transformations des sous-bandes d'ondelettes de haut niveau. L'objectif principal de notre schéma de bouclage de pertinence à base de régions est d'ajuster l'importance visuelle des régions de l'image requête selon les rétroactions négatives des utilisateurs. En effet, le fait d'écarter une image à partir d'une liste est une réaction de rejet et elle est beaucoup plus spontanée que la sélection des exemples positifs. Ceci s'explique par l'intention cachée du cerveau de se focaliser seulement sur les autres exemples et éliminer les intrus qui sont plus distinguables à première vue [15]. Il s'avère ainsi plus facile pour l'utilisateur de décider directement à propos des images qu'il ne cherche pas plutôt que ce qu'il veut exactement avoir [158]. En outre, dans la plupart des cas, le nombre des exemples non pertinents est limité par rapport à celui des exemples positifs affichés après une phase de recherche. De là, le bouclage négatif assure l'élimination des fausses correspondances après un nombre réduit d'itérations . En effet, si l'utilisateur est non satisfait par les images retrouvées, le bouclage de pertinence lui propose d'indiquer un ensemble d'images qu'il trouve non pertinentes dans la récupération déjà produite. La prétention de base derrière notre bouclage de pertinence est que chacune des images déjà retrouvées devrait avoir au moins une région qui est fortement semblable, en termes de descripteurs de bas niveau, à une région de l'image requête. Néanmoins, il est habituel qu'une image soit affichée parce qu'elle a quelques régions qui sont fortement semblables à certaines régions peu importantes dans l'image requête (*e.g.* le ciel et la mer dans Fig. 2.7). Ces régions, qui peuvent être significatives mais ne représentent pas le centre de perception de l'utilisateur, correspondent dans la plupart des cas aux régions du fond [89]. Pour cette raison, ces régions doivent subir une minimisation de leur poids pour qu'elles affectent moins les itérations ultérieures de la recherche. En revanche, les régions qui n'ont pas de correspondants dans les images sélectionnées regagnent un poids supplémentaire. De cette façon, c'est principalement l'objet d'intérêt qui instaure de plus en plus son apport sémantique par rapport aux objets non désirés, sachant que les pondérations des régions de l'image requête changent avec chaque exemple négatif sélectionné. En effet. suite au choix par l'utilisateur de l'ensemble Υ_{NP} des images non pertinentes, la technique proposée identifie les régions les moins importantes de l'image requête, relativement à l'intuition de l'utilisateur, afin de les éliminer pendant la prochaine itération de recherche. Cette élimination est basée sur la réduction du taux de similarité entre la liste des images non pertinentes et l'image requête. En effet, au lieu d'augmenter l'importance des régions d'intérêt par des valeurs empiriques, il s'avère plus raisonnable d'indiquer au début les régions qui doivent être mises en relief et celles à négliger pendant la phase suivante de recherche. De cette manière, les mises à jour des poids visent directement les régions intéressant l'utilisateur même si certaines parmi elles peuvent ne pas avoir une pondération élevée pendant la définition initiale de la signature de l'image. Le fait de multiplier les poids des régions non pertinentes par le poids de la région d'intérêt permet d'affecter, à ces régions, des pondérations réduites par rapport à celles de départ. Nous minimisons cette ressemblance potentielle en l'affectant par le poids le plus pondérant, qui est celui accordé à la région ayant le plus grand poids dans l'image requête $P_{add} = w_1$ (< 1). Cette réduction assure que l'image choisie comme exemple négatif ne sera plus affichée parmi les nouveaux résultats, vu son taux minimal de similarité avec l'image requête. De l'autre côté, les autres régions, supposées être pertinentes, de l'image requête regagnent des poids supplémentaires. Ayant l'ensemble Υ_{NP} des images non-pertinentes interactivement choisies par l'utilisateur, la mise à jour proposée des poids des régions de l'image requête peut être résumée comme suit :

Algorithme 4: Mise à jour des poids des régions de l'image requête suite à une itération de BP

Entrées : $w_1, \ldots, w_n$; /* les poids initiaux des n régions de l'image requête */
Sorties : $w_1, \ldots, w_n$; /* les nouveaux poids des régions de l'image requête */

début
  pour chaque *image* I_{test} *dans l'ensemble* Υ_{NP} **faire**
    $Cpt \leftarrow 0$;
    pour chaque *région* R_i *de l'image requête* ($\in \Re_{req}$) **faire**
      $\vartheta_i \leftarrow \{R'_j \in \Re_{test} / d(R_i, R'_j) \simeq 0\}$;
      si $\vartheta_i \neq \emptyset$ **alors**
        $w_i \leftarrow w_i.P_{add}$;
      sinon
        $Cpt \leftarrow Cpt + 1$;
    pour chaque *région* $R_i \in \Re_{req} / \vartheta_i = \emptyset$ **faire**
      $w_i \leftarrow w_i + \frac{P_{add}}{n - Cpt}$;

Ainsi, les pondérations des régions de l'image requête changent avec chaque exemple négatif sélectionné. Le fait qu'une image, choisie comme exemple négatif, est affichée après une première récupération indique qu'elle a des régions en commun avec l'image requête. Mais ce sont des régions qui n'intéressent pas l'utilisateur. Pour cette raison, ces régions subissent une minimisation de leur poids pour qu'elles affectent moins les itérations ultérieures de la recherche (Fig. 2.9). Par exemple, pour l'image requête I_{req} de la classe "Eléphant" (Fig. 2.7), la troisième image retournée I_{NP} ne correspond pas au concept envisagé par la recherche de l'utilisateur. Ceci s'explique par le fait que l'image requête contient, en commun avec l'image I_{NP}, l'objet "eau" en bas et l'objet "ciel" en haut. De là, comme réponse à la requête I_{req}, l'image I_{NP} a été affichée, bien qu'elle fasse partie de la classe "Bâtiment". Cependant, suite à la sélection de l'image I_{NP} en tant qu'exemple négatif, les régions R_2, R_3 et R_4 de l'image requête vont subir une minimisation de poids pour qu'elles perdent leurs effets sur le calcul de similarité pendant les itérations suivantes (Fig. 2.9). En revanche, les régions qui n'ont pas de correspondants (R_1 et R_5) regagnent un poids supplémentaire. De cette façon, c'est principalement l'objet d'intérêt qui instaure de plus en plus son apport sémantique par rapport aux objets non désirés. Le fait de multiplier les poids des régions non pertinentes par le poids de la région d'intérêt (inférieur à 1) permet d'affecter à chacune de ces régions une pondération réduite par rapport à celle de départ (Fig. 2.9). Ainsi, le choix de I_{NP} comme exemple négatif permet de mieux cerner l'objet d'intérêt ("éléphant") et d'écarter les autres régions en commun avec l'image requête ("eau" et "ciel"). D'ailleurs, chacune de ces dernières régions perd un pourcentage important de sa pondération ce qui va la placer à la fin de la liste des régions lors de la mise en correspondance pendant la prochaine phase de recherche. Dans d'autres cas, la diminution de la pondération peut complètement changer le traitement lié à l'image requête car une nouvelle région d'intérêt sera mise en exergue au lieu de celle définie au début. De cette manière, l'étape de bouclage de pertinence négative sert à améliorer les résultats récupérés mais également à corriger la définition préliminaire de la signature de l'image requête en termes de pondération des régions qui la composent.

Par la suite, la recherche de nouvelles correspondances est lancée dans le reste de la base d'images, tout en élargissant la bande des ondelettes, utilisées comme descripteurs de comparaison, de HH_2 seulement vers HH_2 et HB_2. Ceci permet de prendre en compte plus de détails relatifs aux images traitées surtout ceux liés à l'allure des objets, sachant que la possibilité d'avoir de mauvaises correspondances est

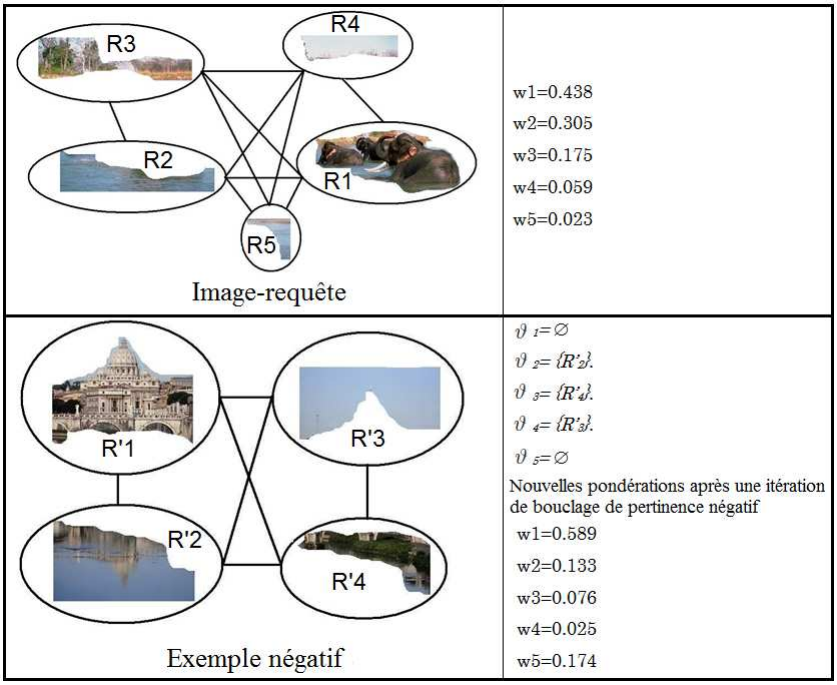


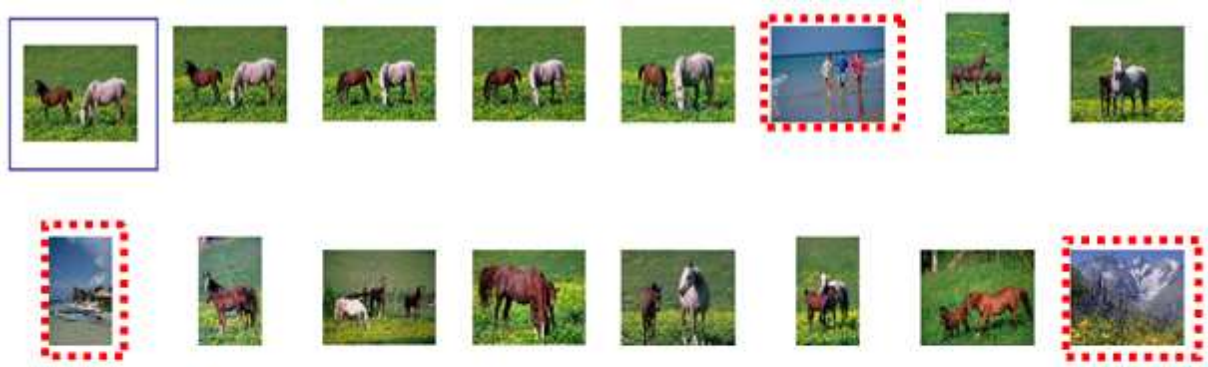
FIGURE 2.9 – Mise à jour des poids des régions de l'image requête après la sélection de I_{NP} comme exemple négatif.

réduite à cette étape et les objets de l'image sont mieux modélisées (leurs poids sont plus importants). Ainsi, l'ajout de la deuxième sous bande d'ondelettes lors des prochaines itérations permet d'améliorer nettement les résultats (Fig. 2.10), puisque la probabilité d'avoir de mauvaises correspondances est réduite à cette étape. Formellement, soient Υ l'ensemble des images de la base de test et Υ_{NP} celui des images non pertinentes sélectionnées par l'utilisateur, alors on a :

$$\forall I_{test} \in \Upsilon \setminus \Upsilon_{NP}, \quad \|\Im_{\Re}(R_i) - \Im_{\Re}(R_i^*)\| = \sqrt{(M_{HH}^{R_i} - M_{HH}^{R_i^*})^2 + (M_{HB}^{R_i} - M_{HB}^{R_i^*})^2}, \tag{2.11}$$

où, $M_{HH}^{R}(x,y) = [H_x * [G_y * I(x,y)]_{\downarrow 2,1}]_{\downarrow 1,2}$ et $G = (G_x, G_y)^t$ désigne un filtre Gaussien passe-bas.

Pour évaluer l'apport de notre technique de bouclage de pertinence négatif, nous présentons l'ensemble de courbes Rappel/Précision (Fig. 2.11). Il est clairement remarquable que l'utilisation d'une seule itération de BP permet d'améliorer considérablement les résultats obtenus lors de la première recherche (Rappel/Précision$\simeq$ 1). D'ailleurs, c'est ce que l'utilisateur souhaite avoir : un meilleur résultat avec le minimum d'itérations de BP. Ceci lui offre la possibilité de communiquer avec le système sans le déborder avec des tâches supplémentaires. En effet, les images obtenues au premier rang (généralement les quatre premières images) après la première recherche répondent sémantiquement au besoin de l'utilisateur. Mais, l'intervention de ce dernier vise à améliorer le reste des résultats affichés. Nous déduisons ainsi que l'approche proposée, même sans bouclage de pertinence, garde un taux de Rappel/Précision élevé pour ce premier rang d'affichage. D'ailleurs, il est clair que les deux courbes de Fig. 2.11 sont confondues au niveau des premiers résultats. Toutefois, la nouvelle signature de l'image requête (de point de vue pondération et quantification des régions) mène à découvrir de nouvelles correspondances, puisque la comparaison des images est guidée principalement par les poids des régions de cette image.



(a) Résultats de la recherche avant le BP, l'image-requête est encadrée par un trait uni et les exemples négatifs sont encadrés par un trait pointillé.

(b) Résultats de la recherche après l'application d'une itération de BP négatif.

FIGURE 2.10 – Amélioration des résultats de la recherche, pour une image requête de la classe "Cheval", après une itération de BP négatif.

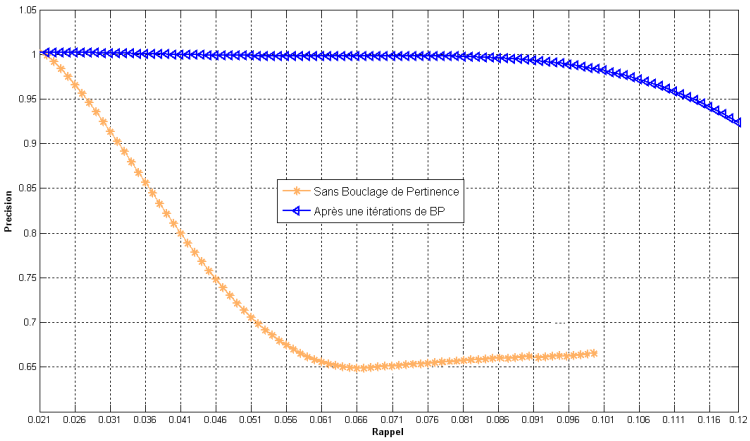


FIGURE 2.11 – Courbes Rappel/Précision illustrant l'amélioration des résultats de la recherche après une itération de bouclage de pertinence négatif.

2.5 Co-segmentation des images

La co-segmentation désigne la segmentation simultanée des objets similaires de l'avant-plan dans deux (ou plusieurs) images. L'objectif est de faciliter la détection d'un objet en exploitant le contraste des objets similaires dans d'autres vues. Plusieurs travaux sur la co-segmentation proposent des approches non-supervisées qui permettent de co-segmenter les objets de l'avant-plan automatiquement sans aucune intervention de l'utilisateur [41]. Ces travaux peuvent être regroupés en trois groupes principaux. Pour le premier groupe, la co-segmentation revient à minimiser une fonction d'énergie, qui évalue les similarités entre les objets de l'avant-plan, dans l'objectif de produire des segmentations lisses tout en imposant la cohérence de l'apparence des avant-plans. L'évaluation de la similitude est souvent basée sur l'histogramme d'apparence commun, qui est utilisé comme une contrainte globale de régularisation dans un cadre de champs de Markov (MRF) modifié. Par ailleurs, la co-segmentation est modélisée sous la forme d'un problème de classification discriminative sur deux classes (fond *vs.* avant-plan) pour les techniques formant le deuxième groupe. Les techniques de co-segmentation du dernier groupe sont basées sur le partitionnement des graphes [91]. Parmi les solutions efficaces de co-segmentation, l'approche dominante est la minimisation d'une fonction d'énergie via un estimateur du maximum a posteriori (MAP) sur MRF [127]. Dans cette approche, la co-segmentation d'une paire d'images (I_i, I_j) revient à l'application d'une segmentation MRF avec un terme supplémentaire qui code la variation dans les histogrammes des régions de l'avant-plan dans les deux images. La seule contrainte est que les objets de l'avant-plan soient similaires et la résolution du problème de co-segmentation revient à la minimisation d'une fonction d'énergie Ξ (2.12), afin de produire un étiquetage binaire de chaque image.

$$\Xi(I_i, I_j) = \Big(D_i + S_i\Big) + \Big(D_j + S_j\Big) + \alpha.G(H_i, H_j), \tag{2.12}$$

où, les deux premiers termes sont les termes d'énergie MRF pour I_i et I_j, respectivement.

Les termes MRF modélisent la cohérence spatiale intra-image, où D est un terme des données intrinsèques qui consiste à pénaliser les solutions qui sont incompatibles avec les données observées et S est un terme de lissage qui favorise une segmentation lisse de chaque image. Le dernier terme G est un terme global de correspondance, qui pénalise la dissimilarité inter-image, et α est le poids de ce terme. Ce terme illustre la différence entre les modèles Gaussien empiriques des deux histogrammes de l'avant-plan H_i et H_j. Trois modèles principaux ont été proposés pour évaluer les différences entre les histogrammes [150] : le modèle de la norme $L1$, le modèle de la norme $L2$ et le modèle de récompense. Les modèles basés sur les normes conduisent à des problèmes d'optimisation NP-difficiles, contrairement à celui de la récompense qui conduit à un problème sous-modulaire qui peut être efficacement et rapidement optimisé avec des coupes de graphes (graph-cuts) [127]. Toutefois, pour évaluer la correspondance entre les régions qui composent les images traitées, qui influe fortement les résultats finaux, les méthodes existantes comparent tout simplement les histogrammes en absence de toute information de cohérence spatiale [74]. Pour cela, nous avons proposé d'intégrer l'information spatiale dans la procédure d'optimisation, ce qui permet en particulier de minimiser l'effet du bruit sur les résultats finaux de co-segmentation. En outre, vu que la co-segmentation permet d'intégrer efficacement l'information temporelle pour la segmentation des objets mobiles dans les vidéos, nous avons adapté la méthode proposée de co-segmentation pour la détection automatique des athlètes, sans aucune hypothèse ou connaissance préalable sur le mouvement de la caméra. Ceci revient à segmenter toutes les instances des mêmes objets de l'avant-plan sur des fonds potentiellement arbitraires. L'objectif est de définir implicitement les régions d'intérêt, sans intervention humaine, via de multiples observations des athlètes sur différents arrière-plans [38].

2.5.1 Intégration de l'information spatiale

Pour évaluer les différences entre les histogrammes dans le cadre du terme global de correspondance (2.12), nous avons fait recours au modèle de récompense [74], tout en évitant les modèles basés sur les normes et qui conduisent à une optimisation complexe en raison de la présence du terme de différence

entre les histogrammes [112]. Ainsi, plutôt que de pénaliser les différences entre les deux histogrammes de l'avant-plan, le modèle utilisé récompense les affinités (2.13) pour segmenter les objets similaires des avant-plans indépendamment des contenus des fonds.

$$\begin{aligned} G(H_i, H_j) &= -\sum_{k=1}^{K} f_i(k) \cdot f_j(k) \\ avec, \quad f_n(k) &= \sum_{(x,y)/I_n(x,y)\in H_n(k)} lab_n(x,y), \end{aligned} \tag{2.13}$$

où, n ($\in \{i,j\}$) dénote l'indice de l'image traitée et $lab_n(x,y)$ ($\in \{0,1\}$) est l'étiquette du pixel (x,y) dans l'image I_n (fond (0) *vs.* avant-plan(1)).

Afin de considérer la cohérence spatiale des pixels voisins, notre stratégie consiste à utiliser l'entropie locale, au lieu de l'intensité, lors de la définition de l'histogramme d'une image [112]. L'entropie locale est une mesure statistique qui prend en compte, d'une manière locale, la distribution spatiale des niveaux de gris. Elle représente la variance d'une région locale et capte exceptionnellement les propriétés naturelles des "régions en transition", souvent situées entre un objet de l'avant-plan et l'arrière-plan, tout en réduisant efficacement les effets du bruit [159]. En effet, le voisinage de chaque pixel est considéré dans le but de caractériser la texture, qui fournit une information grossière sur la variabilité locale des valeurs d'intensité. Ainsi, pour chaque pixel (x,y), l'entropie locale E est calculée sur un voisinage $\aleph(x,y)$ de taille c×c autour de ce pixel (2.14).

$$E(x,y) = -\sum_{l=0}^{L-1} p_l \cdot \log(p_l), \tag{2.14}$$

où, p_l ($= n_l/c^2$) est la probabilité que le niveau de gris l apparaît dans le voisinage $\aleph(x,y)$, L est le niveau de gris maximal dans $\aleph(x,y)$ et n_l est le nombre de pixels dans $\aleph(x,y)$ qui ont un niveau de gris égal à l.

Les petites (*resp.* grandes) valeurs de l'entropie locale indiquent des pixels très prévisibles (*resp.* imprévisibles) et caractérisent ainsi des voisinages homogènes (*resp.* hétérogènes). La prise en compte de l'entropie locale permet l'intégration implicite de l'information spatiale, puisque les pixels voisins devraient avoir en général le même comportement (avant-plan *vs.* fond). Ceci permet d'éviter les pixels isolés sur les objets co-segmentés, en particulier sur les régions en transition aux alentours des objets de l'avant-plan. En plus, une technique de classification floue est introduite afin de modéliser l'ambiguïté de l'appartenance d'un pixel à un bin d'histogramme. Ceci permet d'éviter les faux détections tout en minimisant les effets du bruit. En effet, le principal défi pratique dans le cas de la segmentation du même objet est que les distributions peuvent ne pas correspondre exactement, en raison des changements de l'illumination et du point de vue ou de la présence des objets occultés [150]. Ainsi, la modélisation floue de l'appartenance d'un pixel à un bin de l'histogramme permet de reconsidérer l'imprécision, l'incertitude et le conflit inhérent aux attributs utilisés. En effet, après avoir associé chaque pixel (x,y) de l'image I_i à l'un des K bins de l'histogramme de l'entropie correspondant, tout en produisant les centres des bins à l'image I_j afin d'obtenir les mêmes classes pour les deux images traitées, nous avons proposé de reclasser les pixels ambigus apparaissant principalement sur les frontières entre les bins. Vu que le partage exact des valeurs de l'entropie locale sur K bins équidistants conduit à diminuer la certitude de l'appartenance correcte d'un pixel à un bin, en particulier pour ceux de la zone partagée par deux bins, notre idée consiste à définir un degré flou d'appartenance $\mu(x,y,k)$ pour chaque pixel (x,y) au bin k ($1 \leq k \leq K$), auquel le pixel appartient par défaut, ainsi que ses degrés d'appartenance $\mu(i,j,k-1)$ et $\mu(i,j,k+1)$ aux bins voisins (2.15). Ces degrés sont estimés en fonction de la position de l'entropie locale du pixel (x,y) relativement à la valeur centrale c_k du bin k. Si $E(i,j)$ est supérieur à (*resp.* inférieur à) $E(c_k)$, alors $\mu(i,j,k-1) = 0$, $\beta_1 = 0$ et $\beta_2 = 1$ (*resp.* $\mu(i,j,k+1) = 0$, $\beta_1 = 1$ et $\beta_2 = 0$).

$$\mu(x,y,k) = \frac{|E(x,y) - E(c_k)|}{|E(c_k) - \beta_1.E(c_{k-1}) - \beta_2.E(c_{k+1})|}. \tag{2.15}$$

Ensuite, chaque pixel (x, y) avec un faible degré d'appartenance ($\simeq 0$) à son bin par défaut k, et donc avec un degré d'appartenance élevé à l'un des bins voisins $k-1$ et $k+1$, devrait être associé au plus proche bin selon la distance normalisée entre son entropie locale et celles des centres des bins c_{k-1}, c_k et c_{k+1}. Cette classification floue basée sur l'entropie locale fournit une modélisation pertinente de l'ambiguïté de l'adhésion d'un pixel dans un bin d'histogramme. Elle permet aux pixels de changer leurs bins, ce qui permet la réduction des effets de sur-segmentation principalement dus au bruit et à la présence des zones similaires dans les fonds des images traitées I_i et I_j. Ensuite, les histogrammes obtenus sont comparés tout en maximisant la similarité entre eux de manière à obtenir approximativement le même avant-plan dans I_i et I_j. En effet, deux pixels dans $I_i \times I_j$ ne sont considérés comme similaires que si elles appartiennent tous les deux au même bin. Ceci revient à résoudre le problème d'optimisation suivant :

$$\begin{aligned} \max \textstyle\sum_{k=1}^{K} \big(& |\{(x,y)/lab_i(x,y) = 1 \; et \; I_i(x,y) \in H_i(k)\}| \cdot \\ & |\{(x,y)/lab_j(x,y) = 1 \; et \; I_j(x,y) \in H_j(k)\}| \big). \end{aligned} \tag{2.16}$$

Pour déterminer la solution optimale, nous avons utilisé une technique sous-modulaire d'optimisation quadratique pseudo booléenne (QPBO) [74]. Cette technique non-itérative permet de résoudre le problème d'optimisation dans un temps polynomial par un simple appel à un algorithme de flot maximum (max-flow) appliqué sur un graphe approprié [150]. Ainsi, la co-segmentation est présentée sous la forme d'un problème d'étiquetage sur un graphe complet, tels que les nœuds sont les pixels. En la combinant avec la technique proposée pour l'appariement des histogrammes, la formulation utilisée produit des segmentations précises (Fig. 2.12). En effet, une étude comparative de la méthode proposée avec une méthode de l'état de l'art [74], qui est aussi basée sur un problème de minimisation de l'énergie également résolu en utilisant une procédure max-flow, a confirmé la précision de notre méthode (Fig. 2.12). Nous avons évalué objectivement les résultats de co-segmentation en mesurant l'erreur de segmentation [112] pour sept paires d'images de référence, et nous avons enregistré un moyen d'erreur de 4.18% pour la méthode proposée contre un moyen de 4.68% pour la méthode introduite dans [74]. En particulier, l'intégration de l'information spatiale nous permet d'extraire des objets complets de l'avant-plan sans aucune sur-segmentation ou sous-segmentation (Fig. 2.12).

2.5.2 Application en détection des athlètes

L'analyse automatique des actions des athlètes dans des vidéos s'est révélée être un outil puissant pour la formation sportive et la visualisation [113]. Elle permet le suivi des actions des athlètes afin de fournir des mesures biométriques et cinématiques de haut niveau. Cette analyse permet également de proposer des solutions efficaces pour l'analyse quantitative, l'évaluation et la comparaison des performances des athlètes. La plupart des travaux soulignent trois grands enjeux des systèmes d'analyse des mouvements des athlètes, à savoir la segmentation des athlètes, le suivi des athlètes et la reconnaissance de leurs actions. Cependant, la méthode globale s'appuie sur les silhouettes des athlètes et les erreurs de segmentation ont tendance à affecter le suivi et la reconnaissance de l'action finale. En effet, ces silhouettes, qui fournissent des informations précieuses sur les positions et les formes des athlètes, sont souvent considérées comme entrée pour les prochaines étapes, et devraient ainsi être de haute qualité avec le minimum d'erreurs de segmentation [119]. Néanmoins, peu de travaux se sont intéressés à la détection des athlètes, et la plupart d'entre eux n'ont pas été validés sur de longues vidéos. Ceci

(a) (b)

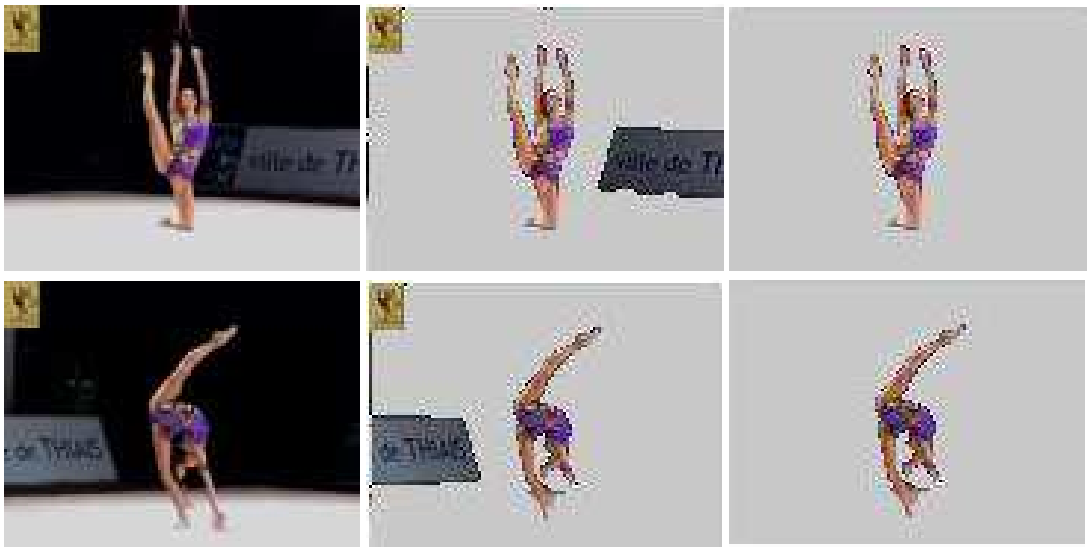
FIGURE 2.12 – Evaluation subjective de la méthode proposée de co-segmentation : (a) comparaison des résultats produits avec ceux de [74] : la première ligne indique la paire d'images, la deuxième et la troisième lignes illustrent les résultats de [74] et nos résultats, respectivement, (b) quelques résultats produits par notre méthode.

est principalement dû à la complexité des arrière-plans dans les vidéos sportives et à la déformation non-rigide des athlètes. En effet, la segmentation automatique des athlètes est un problème difficile en raison de nombreux facteurs, tels que l'occlusion (principalement causée par les arbitres et le public), les mouvements complexes et rapides des athlètes, les changements dynamiques dans les environnements des évènements sportifs (éclairage, arrière-plan...) et les mouvements sans contrainte des caméras. En outre, la plupart des vidéos des activités sportives sont non-calibrées et de mauvaise qualité [124]. Ainsi, afin de pallier aux diverses situations indésirables, nous avons suggéré de réduire ; pour la première fois [126] [38] ; le problème de détection des objets mobiles dans une vidéo à une simple co-segmentation d'images. La principale contribution de la méthode proposée est qu'elle utilise le co-segmentation pour le problème fondamental de segmentation des athlètes dans une grande variété de poses, sans initialisation ou connaissance préalable sur le mouvement de la caméra et les paramètres des athlètes. En effet, vu que le même ensemble d'athlètes apparait dans chaque image de la vidéo, il est intuitivement plus souhaitable de co-segmenter plusieurs images conjointement au lieu de segmenter chacune d'une façon indépendante. Puisque la tâche de co-segmentation suppose que le fond change d'une manière significative dans la paire d'images d'entrée, le premier module de la méthode proposée vise à déterminer, par rapport à chaque image F_t à segmenter ($1 \leq t \leq N$, où N désigne le nombre d'images), une autre image F_{t*} dans laquelle le fond varie suffisament. Sachant que les images successives ont généralement des fonds semblables, nous avons procédé à un échantillonnage temporel afin de définir l'ensemble ζ_t des images candidates pour la co-segmentation avec F_t. Nous avons adapté l'échantillonnage temporel de sorte que l'ensemble ζ_t comprend l'image F_t, afin de garantir que les images sélectionnées soient temporellement éloignées de F_t et les arrière-plans dans ces images sont ainsi suffisamment différents de celui de F_t. Ensuite, parmi l'ensemble ζ_t, l'image F_{t*} avec laquelle F_t sera co-segmentée (2.17) est celle qui maximise la dissemblance, relativement à F_t, en référence à l'intersection des histogrammes.

$$F_{t*} = \arg \min_{F_j \in \zeta_t \setminus F_t} \left(\frac{\sum_{k=1}^{K} \min(H_t(k), H_j(k))}{\min(\sum_{k=1}^{K} H_t(k), \sum_{k=1}^{K} H_j(k))} \right) \tag{2.17}$$

où, $H_t(k)$ (*resp.* $H_j(k)$) est la k-ème composante de l'histogramme H_t (*resp.* H_j).

Le deuxième module co-segmente la paire d'images (F_t, F_{t*}) afin de détecter le masque des athlètes MO_t dans F_t. Cependant, des artefacts de segmentation peuvent être observés dans certains résultats de la co-segmentation. Ceci est principalement dû à la ressemblance des apparences des fonds (surtout en raison des panneaux publicitaires répétitifs) dans les images co-segmentées. En outre, les logos de télévision sont généralement placés de façon statique dans l'un des coins de toutes les images de la vidéo. Ainsi, vu que le contenu vidéo change au fil du temps, sauf pour les régions des logos qui apparaissent plus brillants que le fond [157], les logos sont souvent détectés comme une partie de l'avant-plan. Ces effets pourraient être efficacement évités en appliquant un post-traitement spatial simple. En effet, comme la caméra suit toujours le mouvement des objets mobiles lors de la capture de la scène, il est souvent admis que ces objets sont dans le milieu des images. Cette hypothèse tient pour les vidéos sportives, où l'athlète est généralement l'objet d'intérêt. Ainsi, les régions qui "touchent" les bordures de l'image seront écartées des résultats de la détection des athlètes (Fig. 2.13). Notons que, une fois l'image F_{t*} est déterminée par le premier module, l'image F_t est automatiquement ajoutée à l'ensemble ζ_{t*}, tout en respectant le fait que tous les ensembles de candidats devraient avoir le même cardinal (*i.e.* $\forall t \in \{1, \ldots, N\},\ |\zeta_t| = constante$). Ainsi, si l'image F_t sera plus tard choisie comme image correspondante à F_{t*}, nous évitons de répéter la procédure de co-segmentation afin de détecter les athlètes MO_{t*} dans l'image F_{t*}.



(a) (b) (c)

FIGURE 2.13 – Segmentation des athlètes dans une vidéo de patinage artistique : (a) images d'entrée, (b) résultats de la co-segmentation, (c) segmentation finale après le post-traitement spatial.

L'efficacité de la méthode proposée pour la segmentation des athlètes est démontrée sur une variété de vidéos complexes téléchargées à partir du Web [4]. Ces vidéos ont été acquises avec une seule caméra mobile non calibrée dans des environnements dynamiques et sans contrainte. Dans Fig. 2.14, nous présentons les résultats pour une vidéo de mauvaise qualité de "freefly" (images de taille 100×56, avec une résolution de 72 dpi). Malgré le mouvement libre de la caméra et le manque de l'information texture, la méthode proposée segmente avec précision cette longue vidéo (environ 50 secondes) où les athlètes apparaissent à la fois dans des groupes denses et en tant qu'individus. Fig. 2.15 montre un exemple qui met en évidence la différence des résultats de segmentation, en utilisant la méthode proposée et un

4. http ://www.rsgvideos.com et http ://www.olympic.org

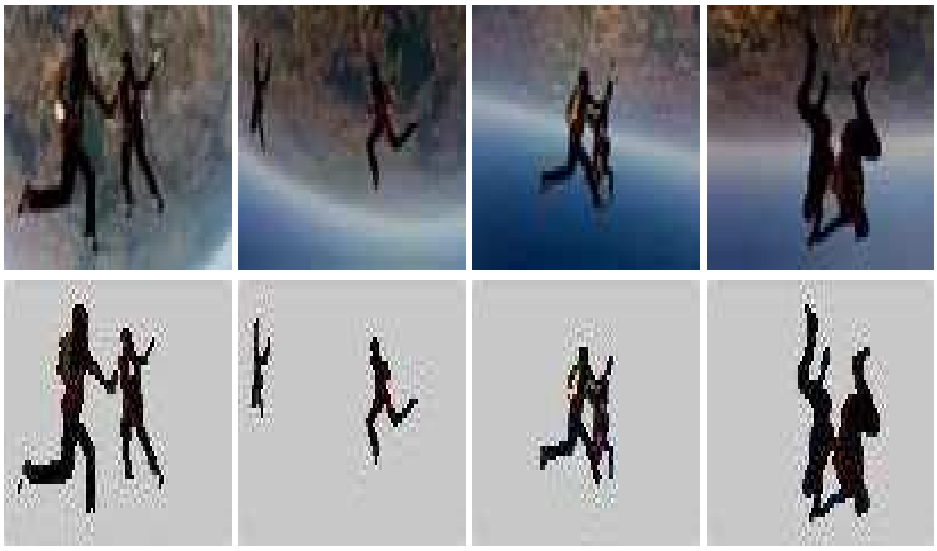


FIGURE 2.14 – Segmentation des athlètes dans une vidéo de "freefly".

procédé classique de soustraction de fond, sur une vidéo de patinage. Cette vidéo, composée de 2993 image de taille 140×105 (environ 130 secondes avec une résolution de 72 dpi), est très complexe en raison de la forte occultation et aux mouvements de la caméra et de l'arrière-plan. La méthode proposée fournit des silhouettes précises des athlètes, comparativement à la méthode comparée, même dans des situations complexes caractérisées par des occlusions étendues et des effets d'ombrage. En effet, les corps des athlètes segmentés, en utilisant des méthodes de soustraction de fond qui sont couramment utilisées pour la détection des athlètes, sont généralement scindés sur plusieurs segments avec beaucoup de détails manquants.

La même vidéo de patinage artistique a été utilisée avec une vidéo du saut en longueur (images de taille 352×288, avec une résolution de 24 dpi), qui est aussi très complexe en raison de l'arrière-plan dynamique et de la vaste mouvement de la caméra, pour l'évaluation objective de la méthode suggérée par rapport à une vérité-terrain (GT) que nous avons généré manuellement (Fig. 2.16). Compte tenu de cette vérité-terrain, nous avons mesuré l'erreur de segmentation e_t (*i.e.* le pourcentage de pixels mal classés sur le nombre total de pixels selon GT) et le pourcentage κ de pixels qui changent de bin par défaut pour chaque image F_t [151]. Ensuite, nous avons calculé les valeurs moyennes $\bar{e}$ et $\bar{\kappa}$ pour les deux séquences, et ceci en faisant varier la valeur du seuil θ (de 0.1 à 0.9) (Fig. 2.17). La méthode proposée extrait précisément les silhouettes d'athlètes puisque nous avons enregistré une erreur de segmentation de 0.047% (*i.e.* une précision de 95.3%), tout en étant peu sensible à la valeur du seuil θ utilisé. En plus, nous avons comparé l'erreur de segmentation e_t de la méthode proposée avec celle obtenue par la méthode utilisée dans [98]. Cette méthode est une version améliorée du procédé de soustraction de fond, qui considère le mouvement de l'objet mobile entre les images en sélectionnant uniquement les images avec un mouvement apparent de l'objet mobile pour construire l'image du fond. Cette comparaison a été réalisée sur un échantillon de 16 images à partir de la vérité-terrain. D'après les résultats enregistrés (Fig. 2.18), il est clair que la méthode proposée ($\bar{e} = 0.0472$) surpasse la méthode comparée ($\bar{e} = 0.1078$). En effet, en utilisant la méthode de [98], les silhouettes des athlètes sont souvent divisées sur plusieurs segments avec beaucoup de détails manquants (trous à l'intérieur des athlètes). En particulier, cette méthode [98] échoue lorsque l'athlète a un léger mouvement pendant les premières images. En effet, des traces fantomatiques apparaissent sur le fond construit, ce qui conduit à une imprécision dans la segmentation de ces images. La méthode proposée est aussi précise que la vérité-terrain. Elle détecte avec précision les athlètes dans des vidéos complexes, ce qui peut être très utile pour l'analyse quantitative de la performance des athlètes. En effet, notre méthode atteint des performances stables, par rapport

(a) (b) (c)

FIGURE 2.15 – Segmentation d'une athlète dans une vidéo de patinage artistique : (a) images d'entrée, (b) et (c) résultats de la segmentation moyennant la méthode proposée et la méthode de soustraction de fond, respectivement.

à la méthode comparée, même si les résultats diffèrent souvent contextuellement de ceux produits par les humains.

2.6 Conclusion

Nos travaux sur la recherche interactive des images par le contenu ont été synthétisés le long de ce chapitre. En effet, nos contributions à ce niveau consistent essentiellement en une modélisation assez riche de l'image traitée, un ensemble d'heuristiques simplifiant la phase de mise en correspondance de deux images, ainsi qu'une technique de bouclage de pertinence opérant sur un RBIR en mettant à jour les pondérations des régions de l'image requête. Le but principal de ce travail était de combiner la recherche d'images à base de régions avec un bouclage de pertinence interactif afin d'améliorer les performances de la recherche des images par le contenu dans des larges collections hétérogènes. Toutes les

FIGURE 2.16 – Évaluation objective de la détection des athlètes : premières images : un échantillon de la vérité-terrain, dernières images : les résultats de la segmentation.

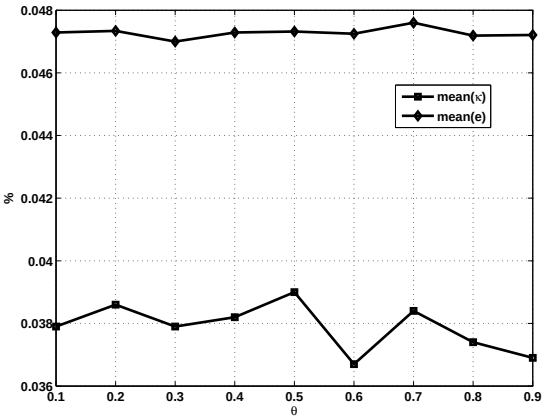


FIGURE 2.17 – Variations de la valeur moyenne de l'erreur e_t et du pourcentage κ, des pixels qui ont changé le bin par défaut, en fonction de la valeur du seuil θ.

images de la collection sont segmentées en régions et des vecteurs descriptifs de l'ensemble des régions extraites sont définis. Ce processus est réalisé hors-ligne et définit l'étape d'indexation de la base. Ensuite, le premier niveau d'interaction en ligne entre l'utilisateur et le système consiste à la présentation d'une image requête. Cette image est ensuite grossièrement segmentée en des régions floues illustrant les objets qui le composent et qui seront indexées par des descripteurs spectraux. La technique la plus utilisée pour décrire une image ou une région sur le plan spectral est celle qui emploie les transformations d'ondelettes afin de réaliser une modélisation sur différentes résolutions. Ces transformations sont basées sur la structuration hiérarchique multi-niveaux qui est implémentée en utilisant une banque de filtrage. Cette modélisation en un ensemble de bandes fréquentielles peut être vue comme étant une segmentation, sauf que cette segmentation ne divise pas l'image en régions mais plutôt en hautes et basses fréquences. Dans notre cas, ces transformations d'ondelettes portent sur les régions de l'image. En effet, nous avons proposé l'utilisation d'un critère fiable pour décrire les régions de l'image qui est l'information spectrale, qui porte essentiellement une combinaison entre l'allure générale et la texture pour la caractérisation des objets composant l'image. Ainsi, une image est représentée sous la forme d'un graphe tel que chaque nœud est marqué par les descripteurs de bas niveau d'une région particulière et il est pondéré selon l'importance visuelle de la région correspondante, alors que les arcs portent l'information de relations spatiales entre les régions. Puis, l'ensemble de matrices spectrales, extrait en ligne de l'image requête et représentant le descripteur de bas niveau, est comparé avec les matrices, enregistrées hors-ligne, des différentes régions de chaque image de la base en utilisant une heuristique de complexité minimale. Notons que nous avons évité la région-requête, que ce soit au niveau de la requête initiale ou au niveau des itérations du bouclage de pertinence, afin de simplifier l'interaction avec l'utilisateur.

En outre, nous avons intégré le bouclage de pertinence dans notre système RBIR afin d'ajouter de la sémantique à la recherche. En effet, notre méthode de bouclage de pertinence est conçue selon la caractéristique de la représentation à base de régions décrites par des sous-bandes d'ondelettes. Le bouclage de pertinence est rarement modélisé avec des descripteurs fréquentiels combinés avec le poids de chaque région. Pour ce faire, nous avons proposé un mécanisme interactif simple, mais efficace, qui se base sur l'adaptation des poids de régions de l'image requête par les rétroactions négatives de l'utilisateur. Ces rétroactions essayent de rapprocher la requête-cible idéale et d'éliminer les mauvaises correspondances tout en essayant de minimiser le taux de similarité entre les objets composants l'image requête et celles des images sélectionnées comme exemples négatifs. Ce bouclage de pertinence, qui fait partie de la classe de représentation vectorielle, perfectionne à un certain niveau les nouvelles collections d'images retrouvées. En plus, le bouclage de pertinence utilisé porte sur l'ajout d'information

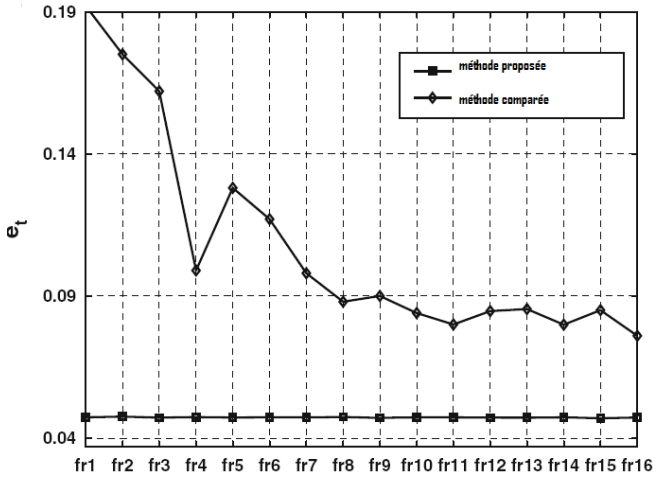


FIGURE 2.18 – Comparaison objective entre la méthode proposés et celle de [98].

fréquentielle afin de s'approcher mieux des attentes de l'utilisateur. C'est-à-dire qu'au lieu d'utiliser seulement une sous bande de haute fréquence, il sera possible d'ajouter les données apportées par les autres bandes fréquentielles à chaque itération de bouclage de pertinence exigée par l'utilisateur. Les expérimentations et l'étude comparative avec des approches similaires de l'état de l'art ("SIMPLIcity" et "Anaktisi") prouvent la robustesse de l'approche proposée pour la recherche des images en termes d'apport sémantique offert par la modélisation riche des images par des graphes complets ainsi que par le bouclage de pertinence. Ceci est essentiellement dû à la combinaison entre une modélisation RBIR, un BP négatif de modèle vectoriel et une description des régions uniquement par la sous-bande de haute fréquence du deuxième niveau des ondelettes. En particulier, grâce à la considération de l'information spatiale, il est toujours possible pour notre approche de procéder à des corrections supplémentaires après l'appariement du contenu visuel. De là, nous pouvons considérer notre approche comme plus simple que les approches comparées tout en étant plus performante vu les résultats produits sur les mêmes bases standards d'images.

Par ailleurs, étant convaincus que la co-segmentation des images peut apporter des solutions pertinentes pour la recherche et la classification des images par le contenu, nous avons abordé cet axe de recherche. Cette co-segmentation revient souvent à la résolution d'un problème d'optimisation qui vise à minimiser une fonction d'énergie, qui intègre un terme de données intrinsèques relatives aux images traitées, un terme de lissage qui favorise la segmentation lisse de chaque image, et un terme de correspondance qui pénalise la dissimilarité entre les images. En plus de l'utilisation de l'entropie locale lors de la caractérisation d'une image par son histogramme, la principale contribution de la méthode proposée réside dans la classification floue de l'entropie locale. Ceci permet de réduire l'ambiguïté d'appartenance d'un pixel à un bin d'un histogramme, tout en imposant la cohérence spatiale des pixels voisins afin d'éviter les fausses détections et les effets du bruit. Les résultats préliminaires ont prouvé l'efficacité de la technique proposée pour la co-segmentation des images. Cette technique a été par le suite adaptée pour la segmentation des athlètes, et ceci en réduisant cette segmentation à une simple co-segmentation d'une paire d'images. En effet, la co-segmentation intègre implicitement l'information temporelle pour la segmentation automatique des athlètes sans aucune hypothèse ou connaissance préalable sur le mouvement de la caméra. Les premier résultats obtenus ont prouvé une amélioration significative apportée par la méthode introduite par rapport aux procédés de soustraction du fond, qui sont couramment utilisés pour la segmentation des athlètes. Notons enfin que même si elle n'a été appliquée que sur des vidéos sportives, la méthode proposée peut être facilement adaptée pour détecter d'autres objets mobiles dans le cadre d'autres types d'applications (*e.g.* surveillance des piétons). En effet, elle fonctionne, selon l'application, sur mono ou multi caméras statiques ou mobiles.

Chapitre 3

Aide au diagnostic médical basé sur l'analyse des images

3.1 Introduction

Développés lentement en cachette, les symptômes cliniques des cancers ne se manifestent qu'à des stades avancés, souvent incurables. Dès lors, un dépistage systématique, dans lequel les personnes passent régulièrement un examen permettant de vérifier leurs états, s'avère incontournable. Dans ce contexte, l'imagerie médicale se présente comme la solution non-invasive la plus efficace permettant de détecter les anomalies à un stade préclinique, voire même précancéreuse. Ainsi, plusieurs pays ont commencé des programmes organisés de dépistage des cancers basés sur cet outil de diagnostic précoce. Toutefois, les limitations liées tant aux images qu'aux cliniciens influent énormément sur la qualité de diagnostic et restreignent par conséquent le succès de ces programmes. D'une part, ces programmes impliquent un nombre continuellement croissant d'images difficiles à analyser, nécessitant souvent le recours à une deuxième lecture. D'autre part, les radiologues retournent souvent des diagnostics incorrects soit par fatigue ou omission, soit par manque d'expérience. De là, un intérêt croissant a été accordé pour les systèmes automatiques et semi-automatiques d'aide au diagnostic médical basé sur l'analyse des images. La décision finale étant prise par le clinicien, les systèmes d'aide au diagnostic ont pour vocation soit d'assister le radiologue, en localisant les éventuelles anomalies ou en quantifiant certaines caractéristiques associées, soit de jouer le rôle du second radiologue. C'est dans ce cadre que s'inscrit notre travail dont l'objectif est de développer un système générique et fiable pour l'aide au diagnostic basé sur les images médicales. La fiabilité concerne l'aptitude du système à retenir tous les cas malins et rejeter tous les cas bénins. En effet, rejeter de cas malins (faux négatifs) signifie que des cancers détectables sont ratés, alors que retentir des cas bénins (faux positifs) résulte en des conséquences morales et des traitements supplémentaires inutiles. En outre, sachant qu'un cancer peut se manifester par différents symptômes, un système est dit générique s'il est capable de détecter toutes ses anomalies. Une synthèse bibliographique des systèmes existants d'aide au diagnostic médical basé sur les images, nous a permis de les regrouper en trois grandes classes, à savoir ceux basés sur la classification, ceux basés sur le recalage et enfin ceux basés sur la recherche par le contenu (CBIR). Toutefois, ces trois approches présentent des limitations lorsqu'elles sont utilisées séparément. En effet, chaque approche est appropriée seulement à des symptômes particuliers et elle est souvent complémentaire avec les autres approches. En l'occurrence, inapte pour décider de l'état d'une lésion à partir d'une seule image, tâche possible en utilisant des techniques de classification, seul le recalage permet de détecter l'évolution temporelle d'une lésion. Ainsi, et afin de profiter de leurs complémentarités, nous avons proposé d'utiliser conjointement les trois approches. En effet, l'architecture proposée comporte quatre modules : un module de classification, un module de recalage, un module de RBIR et un module de fusion. Les trois premiers modules, qui peuvent être exécutés en parallèle, produisent trois opinions floues sur la malignité du patient, et le dernier module fusionne les trois premières opinions afin de produire une opinion finale. Le système

retourne également les résultats de recalage ainsi que les images visuellement similaires à l'image objet de diagnostic, donnant ainsi une aide intuitive supplémentaire aux cliniciens (Fig. 3.1). Profitant de l'expérience requise en diagnostic du mélanome [165] [166], nous avons commencé par valider l'approche de fusion proposée dans le cas de dépistage du mélanome (cancer de la peau) [21]. Pour ce cas, le recalage n'est pas envisageable vu que chaque lésion est décrite par une seule image. En effet, l'architecture proposée se compose d'un module de classification par les réseaux de neurones, d'un module de recherche des images par le contenu et d'un module de fusion par la théorie de Dempster-Shafer. Les résultats produits montrent que la fusion des résultats obtenus par les différentes approches permet d'améliorer la qualité du diagnostic. En outre, dans le même contexte d'aide au diagnostic du mélanome, nous avons proposé une méthode structurale pour la détection automatique du réseau de pigment, qui est un symptôme très important pour la malignité de plusieurs lésions de la peau. Cette méthode a assuré une aire sous la courbe ROC de 0.821 pour détecter avec succès les réseaux de pigment avec un taux de classification correcte de 85% sur un ensemble de 122 images réelles. Ensuite, en visant l'application de l'architecture proposée dans le contexte du cancer des seins, nous nous sommes intéressés au problème de recalage d'images médicales. Ce recalage est souvent heurté à des contraintes de temps de calcul, d'interaction avec l'utilisateur et de qualité de recalage. L'idéal étant de développer des méthodes de recalage automatiques, rapides et capables de gérer les différences inter-images. Une revue de la littérature nous a conduit à déduire que la méthode non-rigide de subdivision progressive offre un bon compromis entre les différentes contraintes. Cependant, des améliorations peuvent être introduites à cette méthode aussi bien en temps de calcul qu'en qualité de recalage. Dans cette perspective, nous avons proposé d'appliquer la stratégie de subdivision à une pyramide d'images, au lieu de l'appliquer à une image. Les résultats obtenus, sur des mammographies et sur des IRMs cérébrales, montrent que la technique proposée permet de diminuer le temps de calcul sans pourtant dégrader la qualité de recalage. D'un autre côté, nous nous sommes récemment intéressés à la reconstruction des modèles 3D dans le contexte de la chirurgie assistée par ordinateur. Nous avons débuté par une amélioration de la méthode de coloration des voxels, qui consiste à affecter des couleurs aux voxels dans un volume tridimensionnel, tout en garantissant la cohérence avec l'ensemble des images 2D en entrée. La qualité de reconstruction de cette méthode dépend fortement d'une étape de seuillage permettant de déterminer, pour chaque voxel, s'il est photo-consistant ou non. Toutefois, en plus de l'absence de toute information sur le voisinage du voxel pendant le test de la photo-consistance, il est extrêmement difficile de définir les seuils appropriés, qui doivent être précis et stables pour tous les voxels de la surface. Notre contribution réside dans l'automatisation du test de photo-consistance, et ceci en faisant recours à un seuillage adaptatif par hystérésis. En plus, la méthode proposée n'a pas besoin de seuils prédéfinis puisque ceux d'hystérésis sont définis automatiquement et de manière adaptative en fonction du nombre des images sur lesquelles le voxel est projeté. Les résultats préliminaires ont montré que la méthode proposée est capable de produire automatiquement des reconstructions volumétriques précises et lisses.

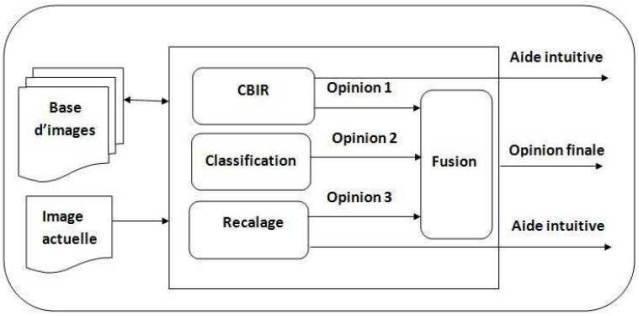


FIGURE 3.1 – Architecture proposée pour l'aide au diagnostic basé sur les images médicales, utilisant conjointement le recalage, la classification et la recherche des images par le contenu (CBIR).

3.2 Aide au diagnostic du mélanome

Le cancer de la peau est devenu l'un des cancers les plus répandus dans le monde, dont la forme la plus mortelle est le mélanome. Ce dernier correspond à une tumeur cancéreuse développée à partir des cellules pigmentaires de la peau et qui peut développer des métastases et envahir d'autres parties du corps. En pratique, la méthode non-invasive la plus efficace pour la détection précoce de cette maladie est le dépistage par des techniques d'imagerie médicale, dont la plus efficace est la dermoscopie. Dans ce contexte, nous avons validé l'architecture standard proposée dans le cas du mélanome. Toutefois, vu que chaque lésion de la peau est décrite par une seule image dermoscopique, le module de recalage n'est pas envisageable dans ce cas. Ainsi, le système proposé est composé d'un module de classification, d'un module de CBIR et d'un module de fusion des résultats produits par les deux premiers modules. Ces derniers incluent trois étapes communes, à savoir le pré-traitement de l'image, l'extraction de la lésion sujet du diagnostic et la caractérisation des signes cliniques de malignité de cette lésion. Pour le pré-traitement, nous avons utilisé un filtre moyen pour le débruitage des images et la technique "*Dullrazor*" pour la suppression des poils [165]. Les lésions sont ensuite segmentées en utilisant une technique floue de croissance des régions appliquée sur une image de haute contraste [166]. Ensuite, chaque lésion segmentée est décrite par 21 attributs caractérisant sa forme, sa texture et sa couleur. Ces attributs permettent de quantifier la règle "*ABCD*" [123] [45], souvent utilisée par les dermatologues pour décider de la malignité d'une lésion [1]. Le contenu visuel de chaque image est ainsi représenté par l'ensemble des 21 attributs, qui fera l'entrée d'une procédure de classification et d'une procédure de recherche par le contenu. Chacune de ces deux procédures cherche à estimer un degré de malignité de la lésion en question. Les résultats des deux procédures sont ensuite fusionnés pour produire une décision diagnostique finale. Par ailleurs, nous avons aussi introduit une méthode structurale pour la détection automatique du réseau de pigment dans une lésion segmentée. En effet, ces dernières années, il y a eu un intérêt croissant pour l'utilisation des caractéristiques de texture, notamment les réseaux de pigment, comme symptômes de malignité pour les lésions de la peau.

3.2.1 Classification des images

La procédure de classification vise à associer la lésion en entrée, en fonction de son vecteur descriptif, à une classe diagnostique (maligne *vs.* bénigne). Cette association est réalisée en deux étapes, à savoir l'apprentissage et la reconnaissance. Le rôle de l'apprentissage est d'éclairer la décision à l'aide des connaissances a priori sur l'organe étudié. En effet, à partir des paramètres spécifiques aux lésions, l'apprentissage définit des modèles de référence caractérisant les classes diagnostiques. Quant à la reconnaissance, elle consiste à identifier la classe diagnostique d'une nouvelle lésion test en fonction des modèles de référence définis par le processus d'apprentissage. Parmi les méthodes de classification largement utilisées dans le contexte du mélanome, les réseaux de neurones fournissent fréquemment des taux d'erreur réduits comparés à des approches statistiques plus conventionnelles [102] [2]. En effet, les cliniciens se basent sur leurs connaissances patho-physiologiques et sur leurs expériences, et une telle expérience n'est pas modélisable par un petit ensemble de relations. Ceci limite l'intérêt de la classification par les approches algorithmiques, vu qu'il est très difficile de créer une base complète de connaissances symboliques à cause des exceptions qui se produisent dans la pratique. Ainsi, la possibilité d'apprentissage fortement autonome par des exemples fait des réseaux de neurones une solution adaptée pour les applications du diagnostic médical. L'architecture des réseaux de neurones la plus commune pour une classification supervisée est le perceptron multicouches, qui est caractérisé par sa puissante capacité d'approximation universelle [2]. De là, nous avons choisi d'utiliser un perceptron multicouche dont l'architecture est définie par n neurones d'entrée représentant les attributs de la lésion ($n = 21$), m neurones cachés et p neurones de sortie, où p est le nombre des classes diagnostiques ($p = 2$). L'apprentissage de ce perceptron est basée sur l'algorithme de rétropropagation du gradient, qui minimise

1. La règle "*ABCD*" analyse quatre caractéristiques cliniques afin d'identifier une lésion maligne, à savoir l'Asymétrie, l'irrégularité des Bords, la variation de la Couleur et le Diamètre.
2. Il peut approximer toute fonction avec une précision qui croit en fonction du nombre des neurones cachés.

l'erreur quadratique entre les sorties réelles du réseau et les sorties désirées. Pour ce faire, nous avons utilisé une base d'apprentissage composée de 200 images avec leurs vérité-terrains. L'apprentissage revient ainsi à adapter les paramètres non-fixés du perceptron (poids et seuils) à travers un processus de simulation supervisé. En effet, l'algorithme d'apprentissage par rétropropagation consiste à réaliser deux passages à travers les différentes couches du réseau : un passage en avant et un passage en arrière. Dans le passage en avant, l'effet de chaque vecteur en entrée se propage couche par couche. Un ensemble de sorties est produit pour donner la réponse courante du perceptron, et les poids du réseau sont ainsi estimés dynamiquement. Le passage en arrière permet de rétro-propager les signaux d'erreurs dans le perceptron en calculant récursivement le gradient local pour chaque neurone, afin d'ajuster tous les poids du perceptron conformément à la règle de correction des erreurs [165]. Toutefois, l'inconvénient majeur de cet algorithme d'apprentissage réside dans son temps de calcul, notamment lorsque la séparation entre les différentes classes de décision est complexe et quand le nombre des couches cachées est grand. Afin d'obtenir des performances satisfaisantes en un temps d'apprentissage acceptable [166], nous avons utilisé une analyse en composantes principales (ACP) qui permet de décrire chaque lésion par q nouvelles variables ($q < n$). Ces nouvelles variables, qui correspondent aux composantes principales, sont des combinaisons linéaires des attributs originaux tout en étant plus discriminantes. Dans notre cas, étant donné les attributs correspondants aux 200 images de la base d'apprentissage, nous avons utilisé l'ACP pour choisir à partir des 21 composantes principales celles les plus efficaces pour le processus de classification. Pour cela, nous avons calculé le taux de classification pour toutes les sous-familles possibles composées des q premières composantes principales ($1 < q < 21$). En outre, pour chaque sous-famille, nous avons calculé le taux de classification tout en variant le nombre m des neurones cachés, sachant que la valeur de m doit être strictement inférieure à celle de q. La valeur de m influence largement la capacité du perceptron à déterminer les propriétés des données qui n'appartiennent pas à la base d'apprentissage. En effet, la distance entre l'approximation obtenue par le réseau de neurones et la fonction de classification recherchée est inversement proportionnelle au nombre m des neurones cachés [26]. Toutefois, ce résultat n'est pas constructif, dans le sens qu'il ne peut donner que des appréciations vagues sur le nombre nécessaire de neurones cachés [53]. Dans notre cas, nous avons testé toutes les valeurs possibles de m jusqu'à l'obtention des performances optimales sur la base de test. Nous avons trouvé que le meilleur résultat est obtenu en utilisant un perceptron avec 16 neurones dans la couche d'entrée et 5 neurones dans celle cachée (Fig. 3.2). En effet, cette architecture assure un taux de reconnaissance supérieur à 87% sur une base de 200 images. Les courbes ROC (*Receiver Operating Characteristic*) obtenues, avec les meilleures architectures pour sept sous-familles différentes ($15 \leq q \leq 21$), confirment l'efficacité de l'architecture retenue (Fig. 3.3).

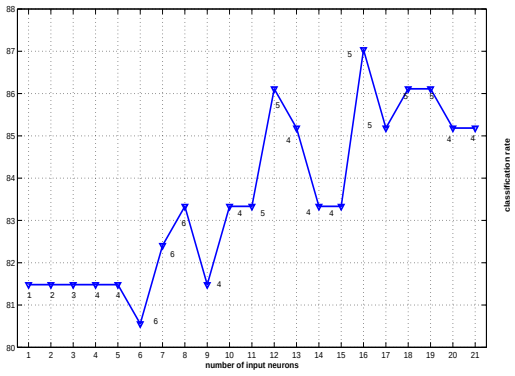


FIGURE 3.2 – Comparaison des taux de classification pour plusieurs architectures du perceptron ($1 \leq q \leq 21$), sachant que le nombre optimal m des neurones cachés est affiché pour chaque valeur de q.

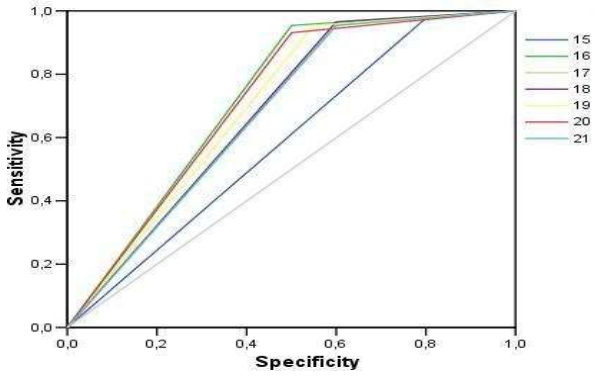


FIGURE 3.3 – Les courbes ROC des meilleures architectures ($15 \leq q \leq 21$).

3.2.2 Recherche des images par le contenu

La nature des lésions de la peau dans les images dermoscopiques (variation de la couleur, structures différentielles...) les présente comme une application idéale pour la recherche des images par le contenu [123]. La procédure de recherche consiste à trouver les images des lésions, dans une base d'images avec sa verité-terrain, qui ressemblent le plus à la lésion dans l'image question du diagnostic, et ceci en comparant leurs vecteurs descriptifs. Dans cette perspective, nous avons comparé quatre distances (Euclidienne, City-block, Cosine et Mahalanobis) et les résultats obtenus montrent que la distance de Mahalanobis est la plus appropriée pour notre cas (Fig. 3.4). Ainsi, les k images $\{I_j,\ j = 1 \dots k\}$ les plus similaires à l'image requête I_{req} sont celles maximisant la mesure de similarité $S(I_j, I_{req})$, qui est l'inverse de la distance de Mahalanobis. Etant donné la petite taille de notre base (400 images), le risque d'avoir des images avec de faibles degrés de similarité, par rapport à l'image en entrée, est élevé pour les grandes valeurs de k. Pour cela, nous nous sommes limités à une valeur de k égale à 5. Ensuite, sachant que les images retournées sont triées selon leurs degrés de similarité vis-à-vis de l'image requête (*i.e.* $S(I_1, I_{req}) \geq S(I_2, I_{req}) \geq \ldots \geq S(I_k, I_{req})$), le degré d'appartenance $\mu_0(I_{req})$ de l'image I_{req} à la classe des mélanomes malins C_0 est défini comme suit :

$$\mu_0(I_{req}) = \frac{1}{15}.\sum_{j=1}^{k}(k+1-j).S_0(I_j, I_{req}), \tag{3.1}$$

avec,

$$S_0(I_j, I_{req}) = \begin{cases} S(I_j, I_{req}) & \text{si } I_j \in C_0, \\ 1 - S(I_j, I_{req}) & \text{si } I_j \notin C_0. \end{cases} \tag{3.2}$$

Ainsi, si $\mu_0(I_{req}) > 0.5$ alors la lésion en entrée est considérée comme maligne, avec un taux de certitude de $\mu_0(I_{req})$, sinon elle est associée à la classe des lésions bénignes, avec un taux de certitude de $(1 - \mu_0(I_{req}))$. Les premiers résultats obtenus ont montré que l'approche CBIR proposée permet un taux de reconnaissance supérieur à 67% sur une base de 200 images (Fig. 3.5). En plus, dans l'ambition d'offrir une aide intuitive aux dermatologues, la solution proposée produit aussi les images qui ressemblent visuellement à l'image diagnostiquée sans pourtant avoir le même état clinique. Pour cela, nous avons appliqué la même procédure en combinant les 21 attributs, que nous avons déjà utilisés lors de la recherche, avec les 7 premiers moments invariants de Hu [77]. L'objectif était de combiner les symptômes de malignité d'une lésion (les attributs) avec son contenu visuel (les moments de Hu) (Fig. 3.6).

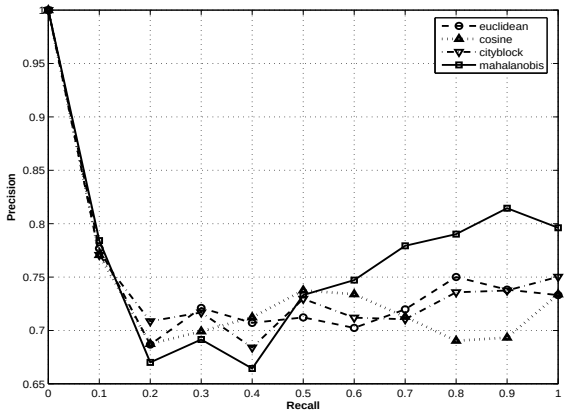


FIGURE 3.4 – Courbes Rappel/Précision pour la recherche des images avec différentes distances.

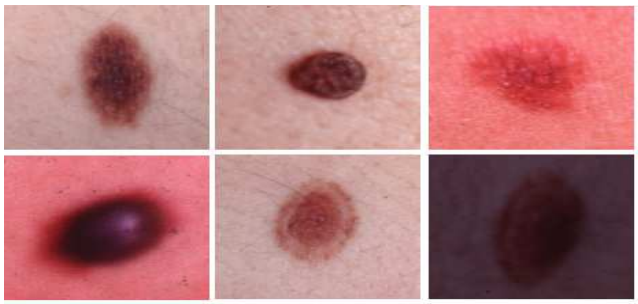


FIGURE 3.5 – Exemple des résultats du CBIR : la première image est la requête et les images suivantes sont les 5 premières images retournées, dont les mesures de similarité successives sont : 0.77, 0.72, 0.71, 0.69 et 0.68 (notons que les 6 lésions sont bénignes et la valeur de $\mu_0(I_{req})$ est égale à 0.18).

3.2.3 Fusion des opinions diagnostiques

Ayant les deux opinions diagnostiques fournies par la procédure de classification et celle de recherche des images par le contenu, l'objectif de cette étape de fusion est de combiner ces deux opinions pour aboutir à une troisième opinion plus précise. La méthode de fusion proposée considère le résultat produit séparément par chaque procédure comme étant une source incertaine sur la malignité de la lésion en entrée. Elle est basée sur la théorie de Dempster-Shafer [135] en considérant les résultats fournis par les procédures de classification et de CBIR comme des probabilités postérieures sur la masse de croyance que la lésion en question appartient à la classe des mélanomes malignes C_0 et à celle des lésions bénignes C_1. Cette théorie est souvent décrite comme la généralisation de l'inférence Bayésienne au traitement de l'incertain, dans le sens qu'elle permet de manipuler des évènements non nécessairement exclusifs. Cette capacité lui confère l'avantage de pouvoir représenter explicitement l'incertitude vis-à-vis d'un événement, ce qui permet d'optimiser la fusion des informations multi-sources afin d'accéder à une information plus fiable [25] [61]. Pour ce faire, la théorie de l'évidence se fonde sur des degrés de croyance permettant de représenter l'incertitude présente sur les différentes propositions (simples et composites). Dans notre cas, étant donné une image d'entrée I_{req}, nous avons considéré les résultats produits par le schéma de classification (*resp.* CBIR) comme une source primaire S^{NN} (*resp.* S^{CBIR}) sur les masses de croyance, $\mu_0^{NN}(I_q)$ et $\mu_1^{NN}(I_q)$ (*resp.* $\mu_0^{CBIR}(I_q)$ et $\mu_1^{CBIR}(I_q)$), de la lésion aux classes C_0 et C_1 [3]. Les

3. Dans notre cas, l'ensemble universel $\{C_0, C_1\}$ est exhaustive ce qui nous permet de travailler sous l'hypothèse du monde clos [106], et le cadre de discernement est ainsi constitué de deux singletons $\{C_0\}$ et $\{C_1\}$.

FIGURE 3.6 – Aide intuitive aux cliniciens : la première image est la requête et les images suivantes sont les 5 premières images retournées, en considérant aussi les moments de Hu, dont les mesures de similarité successives sont : 0.47, 0.45, 0.44, 0.43 et 0.42 (notons que seule la dernière image est maligne).

masses de croyance $\mu_0^{NN}(I_q)$ et $\mu_1^{NN}(I_q)$ correspondent aux valeurs normalisées des deux neurones de la couche de sortie du perceptron, alors que $\mu_0^{CBIR}(I_q)$ et $\mu_1^{CBIR}(I_q)$ correspondent successivement à $\mu_0(I_q)$ et $1-\mu_0(I_q)$. Ce jeu de masses m^{S_k} modélise le degré d'appartenance de la lésion à la classe maligne et à celle bénigne. De là, et vu que le conflit entre les deux sources est rare, nous avons procédé à une combinaison normalisée $\mu_c^{NN\bigcap CBIR}(I_q)$ (avec, $c \in \{1,2\}$) des ces sources via la règle de combinaison de Dempster (3.3), qui est un outil efficace et particulièrement adapté à la fusion des informations imparfaites. Cette règle consiste à construire une fonction de masse unique $\mu_c^{NN\bigcap CBIR}$ par combinaison des deux fonctions de masses issues de deux sources d'information distinctes et indépendantes (μ_c^{NN} et μ_c^{CBIR} dans notre cas). Ainsi, le jeu de masses résultat synthétise la connaissance globale issue des deux opinions diagnostiques (classification et CBIR) sur la malignité de la lésion en entrée.

$$\mu_c^{NN\bigcap CBIR}(I_q) = \frac{1}{1-k_\phi}.\mu_c^{NN}(I_q).\mu_c^{CBIR}(I_q), \tag{3.3}$$

avec,

$$k_\phi = \mu_0^{NN}(I_q).\mu_1^{CBIR}(I_q) + \mu_1^{NN}(I_q).\mu_0^{CBIR}(I_q). \tag{3.4}$$

La valeur de $1/(1-k_\phi)$, évaluant la concordance des opinions diagnostiques fusionnées, peut être considérée comme un taux de confiance en le résultat du diagnostic produit (3.4). Les résultats obtenus ont prouvé que la combinaison des opinions par la théorie de Dempster-Shafer améliore le taux de diagnostic ($> 89\%$), notamment pour les cas douteux lorsque les résultats obtenus par la procédure de classification et/ou celle de CBIR ne sont pas suffisamment discriminants (*i.e.* $\mu_0^{NN}(I_{req}) \simeq \mu_1^{NN}(I_{req})$ et/ou $\mu_0^{CBIR}(I_{req}) \simeq \mu_1^{CBIR}(I_{req})$). En effet, en présence d'information imparfaite (en particulier douteuse), la combinaison des avis multi-sources est une solution intéressante pour accéder, en général, à une information plus fiable. En outre, cette combinaison permet de renforcer la croyance sur les propositions pour lesquelles les sources sont concordantes. En particulier, le taux de confiance permet d'alerter le dermatologue dans le cas d'un conflit considérable entre S^{NN} et S^{CBIR}.

3.2.4 Détection automatique du réseau de pigment

Au cours des dernières années, et afin de réduire le temps de réponse tout en optimisant la précision du diagnostic, il y a un intérêt croissant pour l'utilisation des caractéristiques texturales comme symptômes de malignité du mélanome [108], notamment dans le cadre de la règle du diagnostic "7-point checklist". Le réseau de pigment (RP) est l'une de ces caractéristiques essentielles qui se réfèrent à la fois aux attributs chromatiques, à la forme et aux attributs texturaux de la lésion de la peau. Ce réseau, qui est un type de grille constituée de lignes pigmentées et des trous hypo-pigmentés, peut être absent,

typique ou atypique (Fig. 3.8). Un réseau de pigment typique (RPT) est une grille de clair au brun foncé avec des petits trous uniformément espacés et homogènes en couleur. En plus, il est caractérisé par des lignes minces distribuées plus ou moins régulièrement tout au long de la lésion et généralement éclaircies à la périphérie. Un réseau de pigment atypique (RPA), qui est un signe puissant de malignité, est non uniforme, de couleur noir, marron ou gris avec des lignes foncées et/ou élargies et des trous irrégulièrement réparties et qui sont hétérogènes en surface et en forme. Les lignes sont épaisses et souvent hyper-pigmentées et se terminent brusquement à la périphérie [60]. Quelques travaux basés sur l'analyse de la texture ont été développés pour la détection du réseau de pigment. Ces travaux peuvent être regroupés en deux catégories principales : les méthodes basées sur les caractéristiques statistiques et celles basées sur les caractéristiques structurelles. Les caractéristiques structurelles caractérisent les propriétés texturales des primitives, telles que leurs tailles et leurs formes, tandis que les caractéristiques statistiques mesurent les variations des niveaux de gris. En effet, les méthodes statistiques définissent la texture en termes de statistiques locales, qui varient lentement sur une région texturée, telles que les masques d'énergie de Laws, les matrices de dépendance de voisinage en niveaux de gris (NGLDM) [8] et les descripteurs statistiques de la texture [139]. Les inconvénients majeurs de ces méthodes sont les fausses classifications (dues à la sensibilité du choix des seuils et des distances), le coût élevé en termes de temps de calcul (à cause de l'utilisation de plusieurs mesures), et le calcul local des statistiques. En effet, chacune de ces statistiques décrit la texture en termes des descripteurs locaux sur chaque région de l'image indépendamment des autres régions. Pour remédier à ces inconvénients, les méthodes structurelles essayent de déterminer les primitives qui composent la texture dans la lésion, afin de caractériser leurs propriétés texturales telles que la taille et la forme, sans mesurer les variations d'intensités dans un voisinage. Cependant, la plupart de ces méthodes sont très sensibles aux paramètres utilisés, notamment les seuils qui sont généralement définis d'une manière empirique.

Dans l'objectif de réduire les erreurs de classification tout en minimisant l'effet du choix des seuils, nous avons proposé une méthode structurelle pour décider de la présence ou de l'absence d'un réseau de pigment dans une image dermoscopique. Cette méthode permet de détecter automatiquement le réseau de pigment, en présence d'autres structures telles que les points et les globules, tout en reconnaissant le type du réseau de pigment (typique *vs.* atypique) par analyse de l'uniformité spatiale des mailles. Elle est basée sur le fait que les bords des structures du réseau de pigment forment des graphes cycliques qui correspondent aux structures de texture de la lésion [129]. En effet, suite à un pré-traitement qui vise à rendre la lésion plus nette et la transformer en niveaux de gris, un filtre LoG est appliqué pour de détecter des trous et d'autres structures au sein de la lésion. L'objectif du pré-traitement est de rendre la lésion plus nette et ceci en effectuant deux opérations successives avant de détecter les trous du RP. Tout d'abord, pour mettre en évidence les caractéristiques de texture, un masque flou est appliqué pour mieux apparaitre les bords et les détails au sein de la lésion. Ensuite, uniquement le canal vert est retenu pour étudier les structures de la texture, ce qui permet de réduire la complexité des traitements tout en optimisant la détection et la reconnaissance du RP (Fig. 3.7). En effet, en comparant les différentes transformations de couleur (*e.g.* RGB, NTSC, YIQ, HSV), il a été confirmé que le canal vert produit les meilleurs taux de classification et de reconnaissance [18]. Ayant la lésion dans l'image de luminance verte, un filtre LoG détecte par la suite les trous ainsi que d'autres structures au sein de cette lésion. Grâce à ses propriétés intrinsèques, le filtre LoG détecte avec précision les changements "lumière-obscurité-lumière" de l'intensité dans l'image traitée, suite à une étape de lissage avec un filtre passe-bas afin de réduire les effets du bruit. La sortie du filtre LoG est une image binaire englobant tous les contours fermés (Fig. 3.7) qui correspondent aux passages par zéro de la dérivée seconde. Ceci permet une meilleure localisation de mailles, en particulier lorsque les bords ne sont pas très pointus. Toutefois, cette image binaire contient, en plus des trous potentiels du RP, d'autres trous non significatifs, des lignes et des contours ouverts. Ainsi, afin de ne sélectionner que les trous appartenant au RP, un filtre de taille suivi par un autre filtre d'intensité sont appliqués. Le filtre de taille permet de définir la même résolution pour toutes les images. Dans notre cas, nous avons choisi un intervalle $[a, b]$ de taille plus grande que celle choisie par des travaux antérieurs, afin de trouver autant de mailles du RP que possible. En effet, dans les travaux précédents [108] [129], l'intervalle de taille est utilisé pour filtrer les trous appartenant au RP, tels qu'un trou n'est considéré comme une partie du RP, que si sa taille appartient à cet intervalle. Toutefois, cette

décision binaire peut générer des erreurs de classification, notamment pour les régions avec des tailles autour des limites de l'intervalle de taille. Par exemple, quand une région est composée de $b+1$ pixels, elle est catégoriquement retiré pendant le test fondé sur la taille, contrairement à toute autre région dont le nombre de pixels est b. Ces régions retirées ne peuvent pas être récupérées et peuvent être ignorées par erreur dans un effet de cascade. Pour contourner cette limitation, nous avons introduit une fonction d'appartenance probabiliste ω (3.5), qui produit une décision floue pour chaque région en fonction de sa taille s. Ceci permet une meilleure modélisation de l'incertitude et de l'ambiguïté de la décision de retrait d'une région lors du premier test, et ceci en gardant le maximum de régions candidates pour la prochaine étape tout en reportant cette décision jusqu'à ce que plus d'informations soient disponibles lors des étapes suivantes. En effet, la principale contribution de la méthode proposée réside dans l'évaluation floue du degré d'appartenance d'un trou au réseau de pigment. Les trous retenus sont ensuite reliés, tout en vérifiant une contrainte spatiale, via un graphe représentant le réseau de pigment. Dans notre cas, la taille de l'intervalle a été implicitement intégrée dans la fonction d'appartenance tout en utilisant un intervalle plus grand pour garder le maximum de candidats pour le filtrage basé sur l'intensité. Ainsi, le seuillage proposé évite la suppression des régions du RP, après le filtrage basée sur la taille, qui ne peuvent pas être récupérées ultérieurement (Fig 3.7).

$$\omega(s) = e^{-\frac{1}{2}(\frac{s-c}{\sigma})^2}, \tag{3.5}$$

avec, σ est l'écart type et c est le centre de l'intervalle $[a, b]$.

L'étape suivante vise à traiter les bulles d'huile et les kystes blancs, qui apparaissent avec un très haut niveau d'intensité, tout en évitant l'élimination des régions du RP qui contiennent des pixels de bruit. Dans cette perspective, nous avons enlevé chaque région contenant plus que cinq pixels avec une valeur supérieure à 80% du maximum d'intensité et un degré d'appartenance $\omega(s)$ inférieur à 0.5. Ceci permet d'exclure les bulles d'huile et les kystes blancs et les points, qui sont similaires aux trous du RP en termes d'intensité moyenne de l'intérieur et de la zone de frontière, tout en étant beaucoup plus lumineux à l'intérieur. En particulier, contrairement aux travaux de [129] où un seul pixel est utilisé, le traitement proposé permet d'éviter la suppression des trous bruyants significatifs au sein du RP (Fig. 3.7). En outre, un filtrage basé sur l'intensité est utilisé pour traiter les globules et les points bruns, dont l'intensité moyenne de la zone intérieure de la structure est plus faible que celle sur les pixels de la frontière. Pour cela, nous avons choisi un seuil pour la différence entre l'intensité moyenne des pixels intérieurs et l'intensité moyenne sur les frontières. Une région n'est considérée comme faisant partie du RP (Fig. 3.7), que si son intensité est plus élevée dans la zone incluse dans les structures du réseau (trous) que sur les bords du réseau (lignes). Après l'extraction des trous et pour visualiser le réseau de pigment, un graphe est défini. Les nœuds sont les centres des trous du RP, et les nœuds enregistrant une distance maximale (MDT, pour Maximum Distance Threshold) sont reliés entre eux. La valeur du MDT est calculée sur la base du diamètre moyen de tous les trous de la lésion. En effet, pour définir le diamètre moyen des mailles, nous avons évalué la distance Euclidienne entre les trous. Enfin, nous avons évalué la densité (3.6) du graphe du RP, en fonction de laquelle nous décidons de la présence ou pas d'un réseau de pigment. Si la densité du graphe est supérieure à un seuil, le réseau de pigment est considéré comme présent, sinon le réseau de pigment est absent.

$$Densite = \frac{NA}{NN.\log(\mathrm{T})}, \tag{3.6}$$

où, NA (*resp.* NN) est le nombre d'arêtes (*resp.* nœuds) dans le graphe et T est la surface de la lésion, qui est utilisée pour normaliser le rapport NA/NN.

Nous avons appliqué la méthode proposée sur des images dermoscopiques complexes [26], en raison des paramètres d'acquisition, tels que l'éclairage et l'amplification de la taille de la lésion, et de la présence d'une quantité déraisonnable d'occlusion par l'huile ou par les poils (Fig. 3.8.). La base utilisée est composée de 122 images ELM (EpiLuminescence Microscopy), où 40 images sont sans RP (absent) et 82 images sont avec RP (présent). Ces images ont été marquées, comme RP absent ou RP présent, par cinq experts pondérés par leurs expériences. Lorsque l'image contient un RP, nous reconnaissons son

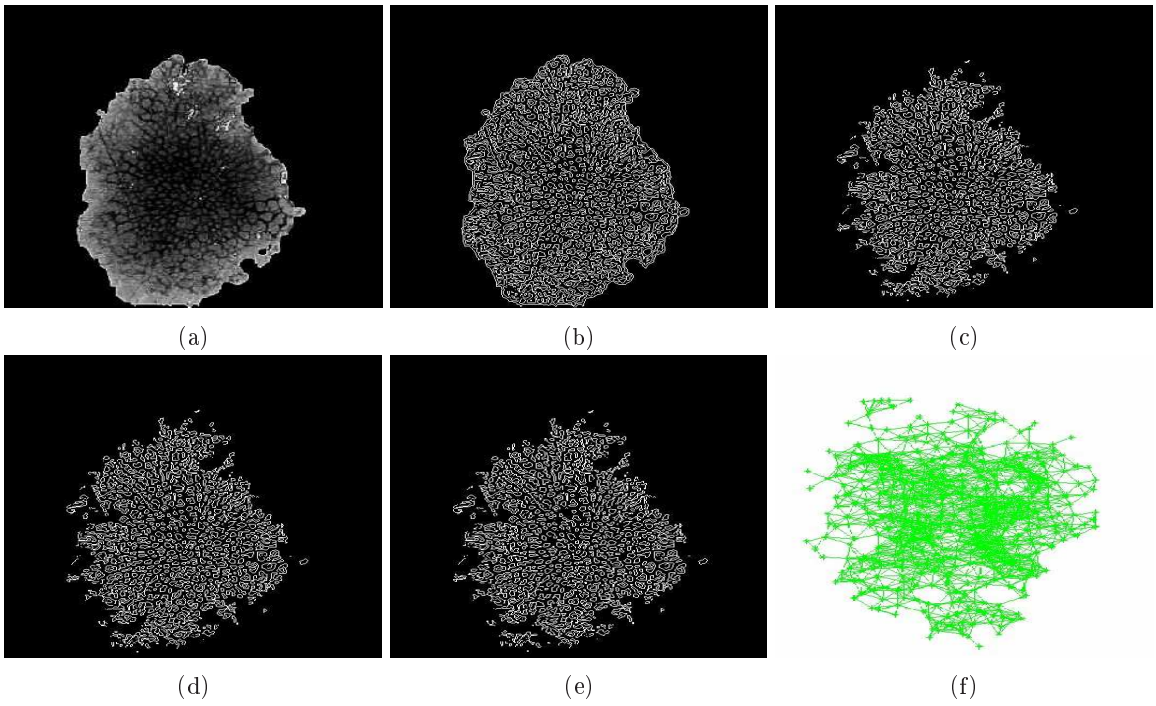


FIGURE 3.7 – Détection du réseau de pigment : (a) le canal vert de la lésion segmentée, (b) filtrage LoG, (c) filtrage basé sur la taille, (d) lissage des bulles d'huile et des kystes blancs, (e) filtrage basé sur l'intensité, (f) graphe représentant le réseau de pigment.

type (typique *vs.* atypique) en calculant l'uniformité spatiale de mailles (*SU*) [18]. Fig. 3.8 illustre trois exemples où le premier exemple illustre la présence du RPT, le second présente un RPA et le dernier illustre l'absence du RP. La méthode proposée détecte correctement la présence du RP et discrimine aussi entre RPT (densité= 0.59 et SU= 0.037) et RPA (densité= 0.29 et SU= 0.047), sachant que la densité est égal à 0.03 pour RP absent. La précision de notre détecteur du réseau de pigment est d'environ 85% pour 122 images pour les deux classes (présent *vs.* absent). La méthode proposée ne parvient pas à détecter le RP dans certaines images à cause du mauvais choix de certains paramètres, qui est basé sur une procédure d'apprentissage sur un ensemble aléatoirement choisi. Une fois le RP est détecté présent, nous pouvons reconnaître son type (typique *vs.* atypique) avec une précision de 100%. En outre, pour évaluer objectivement la performance de la méthode proposée, nous avons analysé la courbe ROC. Cette méthode a réalisé une aire sous la courbe ROC (AUC) de 0.821 pour distinguer correctement les images qui contiennent des réseaux de pigment de celles qui le sont pas (Fig. 3.9). Cette valeur confirme, selon les règles générales de Hosmer et Lemeshow d'interprétation des valeurs AUC [76], que notre méthode permet une discrimination optimale vis-à-vis de la présence du RP. En outre, en comparant nos taux de précision avec ceux enregistrés par [8] et [13], nous pouvons déduire que la méthode proposée les surpasse, bien que nous n'avons pas utilisé la même vérité-terrain (ceci est principalement du au manque de benchmarks d'images dermatoscopiques [1]). En particulier, comparativement à la méthode de [129], qui est la plus similaire à celle proposée, nous sommes en mesure de détecter plus correctement les trous du RP (Fig. 3.10), ce qui optimise par la suite le taux de classification. Par exemple, contrairement à [129], la méthode proposée ne lie pas la plupart des nœuds du RP de la lésion de Fig. 3.10, ce qui permet de l'associer correctement à sa classe diagnostique. Enfin, dans le but d'évaluer les performances de la construction du graphe dans le cas d'une mauvaise segmentation de la lésion, nous avons testé la méthode proposée pour les cas d'une lésion sur-segmentée et d'une lésion sous-segmentée (Fig . 3.11). Dans les deux cas, les filtrages (taille + intensité) permettent de ne garder que les trous qui appartiennent au RP. En plus, l'existence de trous manquants et/ou supplémentaires n'affecte pas la densité du graphe,

et par la suite la décision de présence du RP. En effet, dans les deux cas, les résultats produits pour la détection et la reconnaissance du PN étaient tous les deux correctes.

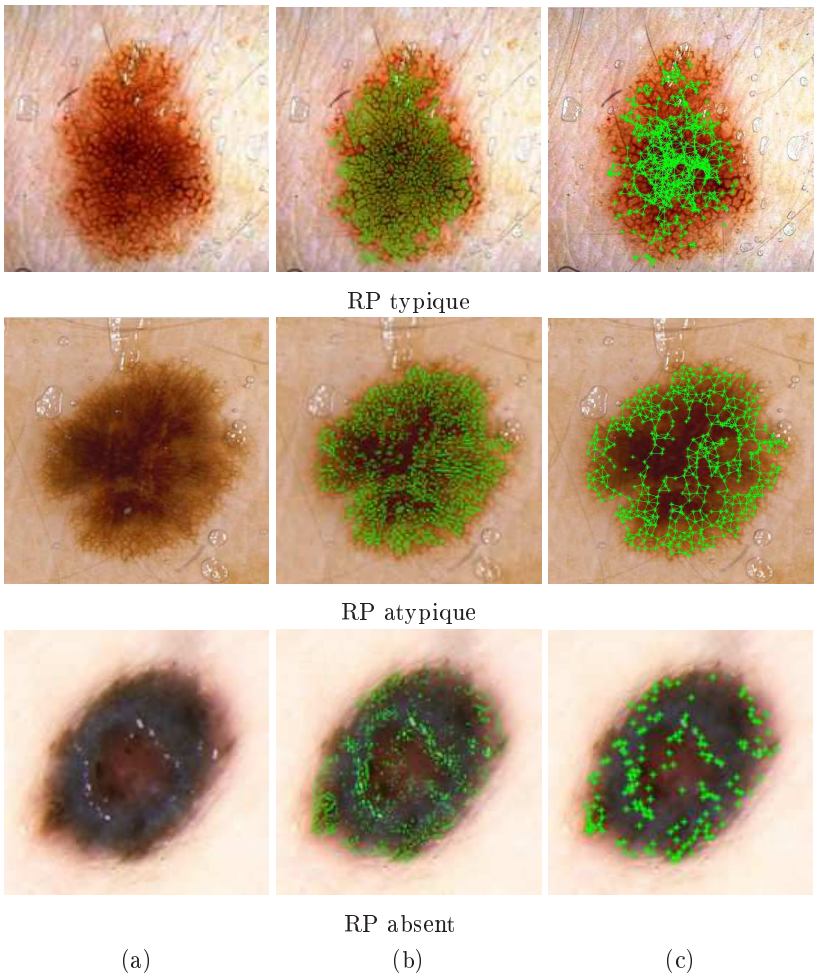


FIGURE 3.8 – Détection du RP : (a) image originale, (b) trous du RP, (c) graphe du RP.

3.3 Recalage non-rigide des images médicales

Étant donné une image source et une image cible, le recalage d'images est le processus qui permet de déformer l'image source afin de superposer les pixels représentant les mêmes structures dans les deux images [172] [144]. Dès lors qu'on veut comparer des images, le recalage s'impose souvent ce qui explique la diversité des applications médicales [110] [149] [118] (évolution temporelle des tumeurs, intégration d'informations complémentaires de différentes modalités ou projections, comparaison avec un atlas...). Les images recalées peuvent être de différentes dimensions, prises avec la même modalité (monomodale) ou avec des modalités différentes (multimodale), appartenant au même sujet (intra-sujet) ou à des sujets différents (inter-sujet). La mise en œuvre d'une technique de recalage revient principalement à répondre aux quatre questions suivantes : Quelles informations faut-il utiliser pour guider le recalage (*les attributs*) ? Comment quantifier le degré d'alignement entre deux images (*la mesure de similarité*) ? Comment déformer l'image source pour qu'elle s'ajuste à l'image cible (*le modèle de déformation*) ? Comment trouver la meilleure déformation qui maximise la mesure de similarité (*la stratégie d'optimisation*) ? Bien que la méthode la plus sûre pour trouver la transformation optimale

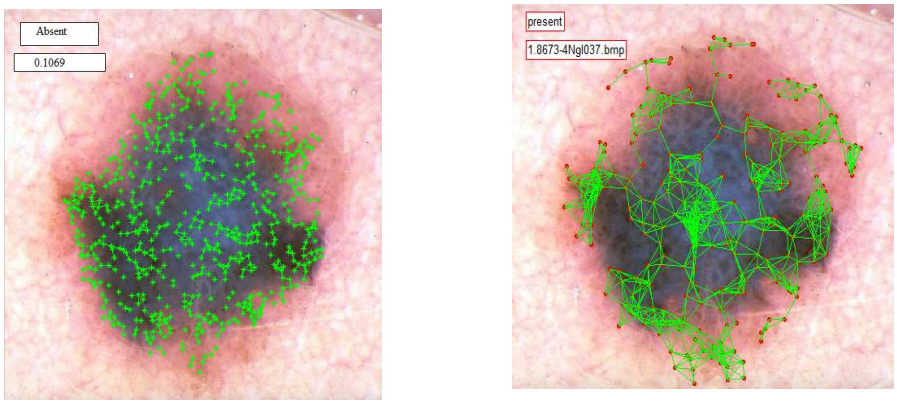


FIGURE 3.9 – Classification du RP par la méthode proposée (a) et par celle introduite dans (b) [129].

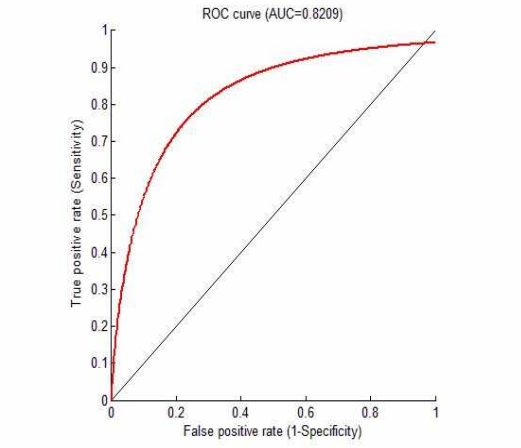
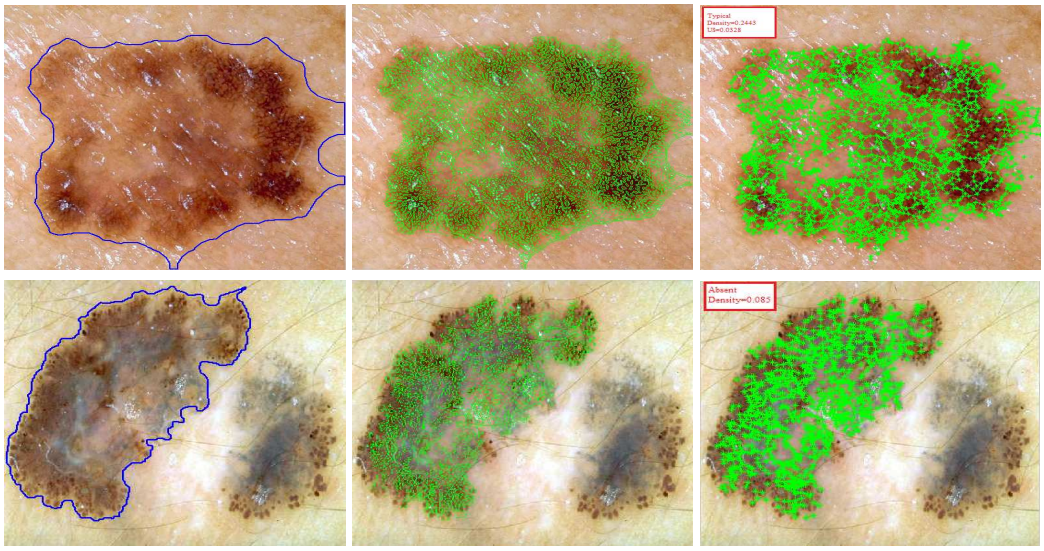


FIGURE 3.10 – La courbe ROC de la détection du réseau de pigment.

soit la recherche exhaustive, des algorithmes d'optimisation sont souvent utilisés afin de réduire le temps de calcul. Ainsi, le recalage revient à résoudre le problème d'optimisation suivant [31] :

$$\hat{T} = arg \max_{T \in \mathbf{T}} S(I, J, T), \tag{3.7}$$

où, S est la fonction qui évalue la similarité entre l'images source I et celle cible J et le recalage consiste alors à trouver la meilleure transformation géométrique $\hat{T}$, parmi l'ensemble $\mathbf{T}$ des transformations, qui maximise la mesure S en déformant la source I (3.5). Toutefois, malgré les avancés considérables réalisées dans le domaine de recalage d'images médicales [46] [163], les méthodes proposées n'arrivent souvent à convaincre les médecins de les utiliser dans leurs routines cliniques. Ceci est principalement dû au fait que les transformations obtenues ne reflètent pas les différences réelles entre les images recalées. En outre, les méthodes existantes souffrent d'un temps de calcul élevé et nécessitent souvent l'interaction avec l'utilisateur, alors que les cliniciens sollicitent des solutions rapides, automatiques et qui reflètent correctement les déformations non-rigides entre les images médicales. En effet, en se basant sur les attributs utilisés, les méthodes existantes sont souvent classées en deux catégories, à savoir les méthodes géométriques, s'appuyant sur l'appariement de primitives extraites des images, et celles iconiques utilisant directement les intensités des images. Les méthodes géométriques sont plus



(a) (b) (c)

FIGURE 3.11 – Performance de la détection du RP dans le cas d'une sur-segmentation (première ligne) et le cas d'une sous-segmentation (deuxième ligne) de la lésion : (a) extraction de la lésion, (b) trous du RP, (c) résultats de la détection et de la reconnaissance du RP (RPT pour la première ligne et RP absent pour la deuxième ligne).

rapides, mais elles sont moins précises et imposent une étape d'extraction de primitives. En plus, ces méthodes nécessitent la présence de primitives simples à détecter et dispersées sur toute l'image, ce qui n'est pas le cas pour les images médicales. L'utilisation de toute l'information portée par l'image permet aux méthodes iconiques d'une part d'échapper à l'étape de segmentation et les problèmes qui y sont liés et d'autre part de développer des solutions automatiques, plus précises, et plus adaptées au recalage multimodale. Toutefois, vu que tous les pixels sont considérés, les méthodes iconiques nécessitent une charge calculatoire élevée. En plus, la relation entre les intensités des deux images n'est pas forcément triviale, notamment dans le cas d'images multimodales. Ainsi, l'utilisation des méthodes hybrides, permettant de combiner les avantages des deux approches, s'avère cruciale afin de produire des solutions automatiques, précises et rapides. Par ailleurs, le choix de la mesure de similarité est souvent guidé par les modalités impliquées ainsi que par la contrainte du temps de calcul. L'information mutuelle permet de tenir compte des dépendances statiques entre les intensités des images recalées aux dépens d'un temps de calcul énorme et une fonction difficile à optimiser (grand nombre de maxima locaux). Par contre, l'utilisation du rapport de corrélation permet de diminuer le temps de calcul et de simplifier la fonction à optimiser. Cependant, ce rapport risque de ne pas satisfaire l'hypothèse de dépendance fonctionnelle entre les intensités des images, notamment pour le cas multimodal. Parmi les modèles de déformations, qui conditionnent fortement la manière avec laquelle l'image est géométriquement modifiée, les opérateurs non-linéaires sont plus adaptés, par rapport à ceux linéaires, au contexte général des images médicales. En effet, lorsque l'organe sujet de diagnostic subit des déformations non-rigides (*e.g.* compression des seins lors de l'acquisition, recalage inter-sujet...) les modèles non-linéaires permettent de mieux modéliser les différences inter-images. Nous avons classifié ces modèles selon la manière d'introduire des contraintes sur la transformation pour assurer sa régularité, et ceci pour mettre en exergue la relation entre le modèle de déformation et le temps du calcul [51]. Nous avons distingué les modèles décrits par des équations aux dérivées partielles, les approches compétitives, les approches itératives et les approches implicites. Les approches compétitives et celles décrites par des équations aux dérivées partielles sont les plus précises mais nécessitent une charge calculatoire très élevée. De là, les approches implicites et itératives, étant plus rapides, sont plus appropriées puisqu'elles dissocient le problème de

régularisation de la transformation de celui de la recherche de cette transformation. En particulier, les modèles itératifs, divisant le problème de recalage en deux étapes (calcul de la transformation suivie d'une étape de lissage) tout en optimisant le temps de calcul, s'avèrent les plus prometteurs, bien que la transformation finale obtenue n'a pas de sens physique. Enfin, dans l'objectif d'optimiser le temps de calcul ainsi que la qualité de la déformation, les stratégies hiérarchiques sont les mieux adaptées pour la résolution du problème d'optimisation, et l'association des stratégies hiérarchiques de données et de déformation s'avère désormais prometteuse. Ainsi, dans l'objectif de développer une méthode automatique de recalage non-rigide qui permet un bon compromis entre le temps de calcul et la précision, nous avons opté pour une approche hybride, itérative et hiérarchique. La méthode de subdivision hiérarchique de [101] satisfait ces différents critères. Elle consiste à diviser progressivement les images à recaler en des imagettes de tailles de plus en plus réduites. Les imagettes correspondantes dans deux images sont recalées indépendamment des autres imagettes en utilisant des transformations rigides. Les centres des imagettes recalées sont ensuite utilisées pour construire une transformation globale lisse. Néanmoins, bien que cette méthode de recalage est l'une des plus rapides [93], l'amélioration de son temps de calcul s'avérait nécessaire pour l'intégrer dans des applications cliniques. Dans ce qui suit, nous présentons cette méthode avec les principales améliorations proposées dans la littérature. Ensuite, nous détaillons notre contribution à cette méthode qui consiste essentiellement à l'utilisation conjointe des hiérarchies de données et de déformation via l'intégration de l'approche de subdivision progressive dans une pyramide Gaussienne. Ceci permet de diminuer considérablement la charge calculatoire tout en garantissant une bonne qualité de recalage.

3.3.1 Recalage non-rigide par subdivision hiérarchique

L'approche la plus simple pour le recalage non-rigide est de diviser les images à recaler en des imagettes qui sont ensuite alignées localement. Le recalage local des imagettes utilise des transformations rigides ou affines et il se fait indépendamment des autres imagettes, sans aucune contrainte de régularité de la transformation globale. Cette transformation globale est itérativement définie par une procédure de lisage. Little et al. [103] ont validé cette approche en recalage des images de la colonne vertébrale, tout en utilisant un segmenteur en régions pour la subdivision des images. Bien que cette méthode s'avère appropriée au cas où on s'intéresse seulement au voisinage des structures rigides, tel que le cas de la colonne vertébrale, elle ne permet pas de tenir compte des grandes déformations. De là, plusieurs travaux ont utilisé des stratégies hiérarchiques afin de tenir compte aussi bien des déformations globales que de celles locales [109]. Pour ce faire, les images sont progressivement subdivisées en des imagettes de tailles de plus en plus petites, qui sont ensuite recalées localement d'une façon indépendante et la transformation globale est estimée par interpolation des résultats des recalages rigides locaux. Dans ce cadre, Likar et Pernus [101] ont proposé une méthode hiérarchique pour le recalage non-rigide des images microscopiques des fibres musculaires. Cette méthode est basée sur l'utilisation de l'information mutuelle pour l'évaluation de la similarité et la technique des splines à plaques minces [29] pour l'interpolation des centres des imagettes recalées. Le processus de division progressive des images est répété jusqu'à ce que les imagettes atteignent une taille minimale prédéfinie (Fig. 3.12). Plusieurs modifications ont été proposées à la méthode de [101] afin d'améliorer sa précision et son temps de calcul. En l'occurrence, sachant que la précision du recalage globale dépend énormément de celles des recalages locaux, plusieurs travaux ont montré que l'utilisation du rapport de corrélation au lieu de l'information mutuelle lors du recalage local des imagettes permet d'améliorer le résultat du recalage local, et par suite la précision du recalage global [9]. D'autre part, pour optimiser le coût de recalage, d'autres techniques ont utilisé une pyramide d'images dans le niveau le plus grossier [11]. L'objectif est de procéder à une étape rapide de recalage grossier d'une façon rigide, suivie d'une étape de recalage non-rigide plus fine mais plus lente.

3.3.2 Combinaison des hiérarchies de données et de déformation

Afin de réduire le coût de calcul du recalage non-rigide, plusieurs travaux ont tenté de combiner les hiérarchies des données avec ceux de déformation. Par exemple, Hellier et al. [72] ont proposé un cadre multirésolution hiérarchique et multigrille pour le recalage non-rigide des IRMs cérébrales.

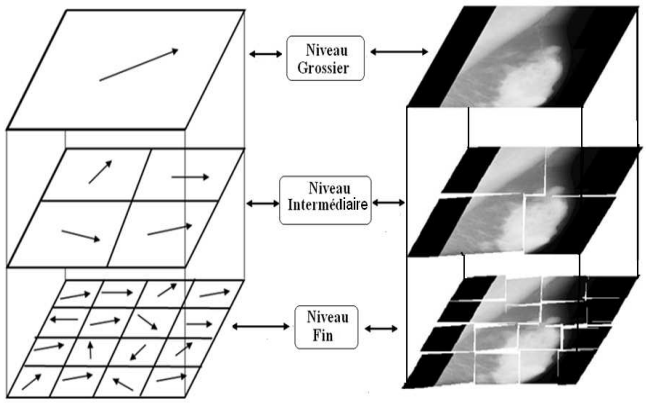


FIGURE 3.12 – Subdivision hiérarchique progressive : le nombre des imagettes à recaler augmente en allant du niveau grossier au niveau le plus fin.

A chaque niveau de résolution, une minimisation multigrille basée sur des partitions successives du volume initial est utilisée. Dans [11], les auteurs ont introduit un schéma automatique qui combine une hiérarchie de données et une hiérarchie de déformation, et non pas les deux. Toutefois, pour ces méthodes, soit la complexité des données, soit celle de la déformation augmente à chaque niveau de la hiérarchie. De là, nous avons proposé une approche hiérarchique pour le recalage non-rigide, dans laquelle la complexité de données ainsi que celle de déformation augmentent simultanément à chaque niveau. Afin de réduire la complexité de recalage, cette approche couple l'approche progressive de subdivision avec une pyramide Gaussienne. La subdivision hiérarchique veille à ce que le processus de recalage soit fiable aussi bien pour les petites déformations que pour les grandes, tandis que l'utilisation de la pyramide Gaussienne vise à optimiser le temps de calcul. L'idée est d'appliquer l'approche de division progressive sur une pyramide d'images au lieu de l'appliquer sur une image de taille fixe. Ainsi, les transformations les plus simples (affines) sont estimées à partir des images de petites tailles, alors que les transformations les plus complexes sont estimées à partir des images de grandes tailles. Pour ce faire, la méthode proposée commence par générer une pyramide Gaussienne pour chaque image (Fig. 3.13), tel que chaque niveau de la pyramide est obtenu par filtrage Gaussien et sous-échantillonnage du niveau précédent. Le processus de recalage commence par aligner les images dans le niveau le plus grossier dans la pyramide. La méthode proposée est générique et les choix de la mesure de similarité et de la technique d'optimisation dépendent dès lors du type des images à recaler (monomodale/multimodale, 2D/3D...). Dans le niveau suivant, les images sont subdivisées en quatre imagettes, qui seront recalées localement et indépendamment. Ce processus est répété jusqu'à atteindre le dernier niveau de la hiérarchie (Fig. 3.14). En associant les centres des imagettes recalées, les splines à plaques minces permettent enfin de construire la transformation globale lisse.

Etant donné que la qualité du recalage est un compromis entre le temps de calcul et la précision de la transformation obtenue, nous avons comparé la méthode proposée par rapport à celle de [101]. Comme la tâche la plus consommatrice dans le recalage est celle de l'évaluation de la similarité, nous avons focalisé notre étude de complexité sur l'estimation de la mesure de similarité, qui est dans notre cas le coefficient de corrélation CC (3.8).

$$CC = \frac{\sum_m \sum_n (I(m,n) - \overline{I})(J(m,n) - \overline{J})}{\sqrt{(\sum_m \sum_n (I(m,n) - \overline{I})^2)(\sum_m \sum_n (J(m,n) - \overline{J})^2)}}, \tag{3.8}$$

avec, I (*resp.* J) et $\overline{I}$ (*resp.* $\overline{J}$) représentent successivement l'image source (*resp.* l'image cible) et la moyenne des intensités de ses pixels. Le calcul du numérateur dans la formule de CC (3.8) nécessite T^2 multiplications, avec T×T est la taille des imagettes. En plus, afin de simplifier le calcul de la complexité, nous avons supposé que les transformations rigides locales sont composées seulement des translations

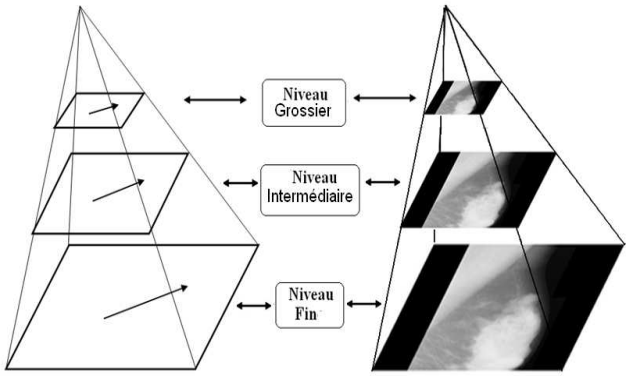


FIGURE 3.13 – Approche multirésolution utilisant une pyramide Gaussienne : la taille des images à recaler augmente en allant du niveau grossier au niveau le plus fin.

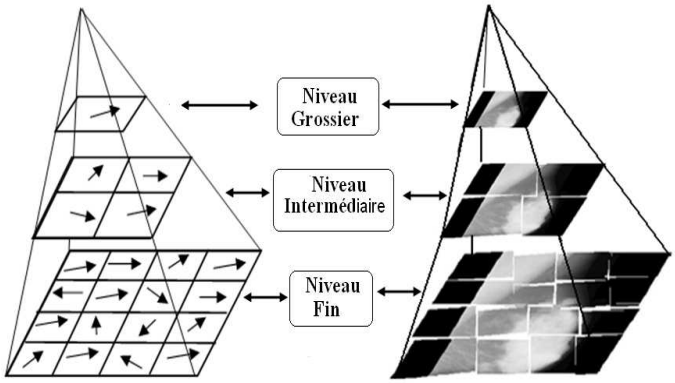


FIGURE 3.14 – Méthode proposée : la taille des images et le nombre des imagettes à recaler augmentent en même temps en allant du niveau grossier au niveau le plus fin.

selon les axes de x et de y, et les translations permises dans chaque direction ne doivent pas dépasser le quart de la taille de l'image. Si nous utilisons une recherche exhaustive pour trouver les paramètres des transformations locales, alors la complexité pour calculer CC lors du recalage local est $\frac{T^2}{4}$. Ainsi, étant donné deux images à recaler de tailles N×N et soit le nombre l des niveaux (la taille de la plus petite imagette est alors $\frac{N}{2^{l-1}}$), le nombre d'imagettes à recaler dans chaque niveau j $(1 \leq j \leq l)$ est 4^{j-1} aussi bien pour notre méthode que pour celle de [101]. Cependant, la taille de chaque imagette est $\frac{N}{2^{l-1}}$ (*resp.* $\frac{N}{2^{j-1}}$) dans le cas de notre (*resp.* l'autre) approche. Ainsi, Le coût du niveau j est $N^4.2^{(2j-4l)}$ (*resp.* $N^4.2^{(-2j)}$) dans notre (*resp.* l'autre) cas. Par conséquent, le coût du calcul total de l'approche proposés est déterminé par (3.9) et celui de l'approche de subdivision progressive est défini par (3.10). En outre, nous avons comparé le temps de calcul total pour les deux méthodes comparées en précisant l'évolution de ce temps d'un niveau de la hiérarchie au suivant (Fig. 3.19). Cette comparaison est réalisée sur des IRMs cérébrales de tailles 512×512, avec une valeur de l égale à 4. En comparant la complexité et le temps de calcul des deux méthodes [4], il est clair que la méthode proposée est beaucoup plus rapide que celle de [101].

4. Les deux méthodes ont été codées en Matlab et exécutées sur une machine dotée d'un processeur Intel® *Pentium*4 2.40 GHz et de 1 GO de RAM.

$$\sum_{j=1}^{l} N^4.2^{(2j-4l)} = \frac{4.N^4}{3}4^{-l}. \tag{3.9}$$

$$\sum_{j=1}^{l} N^4.2^{(-2j)} = \frac{4.N^4}{3}(4^l - 1). \tag{3.10}$$

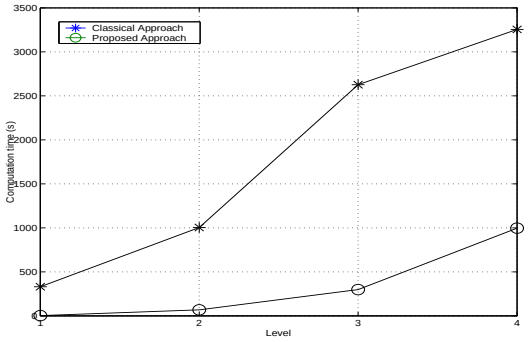


FIGURE 3.15 – Évolution du temps de calcul en fonction du niveau de la hiérarchie.

Afin d'évaluer les précisions des méthodes comparées, nous avons introduit des déformations aléatoires sur des mammographies des seins et des images IRM cérébrales (de taille 512×512). Les résultats obtenus ont été validés visuellement et quantitativement (Fig. 3.16). La validation visuelle est basée sur les images de différence, alors que celle quantitative est basée sur trois métriques hétérogènes, à savoir une mesure d'erreur (l'erreur quadratique moyenne MSE (3.11)), une mesure de la qualité d'images (le rapport PSNR) et une mesure de similarité (le coefficient de corrélation CC (3.8)), sachant que les petites (*resp.* grandes) valeurs de MSE (*resp.* de CC et PSNR) sont des signes d'une bonne qualité de recalage. Nous avons évalué ces métriques pour des mammographies (Tab. 3.1) et des IRMs cérébrales (Tab. 3.2) avant le recalage, après un recalage affine, après un recalage par [101] et après un recalage par la méthode proposée. Les résultats obtenus confirment que la qualité de recalage de la méthode proposée est assez comparable à celle de subdivision progressive de [101], et ceci aussi bien pour les mammographies que pour les IRMs. Ces résultats prouvent que le recalage non-rigide, aussi bien par notre approche que par celle de [101], produit des résultats meilleurs que le recalage affine, notamment lorsque les images recalées présentent d'importantes différences. Il est ainsi clair que notre méthode, qui est complètement automatique, permet de diminuer considérablement le temps de calcul, par rapport à [101], sans pourtant dégrader la qualité de recalage.

$$MSE = \frac{1}{N.M}\sum_{m=1}^{M}\sum_{n=1}^{N}[I(m,n) - J(m,n)]^2. \tag{3.11}$$

	CC	MSE	PSNR
Pré-recalage	0.79	2729	31.70
Recalage affine	0.951	661	45.87
Recalage par la méthode de [101]	0.953	636	46.27
Recalage par la méthode proposée	0.953	635	46.28

TABLE 3.1 – CC, MSE et PSNR obtenus suite au recalage des mammographies.

	CC	MSE	PSNR
Pré-recalage	0.64	3423	29.44
Recalage affine	0.89	1116	40.64
Recalage par la méthode de [101]	0.916	847	43.39
Recalage par la méthode proposée	0.911	897	42.82

TABLE 3.2 – CC, MSE et PSNR obtenus suite au recalage des IRMs cérébrales.

3.4 Test automatique de photo-consistance pour la coloration des voxels

La reconstruction volumétrique des organes à partir d'un ensemble d'images 2D est un problème ouvert dans le domaine de l'imagerie médicale [16]. Plusieurs méthodes ont été proposées pour résoudre ce problème et elles peuvent être regroupées en deux classes principales : les méthodes passives et celles actives. De nos jours, les méthodes passives sont beaucoup plus utilisées dans le contexte de l'imagerie médicale. Dans ce cadre, l'utilisation de plusieurs caméras optimise la qualité de la reconstruction, puisqu'elles permettent d'améliorer, comparativement aux méthodes monoculaires, la qualité des cartes de profondeur en utilisant plusieurs images sources, ce qui produit des modèles complets des objets 3D. En particulier, l'existence de plusieurs caméras permet une gestion optimale de l'occlusion, puisque les régions qui sont occultées dans une image peuvent être visibles dans d'autres images [85]. En l'occurrence, les méthodes multioculaires basées sur la stéréovision produisent des reconstructions précises de haute résolution, mais elles nécessitent des algorithmes complexes et ainsi un coût de calcul élevé. Pour les méthodes "Shape from Silhouette", la forme 3D d'un objet est principalement récupérée à partir de son contour. Ces méthodes sont très robustes contre les changements d'éclairage et les variations photométriques entre les caméras. Cependant, elles ignorent totalement l'information couleur, ce qui provoque souvent des reconstructions incomplètes puisqu'elles ne décrivent pas suffisamment les concavités des surfaces. Dans notre cas, nous nous sommes intéressés à la méthode de coloration des voxels [134], qui est une amélioration de l'approche de balayage de l'espace, offrant une solution aux problèmes de visibilité et d'occlusion générés par les parallaxes entre les images. Le principe de base de cette technique est de profiter de la quasi-totalité des informations contenues dans les images d'entrée [104]. Elle classe chaque élément 3D dans un volume discrétisé de voxels opaques qui encapsule la scène à reconstruire, afin de décider si cet élément appartient à la surface de l'objet $3D$ ou non. Cela renvient à tester la photo-consistance de chaque voxel, qui est la plus ancienne propriété photométrique d'une scène dans la littérature de la stéréovision. Ce test suppose que si un voxel appartient au volume 3D, alors les pixels correspondants dans les projections 2D doivent avoir presque la même couleur. Les voxels inconsistants sont éliminés itérativement jusqu'à l'arrêt sur la surface de l'objet 3D. En effet, un point 3D doit apparaître avec des couleurs similaires, sur ses différentes projections, lorsqu'il n'est pas occlus. Cette propriété est basée sur un couple d'hypothèses. D'une part, les objets de la scène doivent avoir des surfaces Lambertiennes, afin de simplifier la corrélation des pixels sur différents points de vue. Cette loi affirme que la quantité de lumière, émise par une surface dans des directions différentes, est proportionnelle au cosinus de l'angle entre la direction et la normale de la surface. Ainsi, la quantité de lumière émise par une surface Lambertienne est indépendante de la position de l'observateur. Bien que cette hypothèse fournisse un modèle raisonnable pour des surfaces mates qui dispersent la lumière approximativement d'une manière uniforme dans toutes les directions, elle n'est pas bien adaptée pour la reconstruction des surfaces spéculaires [134]. D'autre part, la deuxième hypothèse, dite de visibilité, suppose que la projection d'un point 3D sur les images correspondantes peut être calculée efficacement. Etant donné qu'elle est capable de produire une reconstruction 3D photoréaliste à partir des images 2D sans passer par une étape de mise en correspondance, la méthode de coloration des voxels est devenue une technique populaire pour une reconstruction de faible coût d'un modèle 3D à partir d'une série d'images calibrées. En effet, vu que chaque voxel n'est visité qu'une seule fois, la complexité spatiale et celle temporelle sont toutes les deux linéaires [116]. En outre, cette méthode n'a pas besoin d'un

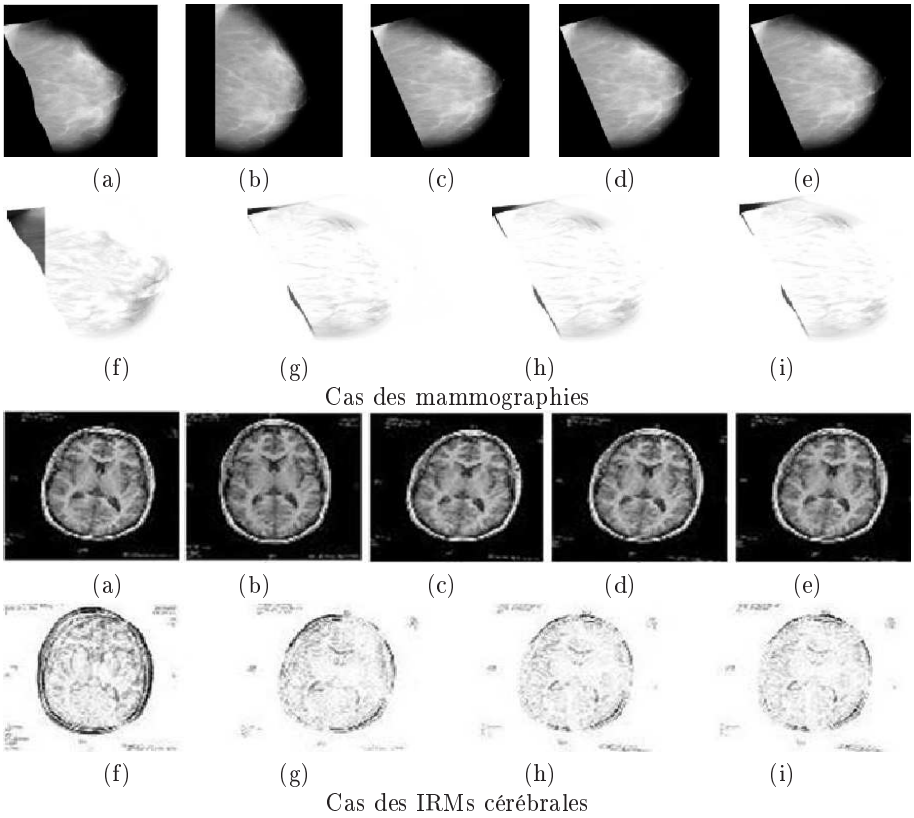


FIGURE 3.16 – Résultats du recalage : a) image cible, b) image source, c) recalage affine, d) recalage par la méthode de [101], e) recalage par la méthode proposée. Image de différence après : f) pré-recalage, g) post-recalage affine , h) post-recalage par [101], i) post-recalage par la méthode proposée.

grand nombre d'images pour produire une reconstruction de haute qualité. Toutefois, la qualité de reconstruction est fortement dépendante de l'estimation de la photo-consistance de chaque voxel. Cette estimation est basée sur la projection de chaque voxel afin de récupérer les valeurs des couleurs à partir du patch surfacique où le voxel est projeté. Sous l'hypothèse de la réflectance Lambertienne, un voxel ayant presque la même couleur dans toutes les images dans desquelles il est visible, doit être coloré avec cette couleur. Cette décision est basée sur une étape de seuillage. Cependant, en plus de l'absence de toute information sur le voisinage, il est extrêmement difficile de définir un seuil unique approprié. D'ailleurs, même si le seuil est défini de manière optimale, l'utilisation de la même valeur pour tous les voxels de la surface génère un problème de compromis entre la précision et la stabilité [27]. Pour cette raison, nous avons proposé d'intégrer l'information spatiale en utilisant un seuillage par hystérésis qui considère la connexité des voxels colorés. Cela permet notamment de mieux gérer les "voxels flottants" qui peuvent se projeter sur des pixels photo-consistants par hasard. En outre, l'ambiguïté du choix des seuils est extrêmement minimisée par l'utilisation d'un degré d'appartenance flou de chaque voxel à la classe des voxels photo-consistants. Aussi, la méthode proposée n'a pas besoin de seuils prédéterminés puisque ceux d'hystérésis sont définis automatiquement et de manière adaptative en fonction du nombre des images sur lesquelles le voxel est projeté. En effet, après la division de l'espace 3D de la scène en une grille de voxels, cette dernière est parcourue dans un ordre adéquat, sous la contrainte de la visibilité ordinale, pour vérifier la photo-consistance de chaque voxel. Etant donné que les caméras sont placées sur le même côté d'un plan et que les voxels les plus proches sont visités en premier, l'ordre "near-to-far" garantit qu'un voxel ne peut pas être occlus par un voxel non visité, ce qui optimise le traitement des occlusions [96]. La reconstruction résultante ("photo-hull") [95] est définie comme étant le volume maxi-

mum qui est photo-consistant avec l'ensemble des images d'entrée. Pour décider s'il est photo-consistant ou pas, chaque voxel est projeté sur les images d'entrée. Ensuite, tous les pixels correspondants à ces projections, et qui n'ont pas encore été marqués, sont collectés dans un seul ensemble. Si cet ensemble est vide, alors le voxel est supprimé. Dans le cas contraire, après le tri de toutes les intensités des pixels afin d'éliminer les valeurs aberrantes, la méthode proposée est basée sur une métrique de photo-consistance et deux seuils dynamiques pour décider si le voxel est à l'intérieur ou à l'extérieur de l'objet 3D. Si le voxel est photo-consistant, alors il est marqué comme parcouru, afin de gérer avec précision les effets d'occlusion, et il est coloré avec la couleur moyenne des pixels correspondants.

3.4.1 Projection des voxels

Une fois que le volume d'entrée est discrétisé en petits cubes (correspondants aux voxels), la première étape de la méthode proposée vise à projeter chaque voxel V^i sur des patches colorés dans toutes les images à partir desquelles il est visible. Pour ce faire, nous avons utilisé le modèle sténopé idéal d'une caméra, qui est une très bonne approximation de la plupart des caméras du monde réel [14]. En effet, puisque les caméras utilisées sont calibrées, une projection de "sprite" est utilisée pour projeter un point de l'espace 3D sur le plan 2D d'une image [14]. Plus précisément, en tenant compte des huit sommets $\{P_1, \ldots, P_8\}$ du cube correspondant à un voxel V^i et de l'ensemble des images calibrées de l'entrée $\{I_1, \ldots, I_M\}$, il est possible de relier chaque sommet $P_k = (X_k, Y_k, Z_k)^t$, $1 \leq k \leq 8$, à sa projection $p_k^j = (x_k^j, y_k^j)$ sur le plan de l'image I_j, en utilisant les coordonnées homogènes (3.12).

$$(x_k^j,\ y_k^j,\ 1)^t = K_j.R_j.T_j.(X_k,\ Y_k,\ Z_k,\ 1)^t, \tag{3.12}$$

où, R_j et T_j désignent les paramètres extrinsèques de la caméra C_j, qui décrivent sa position et son orientation dans le monde 3D, et K_j désigne la matrice des paramètres intrinsèques de la caméra, qui représente la projection du sommet P_k le long du rayon qui le relie au centre de la caméra (3.13). En effet, T_j est la matrice de translation, qui illustre le vecteur de translation de la caméra $t_j = (t_j^x, t_j^y, t_j^z)$ décrivant sa position en coordonnées du monde réel, et R_j englobe la matrice de rotation r_j $(= [r_j^{11}\ r_j^{12}\ r_j^{13};\ r_j^{21}\ r_j^{22}\ r_j^{23};\ r_j^{31}\ r_j^{32}\ r_j^{33}])$ qui spécifie l'orientation de la caméra C_j. Les paramètres intrinsèques, (σ_j^x, σ_j^y) et f_j (3.13), sont respectivement les coordonnées du point principal de la caméra C_j et sa distance focale. Notons que dans notre cas, nous avons simplement supposé que $\eta = 1$ et $\tau = 0$, tout en négligeant les effets de la distorsion radiale de la caméra, comme la plupart des cas étudiés des caméras CCD typiques [53]. En effet, si la caméra utilise des pixels non carrés, alors la matrice K_j doit être étendue par un paramètre η, désignant le facteur d'échelle vertical, pour rendre la caméra plus adaptée à l'utilisation dans le monde réel. En outre, dans le cas d'une grille d'échantillonnage asymétrique, qui se produit lorsque la caméra utilisée est non perpendiculaire à l'axe optique, un paramètre η supplémentaire doit être également intégré dans la matrice K_j.

$$T_j = \begin{bmatrix} 1 & 1 & 1 & t_j^x \\ 1 & 1 & 1 & t_j^y \\ 1 & 1 & 1 & t_j^z \\ 0 & 0 & 0 & 1 \end{bmatrix},\quad R_j = \begin{bmatrix} r_j^{11} & r_j^{12} & r_j^{13} & 0 \\ r_j^{21} & r_j^{22} & r_j^{23} & 0 \\ r_j^{31} & r_j^{32} & r_j^{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad et \quad K_j = \begin{bmatrix} f_j & \tau & \sigma_j^x & 0 \\ 0 & \eta f_j & \sigma_j^y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{3.13}$$

Après la projection des huit sommets $\{P_1, \ldots, P_8\}$ du voxel V^i sur le plan de l'image I_j, la boîte englobante Ψ_j^i autour de leurs projections $\{p_j^1, \ldots, p_j^8\}$ est utilisée pour estimer la forme réelle du voxel. Cependant, le test de photo-consistance du voxel V^i est effectué seulement sur l'ensemble des images dans lesquelles ce voxel est visible. En effet, un voxel V^i est défini comme étant visible dans l'image I_j, seulement si la ligne sortante du voxel vers l'ensemble Ψ_j^i, des pixels correspondants dans cette image,

n'est pas obstruée. Ce problème d'occlusion a été résolu par l'ordre dans lequel les voxels sont parcourus, tout en limitant l'emplacement possible des caméras (la contrainte de visibilité ordinale). En effet, en balayant la grille, les voxels les plus proches de l'ensemble des caméras sont visités en premier, tout en marquant les pixels correspondants comme "déjà affectés". Cet ordre assure que les voxels qui peuvent être occlus sont testés pour la photo-consistance après les voxels qui peuvent les occulter. Ainsi, parmi les pixels appartenant à la boîte englobante Ψ_j^i dans I_j, seul l'ensemble $\prod_j^i$, des pixels qui ne sont pas déjà affectés lors du traitement des voxels précédents V_l $(1 \leq l < i)$, est considéré (3.14).

$$\prod_j^i = \Psi_j^i \cap (I_j / \bigcup_{l=1}^{i-1} \prod_j^l). \tag{3.14}$$

Ainsi, les couleurs des pixels retenus dans les images d'entrée sont collectées dans le même ensemble $\prod^i$ afin d'évaluer la photo-consistance du voxel V^i ($\prod^i = \prod_1^i \cup \prod_2^i \dots \prod_M^i$). En outre, le nombre N^i des images, dans lesquelles V^i est visible (*i.e.* les images I_j pour lesquelles l'ensemble $\prod_j^i$ n'est pas vide), est défini comme suit (3.15) :

$$N^i = \left| \left\{ I_j \in \{I_1, \dots, I_M\} / \Psi_j^i \not\subset \bigcup_{l=1}^{i-1} \Psi_j^l \right\} \right|, \tag{3.15}$$

où, $|\ |$ est l'opérateur de cardinalité d'un ensemble. En d'autres termes, un voxel V^i n'est considéré comme visible dans une image I_j que s'il existe au moins un pixel dans la boite englobante Ψ_j^i qui est à l'extérieur de la projection des voxels déjà parcourus V_l $(1 \leq l < i)$.

3.4.2 Evaluation de la photo-consistance

Après avoir défini les patches surfaciques, $\prod_1^i \dots \prod_M^i$, résultants de la projection du voxel courant V^i sur les N^i images où il est visible, l'étape suivante de la technique proposée pour la coloration des voxels consiste à évaluer la photo-consistance du voxel V^i. Ceci est réalisé à travers l'estimation de l'homogénéité de l'ensemble des couleurs $\prod^i$ englobant tous les patchs relatifs à V^i. La méthode standard de coloration des voxels estime cette homogénéité en fonction de la variance des couleurs de tous les pixels de $\prod^i$. Cette évaluation de photo-consistance basée sur la variance, qui est efficace en termes de coût de calcul, est très sensible aux surfaces non Lambertiennes et faiblement texturées ainsi qu'aux variations d'éclairage [75]. Dans notre cas, nous avons introduit une technique floue permettant la définition d'un degré d'appartenance $\mu_{CV}(V^i)$ $(\in [0,1])$ pour chaque voxel V^i à la classe des voxels photo-consistants (avec, $\mu_{\neg CV}(V^i) = 1 - \mu_{CV}(V^i)$). L'idée principale derrière l'évaluation floue de la photo-consistance d'un voxel est de minimiser la dépendance des résultats de reconstruction 3D aux seuils utilisés. En effet, ayant l'ensemble $\prod^i$ des couleurs des pixels inclus dans les patchs correspondants à un voxel V^i, le degré d'appartenance $\mu_{CV}(V^i)$ de V^i à la classe des voxels photo-consistants (3.16) est défini en deux étapes. En un premier lieu, l'ensemble non vide $\prod^i$ est trié selon l'intensité de couleur afin d'exclure les valeurs aberrantes. Pour ce faire, ayant la version triée ζ^i de l'ensemble $\prod^i$, seuls les pixels dont les couleurs sont dans l'intervalle $[d_1^i, d_9^i]$, où d_1^i et d_9^i sont respectivement le premier et le neuvième décile, sont pris en compte pour l'évaluation de la photo-consistance du voxel V^i (3.16). L'objectif du rejet des bords de l'intervalle est de mieux gérer les erreurs de quantification, la variation d'éclairage et le bruit des capteurs. En outre, cela permet de tenir compte du fait qu'un pixel d'une image peut contenir la couleur mélangée de plusieurs voxels [104]. En effet, ces effets indésirables se reflètent généralement par des couleurs qui sont très différentes de toutes les autres couleurs dans l'ensemble $\prod^i$, qui contient les patchs relatifs au même voxel V^i. Ensuite, le degré d'appartenance $\mu_{CV}(V^i)$ du voxel V^i à la classe des voxels photo-consistants est défini comme le rapport entre la valeur la plus faible et la plus élevée parmi toutes les valeurs dans l'ensemble des couleurs retenues (après le rejet des valeurs extrêmes). Ce degré représente les voxels qui sont fortement photo-consistants avec un degré d'appartenance élevé (*i.e.* $\mu_{CV}(V^i) \simeq 1$). Cependant, les voxels qui enregistrent des degrés d'appartenance faibles (*i.e.* $\mu_{CV}(V^i) \simeq 0$) sont considérés comme non photo-consistants et doivent être ensuite retirés de la reconstruction résultante. Le test flou de la photo-consistance offre une meilleure

modélisation de l'incertitude et de l'ambiguïté, puisqu'il permet de reporter la décision du coloration ou de la suppression d'un voxel douteux (*i.e.* $\mu_{CV}(V^i) \simeq 0.5$) jusqu'à ce que plus d'informations soient disponibles pour prendre la décision finale. Dans notre cas, ces informations sont principalement déduites à partir des décisions prises pour les voxels connexes parmi les voxels précédents V^l $(1 \leq l < i)$. De cette façon, nous évitons le fait que certains voxels peuvent être éliminés par erreur, dans un effet de cascade, surtout que ces voxels supprimés ne peuvent plus être récupérés.

$$\mu_{CV}(V^i) = \frac{min\{couleur(x,y)/(x,y) \in \prod^i \ et\ couleur(x,y) \in [d_1^i, d_9^i]\}}{max\{couleur(x,y)/(x,y) \in \prod^i \ et\ couleur(x,y) \in [d_1^i, d_9^i]\}}. \tag{3.16}$$

3.4.3 Estimation des seuils pour le seuillage par hystérésis

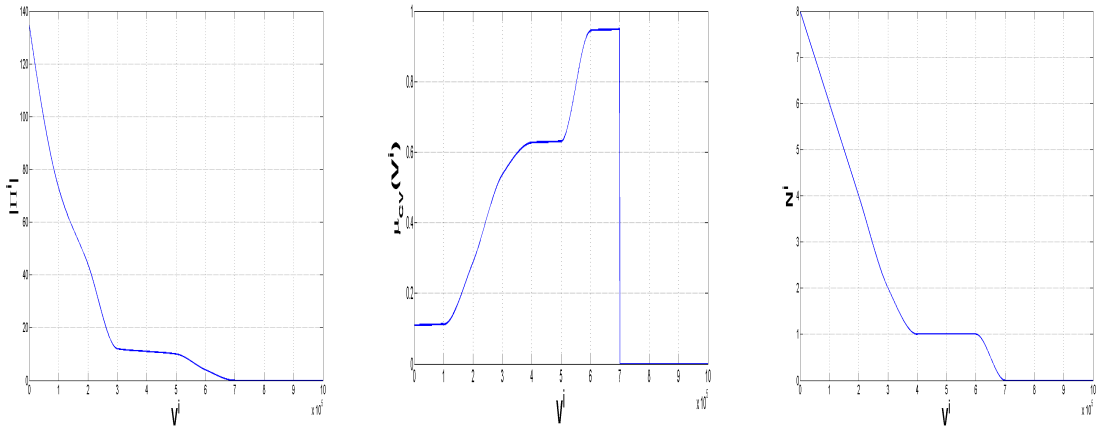
Ayant le degré d'appartenance $\mu_{CV}(V^i)$ du voxel courant V^i à la classe des voxels photo-consistants et le nombre N^i des images d'entrée dans lesquelles V^i est non occlus, la dernière étape de la technique de coloration des voxels proposée vise à déterminer si le voxel V^i est photo-consistant ou pas. Cela revient à déterminer le type du voxel V^i parmi deux classes ($type(Vi) \in \{Consistent, \neg Consistent\}$). Pour ce faire, nous avons introduit un seuillage par hystérésis (3.17), dans lequel deux seuils adaptatifs sont définis dynamiquement. Ce seuillage par hystérésis vise à intégrer la cohérence spatiale dans la reconstruction 3D qui en résulte tout en minimisant les effets du bruit. Le seuil haut T_h^i est utilisé pour sélectionner la plupart des voxels photo-consistants, de sorte qu'un voxel V^i est reconnu comme photo-consistant si $\mu_{CV}(V^i) \geq T_h^i$. Dans ce cas, le voxel V^i doit être coloré avec la valeur moyenne des couleurs retenues dans $[d_1^i, d_9^i]$. Dans le cas contraire, si $\mu_{CV}(V^i) \leq T_l^i$, le voxel V^i est identifié comme non photo-consistant et doit être supprimé. Tous les autres voxels V^i, pour lesquels $\mu_{CV}(V^i) \in]T_l^i, T_h^i[$, sont considérés comme des voxels candidats et doivent être ensuite traités de façon récursive pour décider de leur photo-consistance. En effet, si un voxel candidat V^i est relié à un voxel photo-consistant déjà visité, alors ce voxel candidat doit être coloré ; sinon, le voxel est supprimé (3.17). L'application de l'hypothèse de la connexité des voxels photo-consistants nous permet de vérifier le voisinage d'un voxel donné et d'éliminer les voxels qui reflètant des effets de bruit et qui ne sont pas photo-consistants.

$$\begin{aligned} type(V^i) = Consistent \ \Leftrightarrow \ & \left[\mu_{CV}(V^i) \geq T_h^i\right] \ OU \\ & \left[\left(\mu_{CV}(V^i) \in]T_l^i, T_h^i[\right) \ ET \ \left(\exists V^l \in \Delta / type(V^l) = Consistent\right)\right], \end{aligned} \tag{3.17}$$

où, Δ est un composante homogène qui inclut le voxel V^i et d'autres voxels connexes déjà reconnus comme photo-consistants. En plus de la considération explicite de l'information de connexité, la principale contribution du seuillage par hystérésis proposé réside dans le fait que les deux seuils utilisés, T_l^i et T_h^i $(0 < T_l^i < T_h^i < 1)$, sont automatiquement définis d'une manière adaptative en fonction du nombre N^i des images d'entrée dans lesquelles le voxel V^i est visible. Cela permet aux seuils d'hystérésis d'être liés à la visibilité du voxel traité V^i, qui peut être à son tour dérivée à partir des statistiques des projections de V^i. En effet, puisque les voxels visités en premier sont moins occlus que les suivants, ils sont projetés sur plus de pixels (Fig. 3.17.a) que les autres (*i.e.* $\forall i,\ |\prod^i| \geq |\prod^{i+1}|$), et par conséquent leurs hétérogénéités (Fig. 3.17.b) sont plus strictes (*i.e.* $\forall i,\ \mu_{CV}(V^i) \leq \mu_{CV}(V^{i+1})$). En outre, le nombre N^i des images dans lesquelles le voxel V^i est visible change de manière opposée (Fig. 3.17.c) à l'ordre dans lequel l'espace de voxels est parcouru sous la contrainte de la visibilité ordinale (*i.e.* $\forall i,\ N^i \leq N^{i+1}$). Ainsi, nous avons supposé que si le nombre N^i, des images dans lesquelles le voxel d'entrée V^i n'est pas occlus, est élevé (*resp.* faible), alors le seuillage devrait être plus strict, *i.e.* $T_l^i \nearrow$ et $T_h^i \searrow$ (*resp.* plus souple, *i.e.* $T_l^i \searrow$ et $T_h^i \nearrow$). En d'autres termes, si le nombre N^i des observations sur la photo-consistance d'un voxel V^i est relativement élevé, alors seulement les voxels très photo-consistants $(\mu_{CV}(V^i) \geq T_h^i)$ sont automatiquement conservés tout en permettant au maximum de voxels voisins, qui sont moins photo-consistants $(T_l^i < \mu_{CV}(V^i) < T_h^i)$, d'être considérés comme des candidats étant

donné la cohérence spatiale. Pour cela, les deux seuils d'hystérésis, T_l^i et T_h^i, sont modélisés à travers deux suites adjacentes qui dépendent de N^i. Lorsque N^i augmente alors T_l^i croit et T_h^i décroit, ce qui garantit un seuillage strict. En effet, deux suites sont dites adjacentes si l'une est croissante, l'autre est décroissante et si la différence des deux converge vers 0 lorsque N^i tend vers l'infini. Cela signifie qu'elles sont convergentes et elles convergent vers la même limite. Dans notre cas, nous avons supposé que $T_l^i = U_N^i = S(2N^i)$ et $T_h^i = V_N^i = S(2N^i + 1)$, sachant que $S(n)$ est la suite harmonique alternée (3.18), qui est le cas particulier $\eta(1)$ de la fonction de Dirichlet $\eta(s)$. Ainsi, les deux seuils sont modélisés avec deux suites adjacentes de telle sorte que leur valeurs sont comprises dans l'intervalle $]0, 1[$ et elles convergent vers la même limite l (*i.e.* $\forall N^i,\ 0 < U_N^i \leq U_{N+1}^i \leq l \leq V_{N+1}^i \leq V_N^i$). Nous avons opté pour ces deux suites, vu qu'elles convergent toutes les deux vers log(2) lorsque N^i tend vers l'infini. En effet, nos tests ont montré que les meilleurs résultats de reconstruction en 3D, en utilisant le test proposé pour l'évaluation floue de photo-consistance (3.16) lors du seuillage de la méthode standard de coloration des voxels avec un seul seuil T (*i.e.* $type(Vi) = consistent \Leftrightarrow \mu_{CV}(V^i) \geq T$), ont été obtenus avec des valeurs de seuils proches de log(2) (Fig. 3.18).

$$S(n) = \frac{1}{c} \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k}. \tag{3.18}$$



(a) (b) (c)

FIGURE 3.17 – Variation de $|\prod^i|$ (a), $\mu_{CV}(V^i)$ (b) et N^i (c) selon l'ordre dans lequel l'espace des voxels (discrétisé en 10^6 voxels) est parcouru, pour la séquence "Temple" composée de 8 images d'entrée de taille 640×480.

En utilisant ce test de photo-consistance, la technique proposée n'a plus besoin de seuils constants en entrée et elle est capable de produire des reconstructions plus précises que les reconstructions obtenues en utilisant les tests standards de photo-consistance. En effet, les valeurs des seuils sont définies d'une façon appropriée en fonction de l'emplacement du voxel. En particulier, si le voxel est sur les contours ou sur une surface texturée, alors il a un faible degré d'homogénéité dans chaque image, ce qui explique l'intérêt d'un seuil plus élevé. Cela permet d'intégrer la cohérence spatiale lors de la reconstruction du volume tout en évitant les "voxels flottants" et les trous le long de la procédure de coloration des voxels. En effet, l'efficacité de la méthode proposée a été démontrée sur des scènes artificielles. Comme objets synthétiques, nous avons utilisé les ensembles de données multi-vues "Temple" et "Dino" de Middlebury [133]. L'objet "Temple" est une reproduction en plâtre de 159.6 mm de hauteur d'un temple et l'objet "Dino" est un modèle de dinosaure en plâtre de 87.1 mm de hauteur. Ces deux ensembles de données, qui sont largement utilisés dans d'autres travaux, contiennent beaucoup de structures et de textures spatiales, pareillement aux données médicales. Ils sont capturés par une caméra CCD, avec une résolution

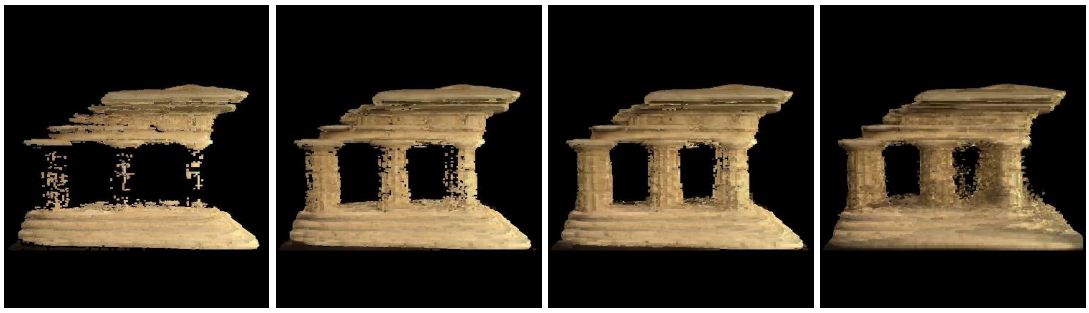
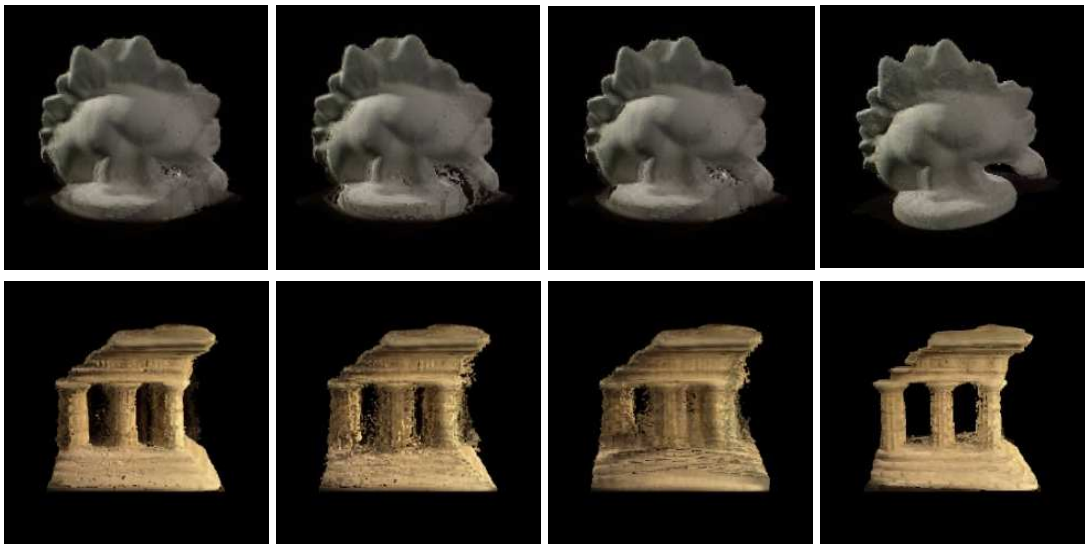


FIGURE 3.18 – Les résultats de reconstruction avec l'utilisation du test de photo-consistance flou avec un seul seuil T, qui est égale à : (a) 0.6, (b) 0.4, (c) 0.3 et (d) 0.2.

de 640×480 pixels, montée sur un support mobile calibré. Les objets filmés ont été éclairés par de multiples sources lumineuses, et les images avec des ombres ont été éliminées de l'ensemble de données. En effet, parmi l'ensemble complet des données "Temple" (*resp.* "Dino"), composé de 312 images (*resp.* 363 images), nous avons utilisé l'échantillon "TempleSparseRing" (*resp.* "DinoSparseRing") qui est une version éparse de l'ensemble de données complet avec 16 vues sur un anneau autour de l'objet. En outre, afin de respecter la contrainte de visibilité ordinale des méthodes de coloration des voxels, nous nous sommes limités aux huit premières images, qui correspondent à des caméras placées sur le même côté d'un plan, à la place des seize caméras de l'ensemble de données éparses. La comparaison des résultats enregistrés par le test de photo-consistance proposé avec ceux produits par d'autres méthodes de l'état de l'art de la coloration des voxels (test de photo-consistance basé sur l'écart-type [28], test de photo-consistance adaptatif [141] et test de photo-consistance basé sur l'histogramme [39]), ont prouvé que notre méthode surpasse les méthodes comparées pour les ensembles de données utilisés (Fig. 3.19). En effet, la reconstruction 3D résultante en utilisant la solution proposée est nettement plus précise et plus lisse, bien que notre méthode est entièrement automatique contrairement aux méthodes comparées (qui exigent une fixation empirique des seuils). Ceci est principalement dû au seuillage flou adaptatif proposé, qui impose le lissage et la cohérence spatiale lors de la reconstruction du volume 3D. En particulier, les "voxels flottants" et les trous sont nettement minimisés par la méthode proposée, notamment pour l'ensemble de données "Dino", malgré l'absence de la texture. En effet, la seule texture dans cet ensemble de données est due à des variations subtiles de l'ombrage sur la surface.

En outre, afin de comparer la méthode proposée avec des méthodes récentes de reconstruction 3D, qui ne font pas partie des méthodes de coloration des voxels, nous avons résumé dans Table. 3.3 les mesures de la précision, de la complétude et du temps CPU consommé. Les résultats complets peuvent être consultés sur la page d'évaluation de Middlebury (*http ://vision.middlebury.edu/stereo/eval/*), où un seuil de précision de 90% (soit 90% des points sont dans 1 mm du modèle de la vérité-terrain) et un seuil de complétude de 95% (soit 95% des points sont dans 1.25 mm du modèle de la vérité-terrain) sont utilisés. La précision et la complétude de notre reconstruction de l'ensemble de données "Temple" étaient respectivement 8.97 mm et 72.1%, et 8.12 mm et 97.2% pour l'ensemble de données "Dino". Il est clair que la méthode proposée permet des reconstructions acceptables mais elle a des performances inférieures, en termes de précision et de complétude, que d'autres méthodes modernes. Ceci est principalement dû à la contrainte de visibilité ordinale de la coloration des voxels. Les résultats devraient être améliorés d'une manière significative si nous avions utilisé toutes les images de l'ensemble de données. Dans notre cas, seules les huit caméras, qui sont placées sur le même côté d'un plan, sont utilisées au lieu des seize caméras utilisées par les méthodes comparées. En outre, l'ordre du parcours est important parce que la visibilité des voxels est prise en compte lors du test de la photo-consistance. Cependant, la méthode proposée est entièrement automatique, simple et à faible coût de calcul. En effet, comparativement à d'autres méthodes récentes [37] [50] [68] [105], elle a une qualité inférieure car ces dernières utilisent plus de vues (16 vues) et des algorithmes plus complexes, ce qui explique le coût de calcul élevé de ces méthodes. En l'occurrence, puisque chaque voxel est visité exactement une fois,



(a) (b) (c) (d)

FIGURE 3.19 – Résultats de la reconstruction pour les scènes "TempleSparseRing" et "DinoSparseRing" par : (a) [28], (b) [141], (c) [39] et (d) la méthode proposée.

les complexités de l'espace et du temps de calcul de la méthode proposée sont toutes les deux linéaires par rapport au nombre des images d'entrée, qui sont extrêmement réduites dans notre cas [19].

	"TempleSparseRing"	"DinoSparseRing"
Méthode proposée	8.97, 72.1%, 310	8.12, 70.0%, 375
[152]	2.77, 79.4%, 2423	1.18, 90.8%, 2461
[148]	1.53, 85.4%, 8400	1.26, 89.3%, 10079

TABLE 3.3 – Comparaison objective de la méthode proposée avec d'autres modèles modernes de reconstruction 3D (qui ne sont pas des méthodes de coloration des voxels). Le premier nombre $x.xx$ mesure la précision (distance en mm), le second nombre $xx.x\%$ spécifie la complétude et le troisième nombre indique le temps d'exécution (en secondes) sur un processeur Intel® *Pentium*4 3 GHz.

3.5 Conclusion

J'ai synthétisé le long de ce chapitre nos travaux de recherche sur l'aide au diagnostic médical basé sur l'analyse des images. Malgré la diversité des architectures proposées dans la littératures, les systèmes développés ne produisent pas en général des résultats satisfaisants, décourageant ainsi les cliniciens à les utiliser dans leurs analyses quotidiennes. En effet, les techniques de recalage, de classification et de recherche des images sont utilisées séparément, contrairement au radiologue qui combine les différentes tâches dans un seul examen. En effet, ces techniques sont complémentaires et aucune technique utilisée seule ne permet de diagnostiquer tous les cas. Prenons l'exemple du cancer des seins, l'état d'une lésion peut être déduite en utilisant une méthode de classification, alors que la détection d'une asymétrie entre les seins gauche et droit requiert l'utilisation des techniques de recalage d'images. L'utilisation des informations complémentaires contenues dans des images prises à des dates antérieures, à partir des projections différentes ou avec d'autres modalités, peut aussi améliorer le diagnostic. Dans ce cadre, nous avons proposé une architecture pour les systèmes d'aide au diagnostic, qui utilise conjointement

les techniques de classification, de recalage et de recherche des images par le contenu. L'architecture proposée se compose de quatre modules, dont trois sont exécutés en parallèle (classification, recalage et recherche d'images) et un quatrième module qui combine les décisions des premiers modules. Cette architecture a été validée dans le cadre du diagnostic du mélanome. La solution proposée estime la probabilité de malignité d'une lésion de la peau en utilisant séparément la classification et la recherche des images par le contenu, et les deux opinions obtenues sont ensuite fusionnées en utilisant la théorie de Dempster-Shafer. Les résultats obtenus montrent que la fusion permet d'améliorer la performance du diagnostic du mélanome. En effet, en utilisant une base de 200 images, l'architecture proposée assure un taux de reconnaissance supérieur à 89%. En particulier, le but ultime du module de CBIR serait de fournir aux dermatologues un outil d'aide à la décision sous la forme d'un affichage des anciens cas pertinents, avec leurs pathologies avérées et d'autres informations appropriées. Par ailleurs, ce module pourrait être utilisé afin de présenter les cas qui ne sont pas seulement semblables de point de vue diagnostic, mais aussi semblables en apparence visuelle, même s'ils correspondent à des cas cliniques différents. Dans le même contexte de diagnostic du mélanome, nous avons aussi proposé une méthode structurale pour la détection automatique du réseau de pigment dans les images dermatoscopiques, qui représente un symptôme très fibale de malignité. Suite à l'extraction de la lésion à partir de l'image, la méthode proposée est basée sur un filtrage LoG afin de détecter les trous et les autres structures au sein de cette lésion. Ensuite, un processus de seuillage permet de filtrer uniquement les trous appartenant au réseau de pigment tout en excluant les autres structures rondes (les points, les globules et les bulles d'huile). L'originalité de la méthode proposée réside dans l'évaluation floue du degré d'appartenance d'un trou au réseau de pigment, ce qui permet de garder le maximum de candidats et de repousser la décision jusqu'à l'obtention de plus amples informations. Les trous retenus sont ensuite reliés, en vérifiant une contrainte spatiale, via un graphe représentant le réseau de pigment. Cette méthode a assuré une aire sous la courbe ROC de 0.821 pour détecter les réseaux de pigment avec un taux de classification correcte de 85% sur une base de 122 images.

Nous avons aussi proposé une méthode rapide pour le recalage non-rigide des images médicales. La méthode introduite améliore le coût de calcul de la méthode de subdivision hiérarchique tout en conservant une bonne qualité de recalage. En effet, bien que cette méthode semble être parmi les méthodes de recalage non-rigide les plus rapides, son coût de calcul est un véritable défi pour être intégrée dans les routines cliniques. Nous avons proposé d'appliquer l'approche de subdivision progressive sur une pyramide d'images (au lieu d'une image) afin d'accélérer l'algorithme. Les résultats obtenus sur des mammographies et des IRMs cérébrales montrent que cette modification permet de diminuer énormément le temps de calcul sans pour autant dégrader la qualité de recalage. La subdivision progressive des images à recaler en un nombre croissant d'imagettes permet à la méthode proposée, qui est complètement automatique, de tenir compte des déformations globales et locales. En effet, le recalage affine effectué au début permet de réduire les différences globales entre les deux images, alors que la subdivision des images en un nombre plus grand d'imagettes permet de récupérer les différences locales. En outre, dans le contexte de la reconstruction des modèles 3D à partir des images médicales, nous avons commencé par proposer une amélioration à la méthode de coloration des voxels, qui est une méthode populaire de reconstruction d'un modèle 3D à partir d'un ensemble d'images 2D calibrées. En effet, la plupart des méthodes de coloration des voxels souffrent de plusieurs limitations. Tout d'abord, elles supposent une texture suffisante et sont généralement très sensibles au bruit et aux erreurs de calibration. En outre, la représentation voxel ignore la continuité de la forme. Un autre inconvénient des tests existants de photo-consistance est que la qualité de reconstruction est très sensible aux valeurs des seuils utilisés. Les seuils stricts peuvent produire des reconstructions précises, mais incomplètes avec beaucoup de détails manquants de l'objet filmé. Cependant, les seuils souples produisent une reconstruction plus complète, mais avec le risque d'inclure des voxels erronés. Ainsi, pour obtenir une reconstruction précise, de nombreuses valeurs doivent être testées pour les seuils . En outre, de nombreux voxels peuvent être éliminés par erreur, dans un effet de cascade, et ces voxels ne peuvent plus être récupérés ultérieurement. Pour surmonter ces limites, nous avons proposé un nouveau test automatique de photo-consistance qui minimise l'influence des seuils à l'aide de la logique floue, et qui intègre l'information spatiale en utilisant un

seuillage adaptatif par hystérésis afin de garantir le lissage et la cohérence spatiale des voxels conservés. Les premiers résultats obtenus ont montré que la méthode proposée est capable de reconstruire avec précision les scènes 3D à partir de quelques vues, en une seule étape et sans nécessité de seuils prédéfinis. En effet, les deux seuils d'hystérésis sont modélisés de manière appropriée à travers des suites adjacentes en fonction du nombre des images sur lesquelles le voxel est projeté.

Travaux en cours et perspectives

Depuis l'obtention de mon doctorat en informatique en Janvier 2006, mes travaux de recherche se sont focalisés sur plusieurs contributions qui ont été situées dans les chapitres précédents sur trois axes. Dans ces mêmes axes et outre les travaux déjà considérés, mes travaux actuels et futurs visent les points suivants. Dans le contexte de l'indexation multidimensionnelle et la structuration des grandes bases d'images, nous nous sommes récemment orientés vers le choix du partitionnement de l'espace de données au lieu du partitionnement des données. Ce choix a été motivé par le fait que les bases peuvent être à tout moment enrichies, et il faut ainsi définir un patron de l'espace de descripteurs où insérer directement les nouveaux éléments. En effet, la tâche de recherche est extrêmement coûteuse dans le cas des bases à grande échelle, car il faut évaluer la similarité entre l'image-requête et toutes les images de la base. Dans le but de résoudre ce problème de la dépendance exponentielle entre la complexité et la montée en échelle (la malédiction de la dimensionnalité), divers travaux ont efficacement adapté l'indexation approximative. Ces travaux se basent sur l'idée de la projection aléatoire des points modélisant les données. Ceci est dans le but de rapprocher la recherche du plus proche voisin en sacrifiant une perte prévisible de précision. Parmi les méthodes approximatives, le LSH (Locality-Sensitive Hashing) se présente comme l'une des techniques les plus performantes. De nombreuses extensions ont été appliquées sur cette technique pour résoudre ses limitations, telles que la définition du meilleur quantificateur, la gestion de l'espace mémoire nécessaire pour le stockage des données et la sélection des plus proches voisins approximatifs. Les différentes solutions basées sur le LSH peuvent être regroupées en deux classes : les méthodes indépendantes de données et les méthodes dépendantes de données. Les méthodes de la première classe n'utilisent pas une étape d'apprentissage contrairement à celle de la deuxième classe. Ceci permet aux fonctions de hachage développées dans le cadre des méthodes indépendantes d'accélérer le traitement et d'éviter surtout les problèmes de classification et des machines d'apprentissage, qui affectent les méthodes LSH dépendantes de données. La composante de base de l'approche LSH indépendante de données est sa fonction de hachage permettant d'optimiser le partage de l'espace vectoriel tout en garantissant que la probabilité de collision soit plus élevée pour des objets sémantiquement similaires. Dans ce cadre, une nouvelle fonction, inspirée du treillis E_8, a été proposée et adaptée au contexte d'indexation des images par les régions. Le choix d'utilisation de ce treillis comme quantificateur est justifié par le fait qu'il offre de meilleures performances de quantification pour les vecteurs uniformes. Ce choix permet aussi une très faible erreur quadratique moyenne. Nous avons mis en œuvre notre version de treillis E_8 en le modélisant selon notre vecteur descriptif de douze dimensions qui correspondent aux différentes résolutions d'ondelettes. La fonction proposée minimise les ressources mémoire en stockant uniquement les buckets non vides, tout en réduisant le nombre des tables stockées. En effet, des tables de hachage sont construites de telle sorte que chacune correspond à une fonction de hachage. Ensuite, puisque le nombre total de buckets peut être très grand, seuls les buckets non-vides sont conservés. Pour ce faire, une fonction de hachage universelle est utilisée pour obtenir, à partir du vecteur descriptif, une clé de hachage de type entier. Cette clé est sélectionnée à partir d'un intervalle suffisamment grand pour éviter, avec une forte probabilité, la collision de deux vecteurs distincts. Quant à la recherche, une étape approximative génère une liste restreinte de candidats, qui feront l'objet d'une recherche exacte par appariement de graphes. Les premiers résultats montrent l'amélioration apportée par la structure proposée en termes de temps de calcul et de qualité de recherche.

En outre, nous sommes en train de valider une technique d'annotation automatique des images, ba-

sée sur le contenu visuel des régions qui composent l'image sujet de l'annotation. L'objectif est d'associer automatiquement les descripteurs visuels de bas niveau des régions, à des caractéristiques sémantiques de haut niveau. Etant donné une base d'images annotées, nous avons introduit un modèle probabiliste pour capturer les relations entre les caractéristiques de bas niveau des régions et leurs classes sémantiques. En effet, l'annotation automatique est associée à un processus d'apprentissage. Ce processus extrait les informations visuelles, à partir d'une base d'images annotées manuellement, et les combine avec les informations sémantiques. Ceci permet de définir des liens entre des informations textuelles et le contenu visuel d'une région. La plupart des travaux existants utilisent différents types de descripteurs sans traiter le problème de la malédiction de la dimensionnalité. Nous résolvons ce problème en introduisant une version floue de la technique d'indexation multidimensionnelle VA-Files. Le modèle d'annotation proposé utilise la technique de VA-Files couplée à un modèle probabiliste afin de stocker l'information visuelle et de permettre une annotation rapide des images. Ceci permet de préparer une vision globale du contenu de la base d'images, ce qui facilite la redirection de l'utilisateur vers la classe d'images similaires à sa requête et accélère par la suite la procédure de la recherche. Les tests préliminaires sur la base "Corel" montrent que la technique proposée permet de réduire le coût de calcul tout en optimisant la qualité de l'annotation.

Par ailleurs, il nous semble très important de poursuivre l'effort de confrontation entre les différentes approches qui ne cessent d'apparaître dans le domaine de la segmentation des images médicales, notamment celles qui sont basées sur la texture. C'est la raison pour laquelle qu'une étape d'évaluation pertinente des techniques existantes doit être bien élaborée. Ceci permettra de choisir la "meilleure" technique en fonction de l'application visée et de la nature des images traitées, afin de mettre en place des techniques de segmentation interactives et souples. En outre, la coopération de plusieurs techniques de segmentation paraît une piste très fructueuse, et ce dans le but de profiter des avantages de chacune. Dans ce cadre, nous avons commencé par intégrer des connaissances spatiales au modèle des formes actives (ASM), pour minimiser la dépendance des résultats de l'étape d'initialisation et de la présence du bruit. En effet, nous avons considéré une connaissance spatiale supplémentaire basée sur les angles entre les objets, pour la localisation des structures anatomiques dans des images médicales. L'objectif est de forcer la forme active à se diriger vers le contour réel, même en présence d'autres objets de même apparence visuelle. La méthode proposée considère deux informations essentielles lors de la déformation de la forme active : une contrainte statistique de forme liée à l'objet sujet de la segmentation et une contrainte de direction. Contrairement aux modèles déformables paramétriques et géométriques, les modèles statistiques se basent sur une analyse statistique préalable sur les variations de la structure étudiée. Cette analyse permet de contraindre l'évolution du contour initial pour qu'il n'épouse que les contours "admissibles" de l'objet. En particulier, les modèles d'apparence actifs permettent d'intégrer des connaissances statistiques sur les formes et sur les textures dans le processus de segmentation. En plus de la modélisation statistique des variations qui permet de représenter, dans un même modèle mathématique, la géométrie et les modes de déformation des structures, ces modèles ont l'avantage de pouvoir être applicables même si les contours ne sont pas apparents. Ceci permet de réduire l'espace des solutions et d'aboutir souvent à une forme admissible. Aussi, étant donné que la structure du modèle est conservée tout au long du processus de convergence, l'influence du bruit et des perturbations locales est nettement moins importante. Notamment, l'intégration de la contrainte de direction permet d'éviter les collisions et l'éloignement des formes et de contrôler l'évolution des estimations proposées par l'ASM au cours de la phase de localisation. Ainsi, le modèle global de la variation des contours permet d'allier des connaissances de forme, de texture et de disposition spatiale. Les tests effectués pour détecter le ventricule gauche dans des images scintigraphiques, ont prouvé la performance du modèle proposé même sur des images peu contrastées et avec des frontières mal définies.

En outre, dans l'objectif d'améliorer la qualité du recalage, une étude sur la relation entre le modèle de transformation utilisé dans le processus de recalage et le type de déformation susceptible de survenir à l'organe, a été réalisée. Cette étude nous a permis de constater l'insuffisance d'une seule transformation à représenter toutes les déformations. Ainsi, pour choisir le modèle de déformation permettant de modéliser au mieux les différences entre les paires mammaires, nous avons mené une comparaison em-

pirique entre les différentes transformations, et ce en se basant sur des critères hétérogènes incluant des mesures de similarité, des mesures de qualité et des mesures d'erreurs. Les premiers résultats montrent la supériorité de filtrage moyen, même par rapport aux splines à plaques minces souvent évoqués dans le cadre du recalage par subdivision progressive.

Nous nous sommes aussi intéressés à l'utilisation des transformations multi-échelles pour la classification des images mammographiques. En effet, les recherches récentes confirment l'efficacité de ces transformations, surtout celles basées sur les curvelettes. Cependant, la manière avec laquelle l'ensemble des attributs est extrait à partir des transformées en curvelettes est critique et peut affecter la qualité de la classification. En effet, un problème de malédiction de la dimensionnalité survient, si tous les coefficients de la transformée en curvelettes sont utilisés. Un tel problème augmente le temps de calcul et diminue les performances du système. Ainsi, notre objectif est d'extraire une famille compacte et discriminante pour la description des mammographies à partir de la transformée en curvelettes. Pour cela, nous avons proposé d'utiliser les moments de premier ordre pour chaque niveau de chaque bande des curvelettes. En plus, une analyse en composantes principales est appliquée afin de réduire le nombre d'attributs tout en optimisant la pertinence de la classification. Par rapport aux techniques qui s'appuient sur la représentation de l'image en utilisant une fraction de la plus grande transformation en curvelettes, la technique proposée a l'avantage d'extraire un ensemble réduit d'attributs. Elle permet non seulement de réduire énormément le coût de calcul, mais aussi d'optimiser la qualité de la classification.

En guise de perspectives, nous entamons le cas des bases d'images dynamiques, sachant qu'aussi bien la quantité d'images d'apprentissage que le nombre de classes augmentent au fil du temps. Pour résoudre ce problème, qui est relativement inexploré, nous envisageons d'adopter les forêts aléatoires de façon que les nœuds de décision soient basés sur une classification par la moyenne de la classe la plus proche, tout en ayant la possibilité d'intégrer de nouvelles classes. L'objectif est de pouvoir intégrer des données provenant des nouvelles classes, de manière à étendre les modèles de classification précédemment établis, à la place de procéder à un nouveau réapprentissage complet. Nous considérons en particulier le cas de la recherche des images médicales par le contenu dans les grandes bases d'images. En effet, nous étudierons les descripteurs qui peuvent indexer une image médicale, ainsi que les mesures de similarité sur ces descripteurs. Nous nous intéressons notamment aux descripteurs de texture au niveau des régions afin de proposer une indexation de haut niveau dans le propre langage des cliniciens. Notre objectif est de récupérer des images avec la même pathologie et en même niveau de sévérité de la maladie que celui de l'image en entrée, ce qui impose la caractérisation des images par les lésions qu'elles contiennent. Pour l'évaluation de la similarité entre les images, nous poursuivrons la nouvelle approche qui consiste à apprendre cette similarité à partir des exemples dispensés par des observateurs humains. Nous explorons également la possibilité d'intégrer les interactions des cliniciens afin d'incorporer un bouclage de pertinence dans le processus d'apprentissage de l'évaluation de la similarité entre les images médicales.

Enfin, grâce aux résultats encourageants produits par notre approche de décomposition d'une vidéo en arrière-plan et avant-plan, nous envisageons d'entamer l'étape de reconstruction volumétrique de la scène dans un environnement multi-caméra, à partir des silhouettes des multiples objets filmés. Ainsi, un nuage de points 3D sera extrait, pour chaque vue, par une technique de corrélation multi-image en multi-résolution. Le résultat est un maillage 3D de la portion de la scène captée dans chaque image. Notre idée consiste à estimer le mouvement relatif entre les poses successives en se basant essentiellement sur le contenu visuel de la scène 3D. Pour ce faire, nous proposons de fusionner les silhouettes en 3D afin de tenir compte d'une manière optimale, de la contribution de toutes les images à la reconstruction des silhouettes finales des objets 3D. Le résultat d'une telle fusion, contient naturellement une information de forme, et peut donc être utilisée pour des applications de modélisation classique à partir des images.

Bibliographie

[1] A.A. Ali et T.M. Deserno : A systematic review of automated melanoma detection in dermatoscopic images and its ground truth data. *SPIE 8318, Medical Imaging*, 1(1):1–11, 2013.

[2] F. Amato, A. López, E.M. Pena-Méndez, P. Vanhara, A. Hampl et J. Havel : Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2):47–58, 2013.

[3] S. Amri, W. Barhoumi et E. Zagrouba : Traitement de l'ombrage en vue d'une détection précise des objets mobiles dans des vidéos complexes. *In Ateliers sur le Traitement et l'Analyse de l'Information : Méthodes et Applications*, pages 543–548, Hammamet, Tunisie, 2009.

[4] S. Amri, W. Barhoumi et E. Zagrouba : A robust framework for joint background/foreground segmentation in complex video scenes filmed with freely moving camera. *Multimedia Tools and Applications*, 46(5):175–205, 2010.

[5] S. Amri, W. Barhoumi et E. Zagrouba : Unsupervised background reconstruction based on iterative median blending and spatial segmentation. *In IEEE International Conference on Imaging Systems and Techniques*, pages 411–416, Thessalonique, Grèce, 2010.

[6] S. Amri, W. Barhoumi et E. Zagrouba : Detection and matching of multiple occluded moving people for human tracking in color video sequences. *International Journal of Signal and Imaging Systems Engineering*, 4(3):153–163, 2011.

[7] S. Amri, W. Barhoumi et E. Zagrouba : A region-based approach for multi people segmentation under occlusion. *In International Workshop on Advanced Image Technology*, pages 1–4, Jakarta, Indonésie, 2011.

[8] M. Anantha, R. Moss et W. Stoecker : Detection of pigment network in dermatoscopy images using texture analysis. *Computerized Medical Imaging and Graphics*, 28(5):225–234, 2004.

[9] A. Andronache, M. Von Siebenthal, G. Szekely et P. Cattin : Non-rigid registration of multi-modal images using both mutual information and cross-correlation. *Medical Image Analysis*, 12(1):3–15, 2007.

[10] M. Arevalillo-Herráez, F.J. Ferri et S. Moreno-Picot : Distance-based relevance feedback using a hybrid interactive genetic algorithm for image retrieval. *Applied Soft Computing*, 11(2): 1782–1791, 2011.

[11] M. Auer, P. Regitnig et G.A. Holzapfel : An automatic nonrigid registration for stained histological sections. *IEEE Transactions on Image Processing*, 14(4):475–486, 2005.

[12] K. Ayari, W. Barhoumi et E. Zagrouba : Online detection of multi-person based on iterative background subtraction. *In IEEE International Conference on Image Information Processing*, pages 721–726, Shimla, Inde, 2013.

[13] A. Ayoub, A. Hajdu et A. Nagy : Automatic detection of pigmented network in melanoma dermoscopic images. *International Journal of Computer Science and Communication Security*, 2(1):58–63, 2012.

[14] B. Babenko, M.H. Yang et S. Belongie : Visual tracking with online multiple instance learning. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 983–990, Miami, USA, 2009.

[15] J.F. Bach, O. Houdé, P. Léna et S. Tisseron : L'enfant et les écrans. Rapport de recherche, Académie des Sciences, 2013.

[16] N. Baka, B.L. Kaptein, M. De Bruijn, T. Van Walsum, J.E. Giphart, W.J. Niessen et B.P.F. Lelieveldt : 2d-3D shape reconstruction of the distal femur from stereo x-ray imaging using statistical shape models. *Medical Image Analysis*, 15(6):840–850, 2011.

[17] M.C. Bakkay, W. Barhoumi et E. Zagrouba : Mise à jour dynamique de l'image de référence pour l'optimisation du résume vidéo en ligne par multiple mosaïques. *In Ateliers de Traitement et d'Analyse de l'Information : Méthodes et Applications*, pages 135–144, Hammamet, Tunisie, 2011.

[18] W. Barhoumi et A. Baazaoui : Pigment network detection in dermatoscopic images for melanoma diagnosis. *Innovation and Research in BioMedical Engineering*, 35(5):128–138, 2014.

[19] W. Barhoumi, M.C. Bakkay et E. Zagrouba : Automated photo-consistency test for voxel colouring based on fuzzy adaptive hysteresis thresholding. *IET Image Processing*, 7(8):713–724, 2013.

[20] W. Barhoumi, M.C. Bakkay et E. Zagrouba : An online approach for multi-sprite generation based on camera parameters estimation. *Signal, Image and Video Processing*, 7(5):843–853, 2013.

[21] W. Barhoumi, S. Dhahbi et E. Zagrouba : A collaborative system for pigmented skin lesions malignancy tracking. *In IEEE International Workshop on Imaging Systems and Techniques*, pages 1–6, Cracovie, Pologne, 2007.

[22] W. Barhoumi, A. Gallas et E. Zagrouba : Effective region-based relevance feedback for interactive content-based image retrieval. *New Directions in Intelligent Interactive Multimedia Systems and Services*, 226:177–187, 2009.

[23] W. Barhoumi et E. Zagrouba : Segmentation en régions guidée par l'appariement pour un couple stéréoscopique non calibré. *In Ateliers de Traitement et d'Analyse d'Information : Méthodes et Applications*, pages 287–292, Hammamet, Tunisie, 2003.

[24] W. Barhoumi et E. Zagrouba : On-the-fly extraction of key frames for efficient video summarization. *AASRI Procedia*, 4:78–84, 2013.

[25] W. Barhoumi, E. Zagrouba, B. Solaiman et F. Ghorbel : Fusion de l'information par la théorie de l'évidence : application en diagnostic du mélanome. *In Conférence Internationale Sciences Electroniques, Techniques de l'Information et des Télécommunications*, pages 1–8, Sousse, Tunisie, 2003.

[26] A. Barron : Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(4):930–945, 1993.

[27] T. Basha, S. Avidan, A. Hornung et W. Matusik : Structure and motion from scene registration. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 1426–1433, Providence, USA, 2012.

[28] J.S.D. Bonet et P.A. Viola : Roxels : Responsibility weighted 3D volume reconstruction. *In IEEE International Conference on Computer Vision*, pages 418–425, Corfou, Grèce, 1999.

[29] F.L. Bookstein : Principal warps : Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[30] S. Boughorbel, J.P. Tarel et N. Boujemaa : Generalized histogram intersection kernel for image recognition. *In IEEE International Conference on Image Processing*, pages 161–164, Gênes, Italie, 2005.

[31] L.G. Brown : A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[32] M. Brückner, F. Bajramovic et J. Denzler : Intrinsic and extrinsic active self-calibration of multi-camera systems. *Machine Vision and Applications*, 25(2):389–403, 2014.

[33] S.R. Bulò, M. Rabbi et M. Pelillo : Content-based image retrieval with relevance feedback using random walks. *Pattern Recognition*, 44(9):2109–2122, 2011.

[34] R.T. Calumby, R. Da Silva Torres et M.A. Gonçalves : Multimodal retrieval with relevance feedback based on genetic programming. *Multimedia Tools and Applications*, 69(3):991–1019, 2014.

[35] J. Chaabani, W. Barhoumi et E. Zagrouba : Recherche interactive d'images sur le web par désambigüisation des requêtes textuelles. *In Conférence Maghrébine d'Extraction et de Gestion de Connaissances*, pages 145–156, Alger, Algérie, 2010.

[36] Y. Chahir, S. Schüpp et L. Chan : Indexation d'images utilisant une segmentation par ensembles de niveaux. *In Extraction et Gestion des Connaissances*, pages 387–392, Montpellier, France, 2002.

[37] J. Chang, H. Park, I. Park, K. Lee et S. Lee : GPU-friendly multi-view stereo reconstruction using surfel representation and graph cuts. *Computer Vision and Image Understanding*, 115(5): 620–634, 2011.

[38] D.J. Chen, H.T. Chen et L.W. Chang : Video object cosegmentation. *In ACM Multimedia*, pages 805–808, Nara, Japan, 2012.

[39] V. Chhabra : *Reconstructing specular objects with image based rendering using color caching*. Thèse de doctorat, Worcester Polytechnic Institute, 2001.

[40] P. Chiranjeevi et S. Sengupta : Neighborhood supported model level fuzzy aggregation for moving object segmentation. *IEEE Transactions on Image Processing*, 23(2):645–657, 2013.

[41] W.C. Chiu et M. Fritz : Multi-class video co-segmentation with a generative multi-video model. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 321–328, Portland, USA, 2013.

[42] A. Ciobanu, M. Costin et T. Barbu : Image categorization based on computationally economic lab colour features. *Soft Computing Applications - Advances in Intelligent Systems and Computing*, 195(1):585–593, 2013.

[43] A. Colombari, A. Fusiello et V. Murino : Segmentation and tracking of multiple video objects. *Pattern Recognition*, 40(1):1307–1317, 2007.

[44] D. Comaniciu, V. Ramesh et P. Meer : Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, 2003.

[45] K.M. Cordoro, D. Gupta, I.J. Frieden, T. McCalmont et M. Kashani-Sabet : Pediatric melanoma : Results of a large cohort study and proposal for modified ABCD detection criteria for children. *Journal of the American Academy of Dermatology*, 68(6):913–925, 2013.

[46] W.R. Crum, T. Hartkens et D.G. Hill : Non-rigid image registration : Theory and practice. *The British Journal of Radiology*, 77(2):140–153, 2004.

[47] R. Cucchiara, C. Grana, M. Piccardi et A. Prati : Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003.

[48] F.S. Da Silva, X.M. Ribeiro, J. Batista Neto, C. Traina-Jr et A. Traina : Improving the ranking quality of medical image retrieval using a genetic feature selection method. *Decision Support Systems*, 51(4):810–820, 2011.

[49] C. Dehais, M. Douze, G. Morin et V. Charvillat : Augmented reality through real-time tracking of video sequences using a panoramic view. *In International Conference on Pattern Recognition*, pages 995–998, Cambridge, UK, 2004.

[50] Y. Deng, Y. Liu, Q. Dai, Z. Zhang et Y. Wang : Noisy depth maps fusion for multi-view stereo via matrix completion. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):566–582, 2012.

[51] S. Dhahbi, W. Barhoumi et E. Zagrouba : A cost efficient approach for automatic non-rigid registration of medical images. *In International Workshop on Medical Image Analysis and Description for Diagnosis Systems*, pages 3–12, Porto, Portugal, 2009.

[52] M. Douze, V. Charvillat et B. Thiesse : Mosaïques d'images par approximations successives. *In Journées Francophones des Jeunes Chercheurs en Analyse d'Images et Perception Visuelle (ORASIS)*, pages 97–102, Cahors, France, 2001.

[53] G. Dreyfus : *Les réseaux de neurones : pourqoui et pour quoi faire.* Eds Eyrolles, 2002.

[54] O. Duchenne, A. Joulin et J. Ponce : A graph-matching kernel for object categorization. *In IEEE International Conference on Computer Vision*, pages 1792–1799, Barcelone, Espagne, 2011.

[55] T. Elamsy, A. Habed et B. Boufama : Self-calibration of stationary non-rotating zooming cameras. *Image and Vision Computing*, 32(3):212–226, 2014.

[56] S.Y. Elhabian, K.M. El-Sayed et S.H. Ahmed : Moving object detection in spatial domain using background removal techniques. *Recent Patents on Computer Science*, 1(1):32–54, 2008.

[57] I. Everts, J.C. Van Gemert et T. Gevers : Evaluation of color spatio-temporal interest points for human action recognition. *IEEE Transactions on Image Processing*, 23(4):1569–1580, 2014.

[58] M.Y Fang, Y.H. Kuan, C.M. Kuo et C.H. Hsieh : Effective image retrieval techniques based on novel salient region segmentation and relevance feedback. *Multimedia Tools and Applications*, 57(3):501–525, 2012.

[59] D. Farin, M. Haller, A. Krutz et T. Sikora : Recent developments in panoramic image generation and sprite coding. *In IEEE International Workshop on Multimedia Signal Processing*, pages 64–69, Cairns, Australie, 2008.

[60] I.H. Fernandez : *Computer-aided detection and classification of pigment network in pigmented skin lesions.* Thèse de doctorat, Faculty of Biomedical Engineering, Politecnico di Milano, 2011.

[61] M. Fontani, T. Bianchi, A. De Rosa, A. Piva et M. Barni : A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence. *IEEE Transactions on Information Forensics and Security*, 8(4):539–607, 2013.

[62] P.F. Gabriel, J.G. Verly, J.H. Piater et A. Genon : The state of the art in multiple object tracking under occlusion in video sequences. *In Advanced Concepts for Intelligent Vision Systems*, pages 166–173, Ghent, Belgique, 2003.

[63] A. Gallas, W. Barhoumi et E. Zagrouba : Bouclage de pertinence négatif pour la recherche des images à base de descripteurs de sous-bandes d'ondelettes. *Traitement du Signal*, 29(1-2):157–177, 2012.

[64] A. Gallas, W. Barhoumi et E. Zagrouba : Image retrieval based on wavelet sub-bands and fuzzy weighted regions. *In International Conference on Communications and Information Technology*, pages 26–30, Hammamet, Tunisie, 2012.

[65] A. Gallas, W. Barhoumi et E. Zagrouba : Image retrieval by comparison between complete oriented graphs of fuzzy regions. *In IEEE International Conference on Intelligent Computer Communication and Processing*, pages 173–180, Cluj-Napoca, Roumanie, 2012.

[66] V.S. Gandyer : Colorogram : A color feature descriptor for human blob labeling. *In International Conference on Signal and Image Processing*, pages 331–341, Coimbatore, Inde, 2012.

[67] R.F.C. Guerreiro et P.M.Q. Aguiar : Global motion estimation : Feature-based, featureless, or both ?! *In International Conference on Image Analysis and Recognition*, pages 721–730, Póvoa de Varzim, Portugal, 2006.

[68] J.Y. Guillemaut et A. Hilton : Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *International Journal of Computer Vision*, 93(1):73–100, 2011.

[69] I. Haritaoglu, D. Harwood et L.S. Davis : W4 : real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.

[70] C. Harris et A. Stephens : Combined corner and edge detector. *In Alvey Vision Conference*, pages 147–151, Manchester, UK, 1988.

[71] R.I. Hartley : Self-calibration of stationary cameras. *International Journal of Computer Vision*, 22(1):5–23, 1997.

[72] P. Hellier, C. Barillot, E. Memin et P. Perez : Hierarchical estimation of a dense deformation field for 3-D robust registration. *IEEE Transactions on Medical Imaging*, 20(5):388–402, 2001.

[73] P.S. Hiremath, S. Shivashankar et J.D. Pujari : Wavelet based features for color texture classification with application to cbir. *Journal of Computer Science and Network Security*, 6(9): 124–133, 2006.

[74] D.S. Hochbaum et V. Singh : An efficient algorithm for co-segmentation. *In International Conference on Computer Vision*, pages 269–276, Kyoto, Japan, 2009.

[75] A. Hornung et L. Kobbelt : Robust and efficient photo-consistency estimation for volumetric 3D reconstruction. *In European Conference on Computer Vision*, pages 179–190, Graz, Autriche, 2006.

[76] D.W. Hosmer et Lemeshow S. : *Applied logistic regression*. John Wiley and Sons, 2004.

[77] M. Hu : Visual pattern recognition by moment invariant. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[78] K.A. Hua, N. Yu et D. Liu : Query decomposition : a multiple neighborhood approach to relevance feedback processing in content-based image retrieval. *In International Conference on Data Engineering*, pages 3–8, Atlanta, USA, 2006.

[79] W. Huang, Y. Gao et K.L. Chan : A review of region-based image retrieval. *Journal of Signal Processing Systems*, 59(2):143–161, 2010.

[80] K. Iqbal, M.O. Odetayo et A. James : Content-based image retrieval approach for biometric security using colour, texture and shape features controlled by fuzzy heuristics. *Journal of Computer and System Sciences*, 78(4):1258–1277, 2012.

[81] Yue. J., Z. Li, L. Liu et Z. Fu : Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modelling*, 54(3-4):1121–1127, 2011.

[82] W.M. Jay : Visual field defects. *American Family Physician*, 24(2):138–142, 1981.

[83] R. Jonker et A. Volgenant : A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[84] R. E. Kalman et R. S. Bucy : New results in linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 83:95–108, 1961.

[85] S.B. Kang et R. Szeliski : Extracting view-dependent depth maps from a collection of images. *International Journal of Computer Vision*, 58(2):139–163, 2004.

[86] O. Kao : On parallel image retrieval with dynamically extracted features. *Parallel Computing*, 34(12):700–709, 1961.

[87] C. Käs et H. Nicolas : Compressed domain indexing of scalable h.264/svc streams. *Signal Processing : Image Communication*, 24(6):484–498, 2009.

[88] M.L. Kherfi et D. Ziou : Relevance feedback for cbir : A new approach based on probabilistic feature weighting with positive and negative examples. *IEEE Transactions on Image Processing*, 15(4):1017–1030, 2006.

[89] D.H. Kim, J.W. Song et J.H. Lee : A hybrid region weighting approach for relevance feedback in region-based image search on the web. *In Conference on Current Trends in Theory and Practice of Computer Science*, pages 705–715, Harrachov, République Tchèque, 2001.

[90] D.H. Kim et S.H. Yu : A new region filtering and region weighting approach to relevance feedback in content-based image retrieval. *Journal of Systems and Software*, 81(9):1525–1538, 2008.

[91] E. Kim, H. Li et X. Huang : A hierarchical image clustering cosegmentation framework. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 686–693, Providence, USA, 2012.

[92] H. Kim et K.S. Hong : Practical self-calibration of pan-tilt cameras. *IEE Vision, Image and Signal Procesing*, 148(5):349–355, 2001.

[93] J.F. Krucker, G.L. LeCarpentier, J.B. Fowlkes et P.L. Carson : Rapid elastic image registration for 3-D ultrasound. *IEEE Transactions on Medical Imaging*, 21(11):1384–1394, 2002.

[94] M. Kunter, A. Krutz, M. Mandal et Sikora T. : Optimal multiple sprite generation based on physical camera parameter estimation. *In SPIE Visual Communications and Image Processing Conference*, pages 1–10, San Jose, USA, 2007.

[95] I.S. Kuo et L.H. Chen : A fast multisprite generator with near-optimum coding bit-rate. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(2):331–353, 2009.

[96] C. Leung, B. Appleton, M. Buckley et C. Sun : Embedded voxel colouring with adaptive threshold selection using globally minimal surfaces. *International Journal of Computer Vision*, 99(2):215–231, 2012.

[97] G.S. Li, C.Y. Hsieh et W.N. Lie : Sprite generation for hole filling in depth image-based rendering. *In IEEE International Conference on Image Processing*, pages 5402–5406, Paris, France, 2014.

[98] H.J. Li, S.X. Lin, Y.D. Zhang et K. Tao : Automatic video-based analysis of athlete action. *In International Conference on Image Analysis and Processing*, pages 295–210, Modè, Italie, 2007.

[99] J. Li et J.Z. Wang : Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[100] J. Li, J.Z. Wang et G. Wiederhold : IRM : integrated region matching for image retrieval. *In ACM Multimedia*, pages 147–156, Los Angeles, USA, 2000.

[101] B. Likar et F. Pernus : A hierarchical approach to elastic registration based on mutual information. *Image and Vision Computing*, 19(1-2):33–44, 2001.

[102] P.J. Lisboa et A.F. Taktak : The use of artificial neural networks in decision support in cancer : A systematic review. *Neural Networks*, 19(4):408–415, 2006.

[103] J. A. Little, D. Hill et D.J. Hawkes : Deformations incorporating rigid structures. *Computer Vision and Image Understanding*, 66(2):223–232, 1997.

[104] X. Liu, H. Yao, Y. Chen et W. Gao : An active volumetric model for 3D reconstruction. *In IEEE International Conference on Image Processing*, pages 11–14, Gênes, Italie, 2005.

[105] Y. Liu, X. Cao, Q. Dai et W. Xu : Continuous depth estimation for multi-view stereo. *In IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2121–2128, Miami, USA, 2009.

[106] Z.G. Liu, Q. Pan et J. Dezert : Classification of uncertain and imprecise data based on evidence theory. *Neurocomputing*, 133(4):459–470, 2014.

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

[108] I. Maglogiannis et C.N. Doukas : Overview of advanced computer vision systems for skin lesions characterization. *IEEE Transactions on Information Technology in Biomedicine*, 13(5): 721–733, 2009.

[109] J.A. Maintz, E.H. Meijering et M.A. Viergever : General multimodal elastic registration based on mutual information. *Medical Imaging*, 3338(2):144–154, 1998.

[110] T. Makela, P. Clarysse, O. Sipila, N. Pauna, Q.C. Pham, T. Katila et I.E. Magnin : A review of cardiac image registration methods. *IEEE Transactions on Medical Imaging*, 21(9):1011–1021, 2002.

[111] J. Martinet : Human-centered region selection and weighting for image retrieval. *In International Conference on Computer Vision Theory and Applications*, pages 729–734, Barcelone, Espagne, 2013.

[112] H. Merdassi, W. Barhoumi et E. Zagrouba : Color images co-segmentation based on fuzzy local-entropy classification. *Multimedia and Signal Processing, Communications in Computer and Information Science*, 346:240–248, 2012.

[113] S. Messelodi et C.M. Modena : Scene text recognition and tracking to identify athletes in sport videos. *Multimedia Tools and Applications*, 63(2):521–545, 2013.

[114] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir et L. Van Gool : A comparaison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2006.

[115] J. Ning, L. Zhang, D. Zhang et C. Wu : Robust mean-shift tracking with corrected background-weighted histogram. *IET Computer Vision*, 6(1):62–69, 2012.

[116] A. Ober-Gecks, M. Zwicker et D. Henrich : Efficient GPU photo hull reconstruction for surveillance. *In International Conference on Distributed Smart Cameras*, pages 1–8, Séville, Espagne, 2014.

[117] P. Ochs, J. Malik et T. Brox : Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, 2014.

[118] F.P Oliveira et J.M. Tavaresa : Medical image registration : A review. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(2):73–93, 2014.

[119] C. Panagiotakis, I. Grinias et G. Tziritas : Automatic human motion analysis and action recognition in athletics videos. *In European Signal Processing Conference*, pages 1–5, Florence, Italie, 2006.

[120] R. Paredes, T. Deselaers et E. Vidal : A probabilistic model for user relevance feedback on image retrieval. *In Joint Workshop on Machine Learning and Multimodal Interaction*, pages 260–271, Utrecht, Pays-Bas, 2008.

[121] S. Philipp-Foliguet, J. Gony et P.H. Gosselin : Frebir : An image retrieval system based on fuzzy region matching. *Computer Vision and Image Understanding*, 113(6):693–707, 2009.

[122] O.P. Popoola et K. Wang : Video-based abnormal human behavior recognition - A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews,*, 42(6):865–878, 2012.

[123] M. Rahman, C.B. Desai et P. Bhattacharya : Image retrieval-based decision support system for dermatoscopic images. *In IEEE Symposium on Computer-Based Medical Systems*, pages 285–290, Utah, USA, 2006.

[124] E. Ramasso, C. Panagiotakis, M. Rombaut, D. Pellerin et G. Tziritas : Human shape-motion analysis in athletics videos for coarse to fine action/activity recognition using transferable belief model. *Electronic Letters on Computer Vision and Image Analysis*, 7(4):32–50, 2009.

[125] A. Rubel, A. Naumenko et V. Lukin : A neural network based predictor of filtering efficiency for image enhancement. *In IEEE Microwaves, Radar and Remote Sensing Symposium*, pages 14–17, Kiev, Ukraine, 2014.

[126] J.C. Rubio, J. Serrat et A. López : Video co-segmentation. *In Asian Conference on Computer Vision*, pages 13–24, Daejeon, Corée de Sud, 2012.

[127] J.C. Rubio, J. Serrat, A. López et N. Paragios : Unsupervised co-segmentation through region matching. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 749–756, Providence, USA, 2012.

[128] Y. Rubner, C. Tomasi et L.J. Guibas : The earth mover's distance as a metric for image retrieval. *Pattern Recognition Letters*, 40(2):99–121, 2000.

[129] M. Sadeghi, M. Razmara, T.K. Lee et M.S. Atkins : A novel method for detection of pigment network in dermoscopic images using graphs. *Computerized Medical Imaging and Graphics*, 35(2): 137–143, 2011.

[130] M.A. Saleh, H. Hashim et N.M. Tahie : A low computational method of secure video streaming in mobile system. *In IEEE Symposium on Computer Applications and Industrial Electronics*, pages 193–197, Penang, Malaisie, 2014.

[131] A. Sanin, C. Sanderson et B.C. Lovell : Shadow detection : A survey and comparative evaluation of recent methods. *Pattern Recognition*, 45(4):1684–1695, 2012.

[132] H. Schweitzer, J.W. Bell et F. Wu : Very fast template matching. *In European Conference on Computer Vision*, pages 358–372, Copenhagen, Danemark, 2002.

[133] S. Seitz, B. Curless, J. Diebel, D. Scharstein et R. Szeliski : A comparison and evaluation of multi-view stereo reconstruction algorithms. *In IEEE International Conference on Computer Vision and Pattern Recognition*, pages 591–528, New York, USA, 2006.

[134] S.M. Seitz et C.R. Dyer : Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(3):151–173, 1999.

[135] G. Shafer : Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 4(5-6):323–362, 1990.

[136] Z. Shaoting, Z. Xiahai, J. Long et G. Lixu : Robust and efficient 3d registration via depth map-based feature point matching in image-guided neurosurgery. *In IEEE International Symposium on Biomedical Imaging*, pages 758–761, Pékin, Chine, 2014.

[137] X. Shen, L. Xu, Q. Zhang et J. Jia : Multi-modal and multi-spectral registration for natural images. *In European Conference on Computer Vision*, pages 309–324, Zurich, Suisse, 2014.

[138] E. Shilat, M. Werman et Y. Gdalyahu : Background recovery from multiple images. *In IEEE Digital Signal Processing and Signal Processing Education Meeting*, pages 135–140, Napa, USA, 2013.

[139] B. Shrestha, J. Bishop, K. Kam, X. Chen, R.H. Moss, W.V. Stoecker, S. Umbaugh, R.J. Stanley, M.E. Celebi, A.A. Marghoob, G. Argenziano et H.P. Soyer : Detection of atypical texture features in early malignant melanoma. *Skin Research and Technology*, 16(1):60–65, 2010.

[140] K. Simonyan, A. Vedaldi et A. Zisserman : Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1573–1585, 2014.

[141] G.G. Slabaugh, W.B. Culbertson, T. Malzbender, M.R. Stevens et R.W. Schafer : Methods for volumetric reconstruction of visual scenes. *International Journal of Computer Vision*, 57(3):179–199, 2004.

[142] A. Sobral et A. Vacavant : A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, 122(1):4–21, 2014.

[143] S. Sorlin et C. Solnon : Similarité de graphes : une mesure générique et un algorithme tabou réactif. *In Rencontres Nationales des Jeunes Chercheurs en Intelligence Artificielle*, pages 253–266, Nice, France, 2005.

[144] A. Sotiras, C. Davatzikos et N. Paragios : Deformable medical image registration : A survey. *IEEE Transactions on Medical Imaging*, 32(7):4–21, 2013.

[145] Z. Su, H.J. Zhang, S. Li et S. Ma : Relevance feedback in content-based image retrieval : Bayesian framework feature sub-spaces and progressive learning. *IEEE Transactions on Image Processing*, 12(8):942–937, 2003.

[146] R. Szeliski et H.Y. Shum : Creating full view panoramic image mosaics and environment maps. *In Annual Conference on Computer Graphics and Interactive Techniques*, pages 251–258, Los Angeles, USA, 1997.

[147] H. Tang, N. Boujemaa, Y. Chen et L. Deng : Modeling loosely annotated images using both given and imagined annotations. *Optical Engineering*, 50(12):127004–127004–8, 2011.

[148] S. Tran et L.S. Davis : 3D surface reconstruction using graph cuts with surface constraints. *In European Conference on Computer Vision*, pages 219–231, Graz, Autriche, 2006.

[149] S. Van Engeland, P. Snoeren, J. Hendriks et K. Karssemeijer : A comparison of methods for mammogram registration. *IEEE Transactions on Medical Imaging*, 22(11):1436–1444, 2003.

[150] S. Vicente, V. Kolmogorov et C. Rother : Cosegmentation revisited : Models and optimization. *In European Conference on Computer Vision*, pages 465–479, Crète, Grèce, 2010.

[151] S. Vicente, C. Rother et V. Kolmogorov : Object cosegmentation. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 2217–2224, Colorado Springs, USA, 2011.

[152] G. Vogiatzis, P. Torr et R. Cipolla : Multi-view stereo via volumetric graph-cuts. *In IEEE Conference on Computer Vision and Pattern Recognition*, pages 391–398, San Diego, USA, 2005.

[153] J.Z. Wang, J. Li et G. Wiederhold : Simplicity : Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[154] X. Wang : Intelligent multi-camera video surveillance : A review. *Pattern Recognition Letters*, 34(1):3–19, 2013.

[155] X.Y. Wang, J.W. Chen et H.Y. Yang : A new integrated SVM classifiers for relevance feedback content-based image retrieval using EM parameter estimation. *Applied Soft Computing*, 11(2): 2787–2804, 2011.

[156] S.P. Wilson, J. Fauqueur et N. Boujemaa : Mental search in image databases : Implicit versus explicit content query. *Machine Learning Techniques for Multimedia, Cognitive Technologies*, pages 189–204, 2008.

[157] G. Xiao, Y. Dong, Z. Liu et H. Wang : Supervised TV logo detection based on SVMS. *In IEEE International Conference on Network Infrastructure and Digital Content*, pages 174–178, Pékin, Chine, 2010.

[158] X. Xu, D. Lee, S.K. Antani, L.R. Long et J.K. Archibald : Using relevance feedback with short-term memory for content-based spine x-ray image retrieval. *Neurocomputing*, 71(10):2259–2269, 2009.

[159] C. Yan, N. Sang et T. Zhang : Local entropy-based transition region extraction and thresholding. *Pattern Recognition Letters*, 24(16):2935–2941, 2003.

[160] H. Yang et N. Ahuja : Automatic segmentation of granular objects in images : Combining local density clustering and gradient-barrier watershed. *Pattern Recognition*, 47(6):2266–2279, 2014.

[161] X. Yang et L. Cai : Adaptive region matching for region-based image retrieval by constructing region importance index. *IET Computer Vision*, 8(2):141–151, 2014.

[162] H.H. Yeh, J.Y. Chen, C.R. Huang et C.S. Chen : An adaptive approach for overlapping people tracking based on foreground silhouettes. *In IEEE International Conference on Image Processing*, pages 3498–3492, Hong Kong, 2010.

[163] L. Zagorchev et A. Goshtasby : A comparative study of transformation functions for nonrigid image registration. *IEEE Transactions on Image Processing*, 15(3):529–538, 2006.

[164] K. Zagoris, S.A. Chatzichristofis, N. Papamarkos et Y.S. Boutalis : img(Anaktisi) : A Web content based image retrieval system. *In International Workshop on Similarity Search and Applications*, pages 154–155, Prague, République Tchèque, 2009.

[165] E. Zagrouba et W. Barhoumi : A preliminary approach for the automated recognition for malignant melanoma. *Image Analysis and Stereology*, 23(2):121–135, 2004.

[166] E. Zagrouba et W. Barhoumi : An accelerated system for melanoma diagnosis based on subset feature selection. *Journal of Computing and Information Technology*, 13(1):69–82, 2005.

[167] E. Zagrouba, W. Barhoumi et S. Amri : An efficient image-mosaicing method based on multifeature matching. *Machine Vision and Applications*, 20(3):139–162, 2009.

[168] E. Zagrouba, S. Ouni et W. Barhoumi : A reliable image retrieval system based on spatial disposition graph matching. *International Review on Computers and Software*, 2(2):108–117, 2007.

[169] C. Zhang, D. Ouyang et J. Ning : An artificial bee colony approach for clustering. *Expert Systems with Applications*, 37(7):4761–4767, 2010.

[170] X. Zhang, T. Huang, Y. Tian et Gao. W : Background-modeling-based adaptive prediction for surveillance video coding. *IEEE Transactions on Image Processing*, 23(2):769–784, 2013.

[171] Y. Zhang, Z. Mao et K. Tian : Salient region detection for complex background images using integrated features. *Information Sciences*, 281(10):586–600, 2014.

[172] B. Zitova et J. Flusser : Image registration methods : A survey. *Image and Vision Computing*, 21(11):977–1000, 2003.