Filtrage de segments informatifs dans des vidéos

Christophe Guilmart

Centre de Mathématiques et de Leurs Applications

Thèse de Doctorat
de l'École Normale Supérieure de Cachan

présentée pour obtenir le grade de Docteur, spécialité « Mathématiques »

par

Christophe Guilmart

Filtrage de segments informatifs dans des vidéos

Thèse présentée et soutenue le 20 décembre 2011 devant le jury composé de :

M.	Serge MIGUET	Professeur des Universités	(Président)
M.	Patrick BOUTHEMY	Directeur de Recherches	(Rapporteur)
M.	Jean-Marc ODOBEZ	Enseignant-Chercheur	(Rapporteur)
M.	Edouard GEOFFROIS	Docteur	(Examinateur)
M.	Patrick PEREZ	Directeur de Recherches	(Directeur de thèse)
M.	Stéphane HERBIN	Docteur	(Co-directeur de thèse)

Remerciements

Je remercie tout d'abord la Direction générale de l'armement de m'avoir donné la chance d'effectuer une thèse. Cette thèse s'est déroulée à l'Office National d'Etudes et de Recherche Aérospatiale au département Traitement de l'Information et Modélisation.

Je tiens à remercier mon directeur de thèse Patrick Perez ainsi que mon encadrant à l'Onera Stéphane Herbin pour leurs différents conseils et la grande latitude qu'ils m'ont laissée dans le déroulement de la thèse. Je remercie notamment Patrick pour ses encouragements dans les moments de doute et Stéphane pour ses avis parfois très critiques mais pertinents.

Je remercie Messieurs Patrick Bouthemy et Jean-Marc Odobez d'avoir accepté de relire ce manuscrit et pour leurs remarques constructives.

Je remercie Serge Miguet d'avoir accepté de présider mon jury.

Je remercie également Edouard Geoffrois pour avoir accepté de faire partie de ce jury en tant qu'examinateur.

Je tiens à remercier l'ensemble du département à l'Onera pour l'ambiance conviviale pendant ces trois ans. Je remercie notamment Martial pour son optimisme et toute l'aide qu'il m'a apportée. Merci à Alain pour m'avoir permis de prendre du recul dans les moments de doute et Fabrice J. pour les différents échanges constructifs ainsi que ses remarques lors de la rédaction du mémoire. Mes remerciements également à Françoise pour sa gentillesse et sa bonne humeur. Je remercie l'ensemble des autres permanents du DTIM pour les différents échanges et les moments passés ensemble : Jonathan, Frédéric, Valérie, Philippe, Olivier, Valentina, Gilles, Fabrice S., Guy, Annie, Christian, Jean, Xavier et tous les autres.

Mes remerciements s'adressent enfin à tous les doctorants qui comme moi se sont lancés dans cette aventure et avec qui j'ai passé d'excellents moments : Joseph, Aurélien, Guillaume, Laure, Pauline, Paul, Thibault et Jeff. Leur présence, leur soutien et les nombreux échanges plus ou moins techniques et philosophiques m'ont notamment permis de surmonter les moments difficiles. J'ai ainsi pu passer ces trois années à leurs côtés dans des conditions positives et amicales.

Le 6 janvier 2012.

Table des matières

Liste des figures

Introduction

Contexte

Vidéo aérienne : disponibilité, applications et difficultés L'imagerie aérienne constitue un moyen efficace de surveillance ou de contrôle de zones autrement difficiles à observer pour des raisons d'accessibilité, de coût ou de sécurité. Les informations obtenues ont progressivement gagné en qualité et en richesse de contenu. Ainsi, les premières photographies en niveaux de gris ont évolué en images couleur de résolution toujours croissante. L'apport de la dimension temporelle avec l'apparition de séquences vidéo a ouvert le champ à de nouvelles applications ou au développement d'applications déjà existantes telles que la détection d'activité ou pistage d'entités. Aujourd'hui, des données vidéo de haute résolution sont maintenant plus communément accessibles. Elles sont exploitées dans le cadre d'applications civiles ou militaires : surveillance de zones étendues, contrôle de trafic routier, compréhension de champ de bataille ou encore missions de recherche et sauvetage. Nos travaux se placent dans ce contexte. Nous nous intéressons donc spécifiquement à des séquences vidéo aériennes, qui présentent des difficultés propres. Ces difficultés sont variées et se traduisent notamment par des problématiques de recalage et de parallaxe ainsi que des images bruitées ou floues, des segments saccadés. Elles ont orienté l'état de l'art présenté ainsi que les axes de travail développés.

Traitement des données : urgence et performances Si le contexte d'application le demande, ces séquences vidéo peuvent être analysées "en ligne", au fur et à mesure de leur acquisition. La réactivité apportée par ces traitements "en temps réel" est essentielle en présence de fortes contraintes temporelles, lors de missions de sauvetage par exemple. Toutefois, les contraintes temporelles limitent la complexité des traitements qui peuvent être réalisés. Cela est encore plus marqué lorsque les traitements doivent être effectués sur les architectures embarquées, par exemple si la transmission d'informations à une éventuelle base au sol pour traitement est délicate voire impossible (obstacles naturels tels que relief ou météo, nécessité d'une approche furtive).

Lorsque le contexte permet un relâchement des contraintes temporelles, l'exploitation des données peut être réalisée "hors ligne" si une analyse plus détaillée est nécessaire. Des traitements plus coûteux en temps voire en mémoire fournissent des résultats plus précis ou à un niveau d'interprétation plus élevé. De plus, les données vidéo peuvent être combinées avec d'autres sources d'information (radar, infrarouge ou cartes par exemple) au sein d'algorithmes de fusion adaptés.

Interprétation humaine : de qualité mais trop lente L'analyse manuelle de séquences vidéo permet de fournir directement une interprétation de haut niveau et de qualité. Dans le cadre d'une tâche de détection (objets ou structures fixes, ou entités mobiles) par exemple, le nombre de fausses alarmes sera particulièrement faible, et l'ensemble des objets sera détecté, hormis d'éventuels manquements de l'interprète dus à la fatigue ou perte de concentration. La visualisation humaine de vidéos dans leur intégralité est en effet lente et fastidieuse. Une telle analyse manuelle, hors ligne, se heurte ainsi à la constante progression du volume des données recueillies, permise par la démocratisation des capteurs vidéo de bonne qualité. Cela conduit à des choix nécessaires parmi l'ensemble des séquences et à la perte d'informations d'intérêt qui n'auront pu être relevées.

Aides automatiques pour l'interprétation Afin d'alléger la charge de travail des experts et minimiser cette perte, des outils automatiques d'aide à l'interprétation sont nécessaires. En effet, les segments véritablement informatifs dans des séquences vidéo sont généralement rares et isolés. La sélection automatique de ces segments pour une analyse approfondie par l'expert humain permet donc de reporter l'effort de celui-ci sur une interprétation de haut niveau plutôt que sur la visualisation fastidieuse de la vidéo et la détection des intervalles d'intérêt.

Axes de travail

Les objets spatio-temporels étudiés, des séquences vidéo aériennes, sont particulièrement riches et complexes. Le contenu statique est complété par un contenu dynamique, tous deux pouvant être d'origine naturelle ou humaine. La détection d'objets mobiles semble particulièrement intéressante en cas de faible activité car elle apporte alors un gain de temps précieux. Une visualisation intégrale de la vidéo est alors en effet inutile et le contenu statique peut être résumé de façon complémentaire par une unique mosaïque. La détection des objets mobiles peut également constituer une étape intermédiaire pour la recherche d'activités anormales. Mais la visualisation du contenu réel de la scène dépend des conditions de prise de vue et de leur évolution, par le biais de l'échelle, de la qualité image et des différents défauts qui dégradent cette qualité.

Conditions de prise de vue L'étude des conditions de prise de vue (CPDV) constitue un premier axe d'analyse de séquences vidéo. En effet, la facilité d'exploitation d'une séquence dépend de sa qualité. Plusieurs critères de qualité sont des conséquences directes des CPDV : le flou [149, 252, 93] et les défauts de compression (effets de bloc, entrelacement, repliement...) [255, 265]. Des segments vidéo trop flous ou bruités ne seront en effet pas exploitables : la résolution effective est alors trop faible pour les opérations classiques de détection, de pistage et *a fortiori* pour une interprétation de plus haut niveau. D'autres difficultés sont également liées aux conditions de prise de vue, tels qu'une illumination des images variable dans l'espace et le temps, des segments vidéo saccadés... L'importance de ces différents défauts dépend de l'application visée : recherche de précision pour une identification précise d'objet ou personne ; détection de mouvement pour un pistage voire une analyse de comportement ; établissement de cartes de profondeur pour appréhender le relief et les infrastructures...

Outre la qualité des images de la séquence, conséquence des CPDV à un instant donné, l'évolution de ces dernières apporte un éclairage complémentaire sur la séquence. De nombreuses études se sont attachées à l'estimation du mouvement apparent dominant ou mouvement global. Ce mouvement correspond généralement à la projection en deux dimensions, sur le plan image, du mouvement de la caméra relativement à la scène [25, 174, 200]. Il est approximé par une transformation globale plus ou moins complexe selon le modèle choisi : translation, rotation, homothétie, affinité, mouvement quadratique ou homographique... La connaissance de ce mouvement global rend possible un recalage des images de la séquence dans un repère commun. Cela permet de produire des mosaïques [144, 165, 231], de pister des objets mobiles détectés [278, 275] ou encore d'améliorer la résolution de la séquence par super-résolution [73, 253]. Le mouvement apparent ne permet toutefois pas de caractériser complètement le mouvement caméra en trois dimensions. Des ambiguïtés subsistent notamment entre rotation et translation ou entre zoom et translation suivant l'axe optique dans le cas d'une ouverture de champ réduite. Une estimation complète du mouvement caméra apporte des précisions supplémentaires et lève dans certains cas les ambiguïtés au coût d'une complexité accrue [159, 185, 94]. Avec ces informations complémentaires, il est possible de produire une carte de profondeur relative [160, 159] voire d'estimer l'incidence du capteur et la résolution image de la séquence pour une

meilleure compréhension de la scène observée.

Le mouvement caméra peut être utilisé comme donnée intermédiaire mais aussi fournir des informations en lui-même. Ainsi, Pan et Deschenes ont réalisé une classification du mouvement caméra estimé notamment grâce à la mesure du flou de mise au point [185]. Waizenegger, Feldmann et Schreer ont également utilisé l'estimation du mouvement caméra en trois dimensions pour annoter des séquences vidéo et pouvoir rechercher dans un ensemble de séquences celles qui comportent un type de mouvement particulier (par exemple zoom ou rotation autour de l'axe optique) [248]. Le nombre de classses de pouvement caméra proposé reste cependant limité.

Détection d'activité L'étude des CPDV d'une séquence donnée se limite à l'influence des conditions d'acquisition sur l'exploitation de cette séquence mais ne fournit que peu d'informations sur le contenu même de la séquence. Ce contenu peut être subdivisé en un contenu statique (environnement, présence de structures particulières telles que des bâtiments, de la végétation, des infrastructures, des réseaux) et un contenu dynamique (évolution temporelle des différentes entités de la scène). La détection d'activité, en particulier d'objets mobiles (principalement véhicules et piétons) est spécifique au contenu vidéo et constitue souvent un critère d'analyse important pour un interprète humain. Elle met en effet en relief des situations "anormales" par rapport à un fond autrement statique. Il semble donc pertinent de s'intéresser aux possibilités de détection automatique d'activité. Cette problématique a été grandement étudiée dans le cadre de caméras fixes.

Le cas de vidéos aériennes avec caméras embarquées complique fortement la détection. Car si le fond peut être estimé après recalage sur des cartes ou des modèles 3D, la présence de structures en trois dimensions (bâtiments, végétation, variations de terrain) introduit des effets de parallaxe brouillant la recherche d'objets mobiles. Un premier ensemble d'études étend des approches de soustraction de fond au cas de séquences vidéo.
Un second groupe d'études s'appuie sur des méthodes d'apprentissage spécifiques à la détection d'objets mobiles. La plupart reposent sur la détection et pistage de points d'intérêt en trois dimensions [44, 131] souvent peu fiables dans le cadre de vidéos aériennes souvent mal résolues.

L'introduction d'informations de contexte local explicité au travers de champs de Markov [241, 116], par la définition de zones de contexte plus globales [101] ou la détection de groupes d'objets d'apparence ou comportement similaires [269, 71] permet de filtrer la parallaxe. Toutefois, ces approches sont spécifiques à certaines situations (nécessité de trafic dense pour la détection de groupes d'objets semblables) ou reposent sur une phase d'apprentissage lourde exploitant d'immenses bases de données.

Démarche

Notre étude a pour but de faciliter l'indexation de séquences vidéo aériennes en fournissant des critères de navigation et de recherche, ceci afin de minimiser le temps nécessaire à l'analyse des séquences. Elle consiste à concevoir des méthodes et outils mathématiques et logiciels d'analyse de séquences vidéo aériennes dans le but d'extraire et de synthétiser des segments d'intérêt dans un cadre opérationnel. La figure 1 présente un squelette partiel de système d'aide à l'indexation de séquences vidéo en précisant les différents points qui seront détaillés dans ce document.

La complexité d'une séquence vidéo en tant qu'objet d'étude requiert de composer deux axes d'analyse : les conditions de prise de vue d'une part ; le contenu de la séquence lui-même d'autre part. De nombreuses études portent sur l'évaluation de la qualité image ainsi que sur le calcul du mouvement apparent. Nous utiliserons ces modalités comme

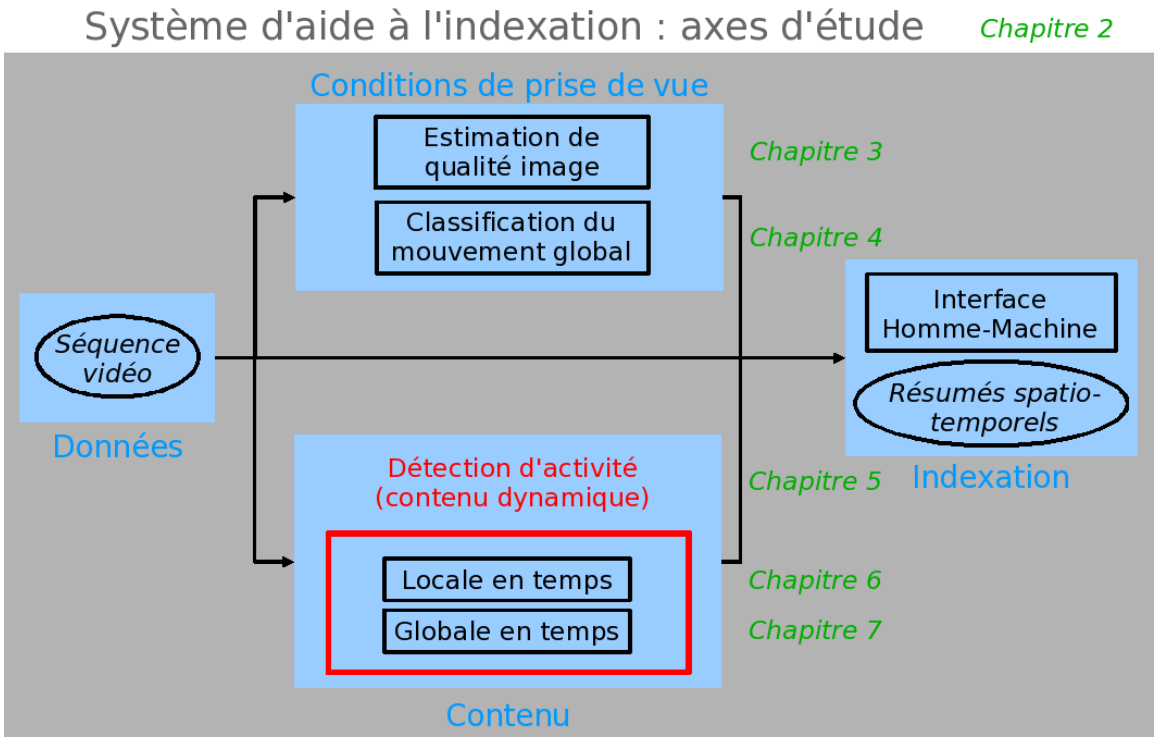


FIG. 1 – *Choix des axes d'étude dans le cadre de systèmes d'aide à l'indexation de séquences vidéo. Le chapitre 1 vise à introduire le schéma général d'un système d'aide à l'indexation. Deux types d'information peuvent être extraits des données (séquence vidéo) : conditions de prise de vue et contenu. Pour chacun de ces types, nous avons choisi des briques de travail développées dans les chapitres suivants. L'indexation tire parti des résumés spatio-temporels construits notamment à partir des briques précédentes et des interfaces homme-machine (IHM) effectuent la liaison avec l'interprète humain. Les ellipses représentent des objets (données d'origine ou données résultats tels que les résumés), les rectangles des briques d'analyse (du contenu ou des conditions de prise de vue). L'interface homme-machine (IHM) n'est pas une brique d'analyse à proprement parler, mais plutôt de communication à l'interprète et retour d'information.*

critères de sélection d'intervalles dits d'intérêt en nous appuyant sur ces études. La caractérisation automatique de segments ou intervalles vidéo comme éléments "d'intérêt" pour un interprète humain est délicate car elle repose sur des critères partiellement subjectifs. Ainsi, à partir de quel niveau de dégradation image peut-on considérer qu'une image ou un segment vidéo sont-ils inutilisables ou de peu d'intérêt ? Quelles sont les différentes stratégies de seuillage et de combinaison des mouvements apparents élémentaires (changement d'échelle, rotations, translations) afin d'obtenir des segments susceptibles de représenter un "intérêt" pour l'interprète ? Il faut ainsi pouvoir ajuster la sélection à la sensibilité de l'interprète, par le biais de seuils manuels par exemple.

Les conditions de prise de vue n'apportent cependant que peu d'informations sur le contenu même de la séquence analysée. C'est pourquoi l'analyse du contenu dynamique pertinent d'une séquence vidéo, à savoir la détection des objets mobiles, permet d'enrichir la compréhension de la séquence.

Notre démarche a consisté dans un premier temps à aborder les systèmes d'aide à l'indexation dans leur ensemble, en identifiant notamment plusieurs étapes dans l'élaboration d'un tel système (chapitre 1).

Nous nous sommes ensuite attachés à analyser les conditions de prise de vue pour caractériser différents critères de qualité image (chapitre 2) et d'obtenir une classification poussée du mouvement apparent (chapitre 3), voire d'en établir une interprétation partielle. Le but de cette approche est de présenter à l'interprète des critères de tri rapide. Ces critères peuvent être de l'ordre de l'intention. Les séquences vidéo aériennes étudiées sont en effet téléopérées et certains mouvements de caméra, traduits par des mouvements apparents

FIG. 2 – *Exemple d'un segment vidéo d'intérêt potentiel (recherche de détail), de haut en bas et de gauche à droite, fréquence* 25*Hz.* 1*ère ligne : zoom de durée 0.7 seconde (images* 548,556,564*).* 2*ème ligne : observation de la zone avec quelques secousses (images* 639,689 *et* 764 *soit respectivement* 3,5 *et* 8 *secondes après le zoom)*

FIG. 3 – *Exemple d'un segment vidéo de "zoom out" ou recherche de plus grande emprise. Les* 3 *images ont été prises à* 2 *secondes d'intervalle*

particuliers, relèvent d'intentions précises de la part de l'opérateur. Ainsi, par exemple, la présence d'un zoom puis d'un suivi de zone consécutif sont susceptibles de dénoter un intérêt de l'opérateur en cet endroit précis. La figure 2 fournit un exemple possible d'une telle situation (le zoom est visible même si le changement d'échelle reste limité). Dans une autre optique, une augmentation de focale ou "zoom out" (cf. figure 3) permet d'accroître le champ de vue afin d'appréhender plus facilement l'environnement de la scène jusque-là observée, au prix d'une diminution de l'échelle des objets et des détails de l'image.

Il peut également s'agir de filtrer les passages pour lesquels la qualité de l'image ou du mouvement est insuffisante : par exemple, un flou trop marqué ou un mouvement trop rapide pour que le contenu vidéo correspondant puisse être correctement interprété. La figure 4 illustre ces deux cas problématiques. Un résumé vidéo exploitant ces phénomènes peut représenter sous forme compacte l'ensemble des informations recueillies.

Dans un troisième temps, nous avons étudié la détection des segments vidéo comportant des objets mobiles (partie 2). Le chapitre 4 propose une présentation générale ainsi qu'un état de l'art des approches existantes, en différenciant approches locales et globales en temps. Dans le chapitre 5, nous développons une approche locale utilisant différentes informations de contexte. Le cas d'étude considéré, à savoir des vidéos aériennes avec caméra embarquée, présente en effet des difficultés spécifiques (bruit et mouvement du capteur, parallaxe). La combinaison d'informations de contexte à différents niveaux (voisinage spatial et temporel, présence de route) vise ainsi à filtrer les résultats d'une première

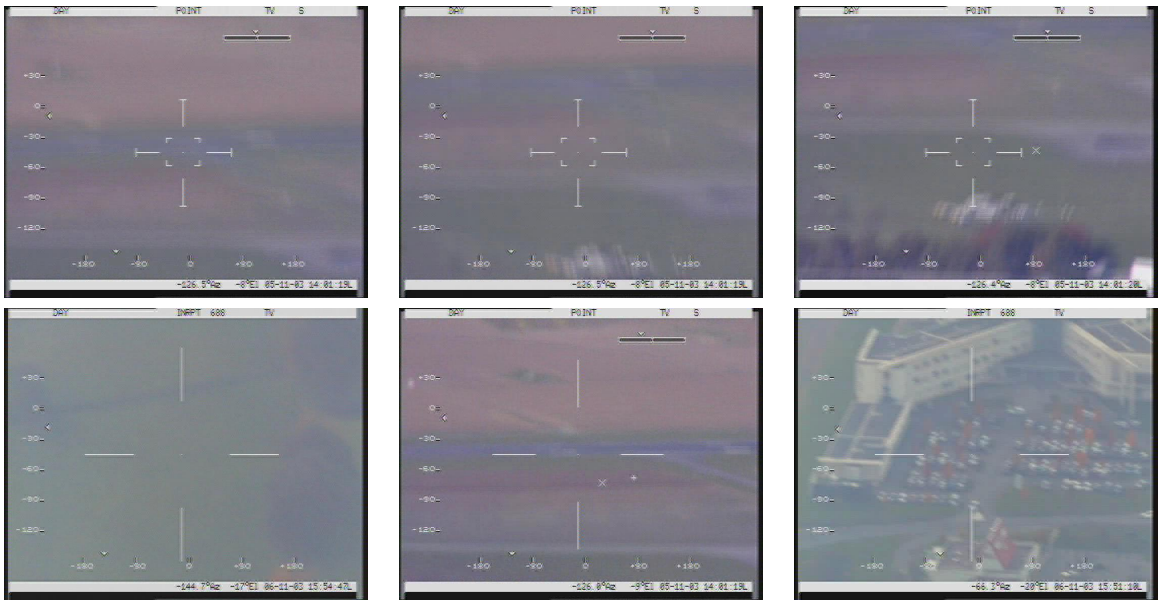


FIG. 4 – *Exemples de segments vidéo de faible qualité.* 1*ère ligne : mouvement rapide (*3 *images avec un échantillonnage à une fréquence de* 8*Hz).* 2*ème ligne :* 3 *exemples d'images floues.*

classification fondée sur des primitives d'apparence et de mouvement résiduel. Le chapitre 6 décrit plusieurs méthodes globales en temps, heuristiques ou tirées de l'état de l'art, et s'efforce de relever les différents avantages et difficultés inhérents à chaque méthode.

Le flot optique est un outil particulièrement utile dans le cadre de notre étude. L'annexe A développe le problème d'estimation du flot optique et s'efforce de décrire les diverses méthodes d'estimation utilisées dans la littérature ainsi que différents critères d'évaluation retenus à ce jour.

1 Systèmes d'aide à l'indexation et recherche de séquences vidéo : chaînes de traitement et solutions existantes

1.1 Introduction

Les séquences vidéo obtenues à partir de capteurs optiques ou infrarouges aéroportés, montés sur des hélicoptères ou des drones par exemple, sont des objets riches et complexes, de par leur dimension temporelle, l'étendue spatiale des régions observées et la diversité des dynamiques apparentes, qui peuvent traduire aussi bien les activités d'objets ou d'êtres animés que du bruit ou encore des variations de profondeur. Leur interprétation comporte de ce fait des difficultés importantes : il s'agit de traduire au sein de ces systèmes d'imagerie, des données physiques spatio-temporelles en des concepts abstraits communément maniés par l'être humain. La réduction de ces difficultés, souvent résumées sous l'appellation de "fossé sémantique", nécessite plusieurs phases de traitement de l'information, du "bas niveau" proche des données physiques à de plus "hauts niveaux" se rapprochant d'une interprétation sémantique.

Le cas de plateformes aéroportées introduit notamment des difficultés tels que le recalage géographique, la parallaxe, une résolution spatiale et temporelle inférieure ainsi qu'une qualité image variable. L'exploitation d'un contexte d'information géospatial pour la caractérisation d'activité (piétons, véhicules notamment) représente un autre défi. Les échelles d'intérêt spatiale et temporelle varient également suivant le scénario et les objets d'intérêt associés. Un scénario correspond ici à un ensemble d'activités particulières dont la nature et l'enchaînement (ou la réalisation simultanée) ont été définis au préalable. Enfin, la définition des scénarios elle-même peut être incomplète ou variable et spécifique à chaque séquence vidéo, et les éléments du scénario sont souvent inégalement répartis temporellement dans la séquence.

La figure 1.1 illustre une représentation possible de la structure générale d'un système d'aide à l'indexation de séquence vidéo.

La première colonne (données) est formée d'une part de la séquence à traiter, d'autre part de résultats éventuels d'apprentissage (modèles ou classifieurs).

Ces données permettent de calculer plusieurs modalités d'indexation (deuxième colonne) : les conditions de prise de vue ainsi que le contenu statique et dynamique.

La dernière colonne consiste à obtenir une interprétation sémantique de la séquence. Cela passe par l'utilisation d'interfaces : composition de requêtes simples ou composées, définition de scénarios. Une interface de navigation est nécessaire afin de permettre une visualisation de la séquence tout en intégrant l'affichage des résultats. Il est également souhaitable de disposer d'une interface d'annotations, à des fins de sauvegarde (métadonnées) voire d'enrichissement des données d'apprentissage. Les résultats de l'analyse, dont sont

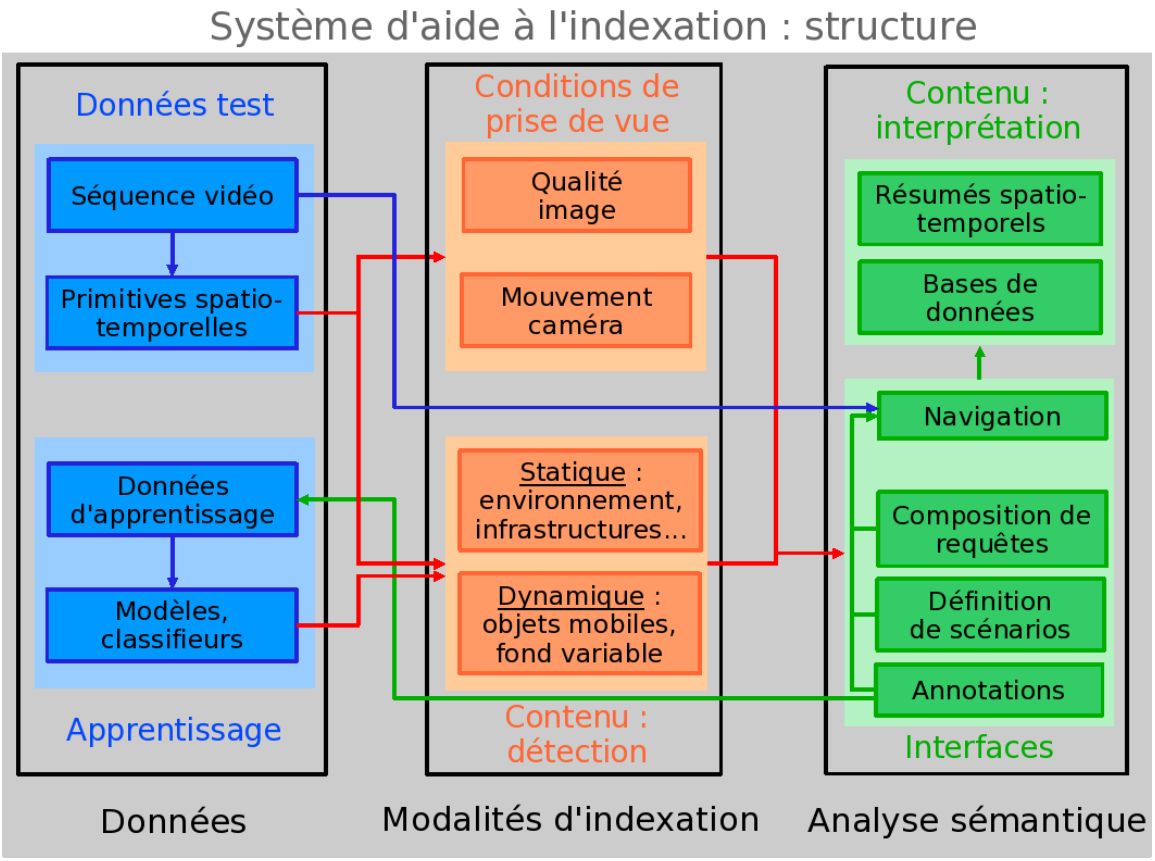


FIG. 1.1 – *Structure générale d'un système d'aide à l'indexation de séquence vidéo. Les données (première colonne) de la séquence et éventuellement résultats d'apprentissage sont traitées pour fournir plusieurs modalités d'indexation (deuxième colonne) : les conditions de prise de vue ainsi que le contenu statique et dynamique. La dernière colonne représente l'interprétation sémantique de la séquence. Différentes interfaces sont nécessaires : composition de requêtes, définition de scénarios, annotation et navigation. Les résultats de l'analyse, réponses aux requêtes et intégrant les définitions de scénarios, peuvent être stockées de diverses manières, comme bases de données multi-critères ou de manière plus graphique, comme résumés spatio-temporels.*

tirées les réponses aux requêtes, peuvent être stockées de diverses manières, comme bases de données multi-critères ou de manière plus graphique, comme résumés spatio-temporels.

L'exploitation de séquences vidéo aériennes nécessite donc plusieurs modèles spatio-temporels reliant capteur physique et scénarios sémantiques. Ces modèles peuvent être présentés en une suite d'étapes d'analyse sur des étendues spatiales et temporelles de plus en plus grandes.

Ainsi, une première étape concerne la détection d'événements ponctuels tels que la présence et localisation d'objets mobiles. Elle nécessite au préalable la stabilisation de la séquence, ou le recalage dans un repère unique, et se heurte notamment aux problèmes de parallaxe et estimation du mouvement global du capteur par rapport à la scène.

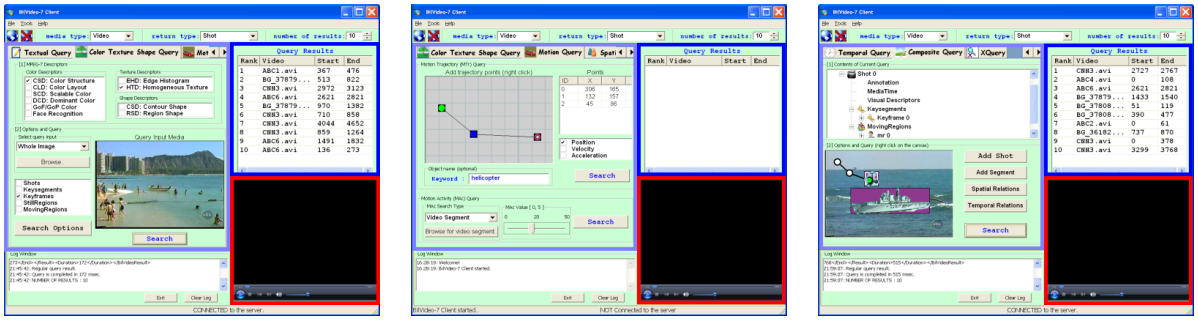
Une deuxième étape vise la caractérisation d'activités telles que des trajectoires de véhicules ou interactions entre plusieurs objets mobiles. Des requêtes, ou demandes dans un langage formalisé intégrant différentes variables de forme et apparence par exemple, sont alors possibles selon différents critères tels que vitesse, direction absolue ou relative dans un groupe d'objets.

Une troisième étape peut être effectuée si suffisamment de séquences vidéo couvrant une zone donnée sont disponibles. Dans ce cas, il est alors possible d'accumuler assez d'informations pour modéliser différentes activités voire des scénarios plus complexes, qu'il s'agisse d'effectuer des requêtes directes sur ces scénarios ou de détecter des situations anormales. Cette dernière étape est impraticable dans le cas de vidéos aériennes avec capteur mobile, prises par des drones par exemple dans le cadre de missions ponctuelles. Ce cadre ne permet en effet pas d'atteindre un nombre suffisamment représentatif de vidéos. Dans le cadre de séquences vidéo avec caméra fixe sur de longues périodes, la redondance permet en revanche une modélisation directe d'actions comme événéments (ou ensembles d'événements) spatio-temporels. Pour les vidéos aériennes, il devient donc nécessaire d'ajouter un niveau d'interprétation afin de pouvoir détecter d'éventuels comportements "anormaux" : ainsi, un algorithme automatique ne pourrait pas analyser aussi facilement les données de manière statistique et devrait extraire des critères de "normalité" tels par exemple des seuils sur la vitesse des objets mobiles. La modélisation automatique d'événements complexes par apprentissage est donc délicate voire impossible. Une autre possibilité consiste alors à apporter des *a priori* de connaissance.

La section 1.2 présente différents points liés à l'utilisation d'interfaces homme-machine (IHM), notamment sur leur évaluation. La troisième étape précédemment citée nécessite l'apport d'êtres humains au travers d'outils interactifs. Elle concerne les aspects "haut niveau" d'interprétation sémantique du système d'aide à l'indexation et fera l'objet de la section 1.3. Les éléments nécessaires pour cette étape sont produits à partir des données physiques d'origine, l'ensemble des images ou pixels de séquences vidéo. Cela nécessite une approche plus proche des données, "bas niveau", qui sera développée dans la section 1.4.

1.2 Approches d'interface homme-machine (IHM) existantes et évaluation

Les différents outils de recherche et environnements d'analyse présentés ci-dessus doivent être inscrits dans un processus d'interaction homme-machine efficace afin d'en tirer le meilleur parti. Les objectifs sont multiples : gain de temps principalement, en facilitant la navigation et recherche de segments ou événements d'intérêt ; aide à la représentation également, en permettant de visualiser sous une forme compacte un ensemble d'informations autrement difficile à regrouper et appréhender manuellement ; précision et développement enfin des modèles d'événement ou critères d'analyse au fur et à mesure des interactions, ce qui permet en retour d'améliorer la qualité des futures requêtes. [67] détaille les différentes applications du domaine.



(a) (b) (c)

FIG. 1.2 – *Exemple d'IHM de requête de BilVideo-7 avec module de navigation (rouge), d'affichage des résultats (bleu) et de définition de requête (bleu-gris) : (a) Requête sur l'apparence d'objets (couleur, texture, forme) (b) Requête sur le mouvement d'objets (position, vitesse, accélération) (c) Requête composite (apparence, mouvement, relations spatiales et temporelles)*

Plusieurs interfaces de manipulation de vidéos existent, avec des objectifs différents. Ainsi, BilVideo-7 [20] est un exemple de système d'indexation et de recherche de séquences vidéo compatible avec la norme MPEG-7. Il comprend notamment un module d'extraction de primitives vidéo et d'annotation. La figure 1.2 montre des exemples d'interface pour l'aspect requête à partir de différents types d'informations : apparence, mouvement ou requête composite mêlant informations d'apparence, de mouvement et de relations spatio-temporelles.

Le projet Advene (Annotations de vidéo numérique (Digital Video) Échangées sur le NEt) [10] vise à fournir un modèle et un format pour partager des annotations de documents audiovisuels numériques. Il apporte pour cela des outils permettant d'éditer et de visualiser des vidéos augmentées (avec inclusion de résultats de détection par exemple) générées à partir des informations liées aux annotations et aux documents audiovisuels. La figure 1.3 illustre le cas d'une telle vidéo (le cercle rouge et le rectangle vert dans le module de navigation représentent des résultats de détection).

L'outil industriel d'indexation et de recherche Ovire (Optimized Video Indexing & Retrieval Engine) [1] réalise une extraction automatique d'images-clefs (keyframes) dans un but de traitement, indexation et recherche automatique de séquences vidéo numériques. La figure 1.4 présente un exemple de l'interface utilisée pour la recherche et navigation.

1.2.1 Interface homme-machine

L'intérêt d'une telle interface réside notamment dans les retours possibles de l'utilisateur au système algorithmique (contrôle de pertinence). En effet, ce mécanisme permet d'optimiser les recherches en cours par rapport au bruit des données en tirant parti de la connaissance de l'utilisateur. Un état de l'art sur les méthodes de contrôle de pertinence pour des systèmes de recherche image à partir du contenu est présenté dans [111]. Dans le cadre d'environnements persistants (applications de surveillance ou analyse de la circulation routière) ou récurrents (événements sportifs présentant des caractéristiques communes par exemple), ce contrôle de pertinence peut être accumulé au cours du temps et contribuer à la définition de représentations sémantiques de haut niveau [97]. Ainsi, l'identification des requêtes les plus fréquentes et leur prise en compte dans les modèles physiques permet de fournir directement les résultats de ces requêtes sur de nouvelles données. De plus, des modèles de données spécifiques peuvent alors être appliqués aux requêtes selon le type de ces dernières afin d'améliorer les performances. La difficulté réside alors dans la fusion des modèles générateurs appliqués aux données et des modèles discriminatifs correspondant aux requêtes. L'analyse des interactions successives homme-machine peut également modéliser le comportement ou emplois-types de l'utilisateur [13] pour par exemple accé-

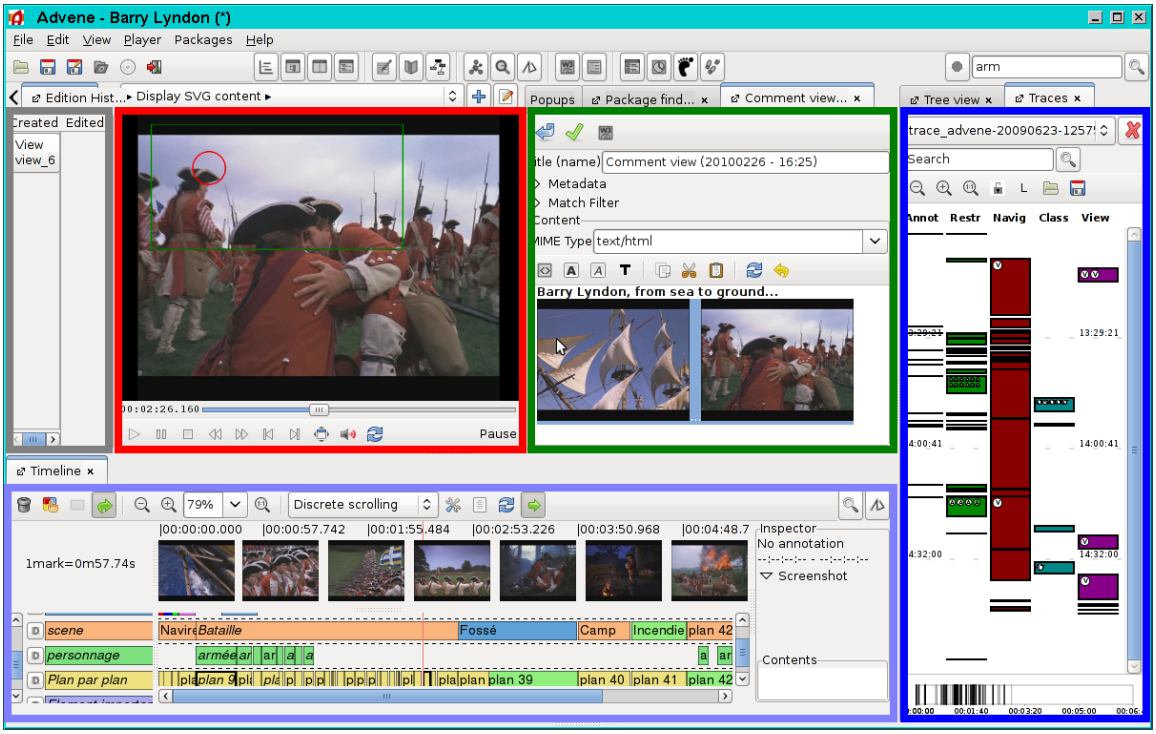


FIG. 1.3 – *Exemple d'IHM d'annotation et de navigation Advene : interface incluant des modules de navigation (rouge), de visualisation de la structure de la séquence (bleu), d'annotation et de segmentation selon plusieurs concepts sémantiques (bleu-gris) et de traitement de métadonnées (en vert)*

FIG. 1.4 – *Exemple d'IHM d'indexation et de requête Ovire : interface incluant des modules de visualisation, navigation et annotation (rouge), navigation par images-clefs (bleu-gris), métadonnées (vert) et de requêtes thématiques et transcription de paroles en texte (bleu)*

lérer de futures requêtes. Les outils de recherche ou d'indexation et les interfaces correspondantes peuvent être adaptés à des cadres spécifiques de vidéos sportives [125] ou de vidéo-surveillance [234, 62].

1.2.2 Évaluation

L'évaluation des interfaces dépend des tâches auxquelles celles-ci sont confrontées. Il importe donc de définir au préalable ces tâches, mais aussi leur cadre d'application : séquences de test, outils et données disponibles (par exemple des bases de données d'objets spécifiques pour des tâches d'identification ou reconnaissance). Trecvid [Tre] est un ensemble d'ateliers de conférences ayant débuté en 2003 et dédié à la recherche dans les domaines de la segmentation automatique, indexation et recherche fondée sur le contenu, sur des données vidéo numériques. Un ensemble important de données de test ainsi que des procédures d'évaluation uniformes sont fournis. Ainsi, en 2011, les travaux ont porté sur plusieurs domaines dont notamment l'indexation sémantique, la recherche d'objets connus, la détection d'événements dans un contexte de vidéo-surveillance ainsi que la recherche par l'exemple. Les résultats portent sur les méthodes de résumé [181] mais aussi sur l'aspect interactif homme - machine [52]. Les critères d'évaluation dépendent de la tâche considérée. Il s'agit de mesures de précision et de rappel à partir de notes portées par un panel de juges humains sur l'ensemble des résultats obtenus.

Données Les données ayant servi aux évaluations consistent en plusieurs centaines d'heures de séquences vidéo extraites de documentaires, bulletins d'information, actualités scientifiques, programmes éducatifs, et archives. La répartition entre ensembles de test et d'apprentissage a été réalisée aléatoirement sous la contrainte d'une répartition équilibrée sur l'ensemble des catégories vidéo.

Extraction de primitives sémantiques La faculté de détecter automatiquement la présence de primitives sémantiques telles que "en intérieur / extérieur", "personnes", "véhicules" est importante en tant que fondation pour la construction de requêtes et la recherche d'événements spécifiques. Elle peut également aider à la navigation au sein d'une séquence vidéo.

Détection d'événements en vidéo-surveillance Cette tâche a pour but la détection d'événements adaptés au domaine de la surveillance, des activités ou des interactions humaines, et ce à une grande échelle afin de traiter des volumes de données importants.

1.3 Approche "utilisateur" ou prise en compte des besoins : des activités aux scénarios

La définition de scénarios dépendant du contexte d'application, il est impossible de fournir un ensemble de scénarios exhaustif. La construction de modèles abstraits de scénarios (événements élémentaires pouvant être associés en parallèle ou séquentiellement selon des relations spatio-temporelles) dans un but d'interprétation de séquences vidéo doit ainsi rester générique et reposer sur un nombre limité d'hypothèses. L'intervention d'experts humains dans le processus d'interprétation est essentielle et les systèmes d'aide automatiques doivent donc intégrer des interfaces homme-machine efficaces. Cela nécessite tout d'abord de définir les différents outils de recherche qui seront utilisés, ainsi que l'environnement d'analyse permettant la recherche, la visualisation et le contrôle des données et des résultats.

1.3.1 Outils de recherche

Les outils de recherche permettent déjà un gain de temps en présentant directement les segments temporels et la localisation des éléments d'intérêt correspondant à la requête. Les performances de ces outils dépendent directement de la qualité des sorties intermédiaires : résultats de détection, de reconnaissance, d'identification, de pistage... L'intérêt des outils est variable suivant le contexte d'étude, une détection performante de structures fixes (bâtiments, végétation, ponts...) étant par exemple inutile dans un but de détection d'activité. Il peut s'agir de rechercher l'ensemble des présences d'objets ou structures d'un type donné, par exemple des piétons, véhicules à 2 ou 4 roues ou bâtiments. Cela suppose que ces différents éléments d'intérêt ont été auparavant détectés, reconnus et classifiés. La recherche peut également concerner des entités mobiles. L'estimation de la densité du trafic sur une portion du réseau routier en est un exemple [91]. Il faut alors toutefois disposer d'assez d'observations pour parvenir à une estimation pertinente : ce n'est pas le cas de vidéos obtenues à partir de drones parcourant d'importantes étendues spatiales.

Les requêtes peuvent être plus précises et faire intervenir les relations entre des pistes multiples, notamment les instants de départ et d'arrivée. Il s'agit de requêtes d'interaction telles que la détection de personnes ou objets ayant traversé le même espace (au même moment ou avec décalage temporel), qui permettent de caractériser le comportement d'ensembles d'objets. Ces requêtes dépendent du nombre d'objets et de l'extension temporelle de la recherche (par exemple le délai maximum autorisé entre les présences de différents objets au même endroit). Hoogs et al. utilisent ainsi des prédicats spatiotemporels permettant de définir différents types d'activités afin de détecter des activités de groupe impliquant plusieurs personnes [102]. Dans le cas d'une circulation dense, la résolution des requêtes peut être délicate en raison de l'incertitude des pistes aux croisements et sur les lieux d'arrêt ou de départ qui conduit à un nombre important de fausses alarmes. L'estimation du taux de fausses alarmes à partir des départs et des arrêts aide à juger la vraisemblance des rencontres [196].

Un autre ensemble de requêtes porte sur les pistes elles-mêmes, voire sur des ensembles coordonnées de pistes. La caractérisation de la trajectoire d'objets, tels qu'un virage en U ou une succession de virages en S, est en effet utile pour la détection de comportements particuliers voire inhabituels. Ainsi, la détection d'un virage en U à une intersection où ce type de manoeuvre est interdit, un *a priori* défini par exemple manuellement sur une région de l'image ou la mosaïque de référence, pourra renvoyer une alerte de comportement anormal. Ce type de requêtes nécessite des pistes complètes combinant de multiples pistes fragmentées et capables de résoudre des occultations partielles. L'association d'éléments de pistes en pistes complètes est délicate et peut être aidée par l'apport de contraintes, des modèles de trajectoires acceptables ou un contexte géospatial. Différents modèles d'association sont fondés sur la similarité d'apparence, et de vitesse ainsi que la cohérence entre les vitesses et les extrémités des fragments à recomposer.

La détection de groupes d'objets de comportements liés présente un cas particulier plus complexe de ce type de requêtes. Le lien peut être donné par la similarité des trajectoires, des lieux de départ ou d'arrivée, ou la conduite en formation particulière telle que "convoi" ou "poursuite" dans le cas de véhicules. Ces requêtes font intervenir de multiples objets sur plusieurs morceaux de pistes et leur coût calculatoire est conséquent, avec une explosion combinatoire en fonction du nombre de segments. L'emploi de modèles statistiques tels que les HMM (modèles de Markov cachés) amène à la détection d'activités physiquement plus précises [39] mais auxquelles il peut être difficile d'associer des événements classiques pour un être humain. Une représentation adaptée des entrées fournies aux modèles statistiques est alors utile pour obtenir des activités plus proches d'une interprétation humaine [51]. La modélisation de comportements sémantiques complexes dans un cadre aérien présentant un grand nombre d'objets mobiles est abordée dans [103]. Des méthodes

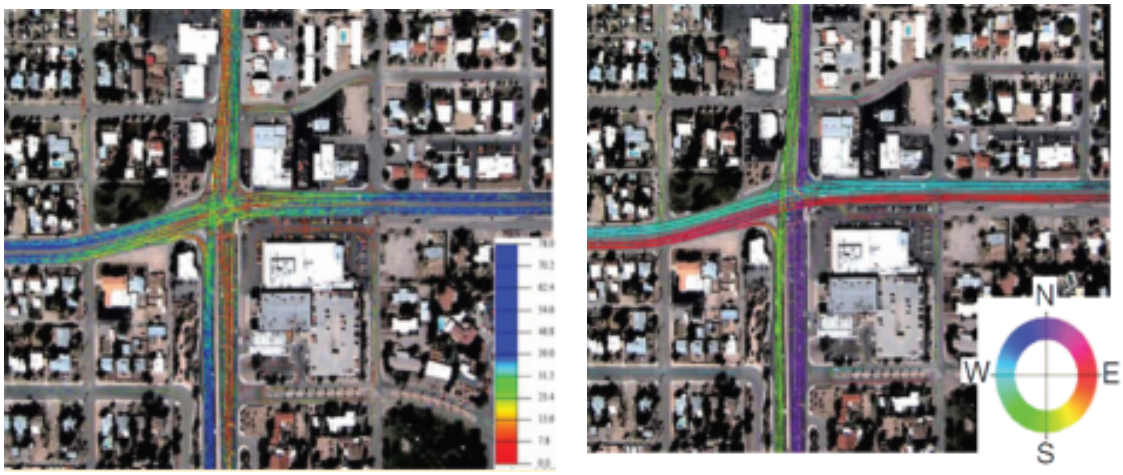
probabilistes sont associées à des *a priori* de connaissance sémantique de haut niveau et permettent de détecter des événements complexes rares par l'intermédiaire de primitives sémantiques dérivées de distances inter-objets. Le cas particulier de séquences vidéo d'événements sportifs, pour lesquelles il est possible de tirer parti de points de repère ou d'événements d'intérêt récurrents, a fait l'objet de plusieurs études [125, 126, 68].

Les requêtes peuvent être précisées grâce à une interaction avec l'utilisateur au travers de multiples choix et corrections. Les différents paramètres éventuels liés à la requête peuvent être ajustés, les différents critères peuvent être choisis dans un ordre défini, les résultats peuvent être validés ou infirmés voire classés comme fausses alarmes. Ces différents retours de l'utilisateur entraînent alors une amélioration de l'algorithme par ajustement des différents paramètres, par un apprentissage automatique ou une modification manuelle lorsque cela est pertinent. L'introduction de ces mécanismes de retour entraîne généralement une amélioration de la précision des résultats et / ou une diminution du temps de recherche [214]. Les requêtes ne représentent souvent qu'une faible partie des événements présents dans une séquence vidéo, ce qui déséquilibre le processus de retour et doit être pris en compte, par exemple dans le cadre d'un apprentissage actif semi-supervisé [109].

Un autre ensemble d'approches consiste à établir des modèles d'événements ou de comportements considérés comme "normaux". Les anomalies par rapport à ces modèles peuvent ensuite être extraites. La définition d'un tel contexte de "normalité" dépend de l'environnement et peut être inférée automatiquement par l'étude d'une base de données ou de la séquence vidéo elle-même [80, 33]. Dans [216], Saleemi et al. apprennent plus particulièrement les motifs de mouvement d'objets de manière non supervisée comme des fonctions de densité de probabilité de variables spatiotemporelles (nécessitant une détection préalable des objets mobiles). Sur des séquences vidéo avec caméra fixe, ces fonctions permettent de dresser ensuite une carte des trajectoires probables dans la scène. Xiang et Gong présentent une méthode [270] capable d'adapter de manière non supervisée des modèles de comportement au cours du temps. Il est cependant délicat d'interpréter les modèles obtenus de manière sémantique. La modélisation des propriétés spatiotemporelles d'objets par des relations floues permet après agrégation de détecter les points d'entrée et de sortie des objets dans la scène, puis les zones d'activité, et enfin de décrire sémantiquement les motifs de mouvement par rapport à ces différents points de repère [190].

1.3.2 Environnement d'analyse

Les outils de recherche doivent être intégrés à un environnement permettant non seulement de les appliquer, mais aussi de visualiser les différentes données d'origine (séquences vidéo) ou produites (objets ou bâtiments détectés, pistes d'objets, groupes d'objets de comportement semblable...) et de contrôler un ensemble accru de données en un temps réduit. Les méthodes de visualisation et d'intégration des données et des outils diffèrent suivant les critères retenus : compression temporelle et spatiale, cohérence temporelle des activités après transformation, compromis entre exhaustivité et concision, outils de recherche retenus... Les systèmes d'information géographiques (GIS) sont particulièrement utiles en environnement urbain car ils fournissent un riche contenu sémantique décrivant l'environnement spatial, des informations d'attribut telles que la nature (route, forêt, plan d'eau...) ou d'autres informations contextuelles (nombre d'habitants, densité, lignes de niveaux...) [271]. Des affichages visuels interactifs fondés sur ces systèmes peuvent être pertinents en regard de motifs spatiotemporels [120]. Dans le cadre de l'exploitation de séquences vidéo aériennes, il doit toutefois rester possible de naviguer de façon fluide d'une représentation sémantique de haut niveau au contenu vidéo original tout en passant par les représentations intermédiaires telles que les pistes d'objets.



(a) (b)

FIG. 1.5 – *Modèle de champ de circulation obtenu à partir d'une accumulation d'observations [213] : (a) vitesse et (b) orientation du mode d'histogramme*

La visualisation analytique [233] a pour objet de combiner la visualisation des données avec des tâches d'analyse et de fouille afin de faciliter la navigation et l'exploitation de données complexes. Les avantages affichés sont multiples : améliorer les capacités de mémoire par le biais de ressources visuelles ; réduire le temps de recherche par la représentation d'une grande quantité d'informations dans un espace réduit ; faciliter la reconnaissance de motifs en présentant spatialement les liens temporels ; contrôler plus aisément un nombre important d'événements potentiels ; apporter un moyen flexible de représentation de données, permettant d'explorer un espace de paramètres de façon interactive. Elle comprend notamment la construction de représentations compactes de séquences vidéos, élément essentiel de l'interface homme - machine.

1.3.3 Représentations compactes

Un premier ensemble de méthodes s'appuie sur la représentation de la séquence vidéo sous la forme de mosaïques après recalage dans un repère commun. Une mosaïque statique permet déjà de visualiser de manière compacte l'ensemble des régions observées lors de l'acquisition de la séquence [231]. Cela est particulièrement utile dans le cadre de vidéos aériennes avec d'importants déplacements : non seulement la compression temporelle est particulièrement efficace, mais le contexte spatial est également beaucoup plus clair et le recalage évite à l'utilisateur l'effort correspondant. En revanche, l'aspect dynamique y est absent. Une extension naturelle consiste à produire des "mosaïques dynamiques" en incorporant les objets segmentés dans l'espace de la mosaïque statique correspondant au fond [165]. Dans [213], Rosten et al. représentent les modes d'histogrammes de la vitesse et de l'orientation des véhicules superposées à une mosaïque statique, à partir des données observées en cette région (figure 1.5).

Les résumés vidéo constituent un deuxième ensemble de représentations. Il s'agit plutôt de compresser l'aspect dynamique tout en conservant le repère original : le ressenti est plus proche de la séquence originale, au prix d'un contexte spatial réduit et d'un référentiel non statique qui nécessite de la part de l'opérateur un effort de recalage qui peut nuire à la détection de structures ou événements d'intérêt. La méthode de compression peut être relativement simple, en extrayant une image de résumé toutes les N images, N étant plus ou moins élevé suivant le taux de compression souhaité. [263] propose une évaluation de différents taux par le biais de plusieurs mesures de performance notées par un ensemble d'utilisateurs et trouve une valeur de $N = 64$ acceptable. Une sélection plus recherchée des images servant au résumé est aussi envisageable [35, 121]. Toutefois les inconvénients

mentionnés ci-dessus subsistent. Il est également possible de recaler la séquence dans un repère de référence et de sélectionner ensuite les images pour la construction du résumé.

Une approche différente consiste à repérer les différents "tubes de mouvement" correspondant aux différentes activités [243]. Il faut ensuite les représenter de façon compacte. Cela nécessite un compromis entre le caractère exhaustif des activités, la durée de résumé ou taux de compression et la cohérence temporelle [198, 211]. D'autres méthodes encore représentent une vidéo par une collection d'images de résumé organisées de manière hiérarchique [113] ou comme un panorama de régions d'intérêt extraites des images-clefs de la séquence [251]. Plusieurs revues de l'état de l'art détaillent les différentes méthodes de segmentation de séquences vidéo [127, 137] lorsque ces dernières peuvent être découpées en prises de vue (films, documentaires, émissions sportives), ainsi que les méthodes de compression et formats de présentation de résumés [168, 16].

1.4 Approche bas niveau ou "bottom-up" : des pixels aux activités

Les premiers traitements réalisés sur les données vidéo brutes comprennent généralement une étape de recalage de la séquence afin d'obtenir des images de qualité supérieure dans un repère de référence commun ; une étape de détection d'objets suivie d'une étape de pistage. Suivant le contexte applicatif, des informations autres que la trajectoire (apparence ou classe d'objet par exemple) peuvent être également extraites. Les approches d'apprentissage automatique fournissent un ensemble de méthodes utiles pour la modélisation et la reconnaissance d'objets élémentaires voire d'activités, plus complexes.

Recalage L'une des premières opérations effectuées sur une séquence vidéo consiste à recaler les différentes images dans un repère commun. Cela permet de détecter et d'analyser plus simplement les objets mobiles et leurs trajectoires. Toutefois, des effets de parallaxe peuvent survenir et doivent être pris en compte. Ainsi, le calcul de cartes d'élévation par SfM (*Structure from Motion*) en ligne [53] ou ajustement de faisceaux [238] est utilisé pour compenser cet effet. De nombreuses méthodes de recalage existent [285, 119] et certaines permettent de produire simultanément des mosaïques [231, 144, 275]. L'évaluation de la qualité du recalage n'est pas évidente et ne présente pas de solution unique [167, 265]. Cette étape peut aussi permettre d'obtenir des images de résolution supérieure par super-résolution [50, 187, 253] afin de mieux distinguer les détails de structures ou d'objets d'intérêt.

Détection d'objets mobiles Cette dernière souffre de plusieurs difficultés, dont une résolution spatiale parfois faible et un bruit induit par les erreurs de recalage, voire une faible résolution temporelle. L'utilisation de détecteurs hybrides alliant soustraction de fond et détection de changement [195] permet de profiter des avantages respectifs de chaque méthode (précision de la localisation ; suppression de bruit de fond). Des primitives d'apparence peuvent résoudre les ambiguités dues à des arrêts, à des occultations, à des densités élevées d'objets autrement indétectables ou indifférenciables par les seules informations de mouvement. Ces primitives rendent aussi possible la classification d'objets tels que des véhicules au travers de modèles implicites [154] ou géométriques explicites [183]. La fusion de données provenant de moyens d'imagerie complémentaires tels que radar, thermique, infrarouge ou hyperspectrale améliore les résultats de détection et étend les contextes d'application possibles à des conditions météo ou d'illumination dégradées.

Pistage d'objets mobiles Il peut être réalisé après détection dans chaque image ou grâce à une approche de type *track-before-detect* [61]. Ces approches permettent d'améliorer la détection en effectuant un peu de pistage préalable, ce qui permet de filtrer les détections non cohérentes dans le temps. De plus, en cas de détection confirmée, elles fournissent des

pistes déjà initialisées. Le pistage après détection est fortement lié au choix du modèle d'objets [276]. Des améliorations ont été apportées sur des filtres connus tels que le filtre de Kalman [117] ou le filtre à particules [58].

Certaines séquences comportent un nombre important d'objets mobiles spatialement proches. Il est alors important de garantir l'intégrité des pistes de chaque objet. Différents algorithmes sont proposés comme le JPDAF (Joint Probabilistic Data Association Filter) [18] ou le MHT [227]. Le pistage à hypothèses multiples (MHT) [227] permet de gérer les pistes multiples en développant de façon exhaustive l'ensemble des associations possibles, y compris les hypothèses de nouvelles pistes. Il s'agit d'un algorithme de pistage optimal, mais qui se heurte à une explosion combinatoire du nombre d'hypothèses possibles. En pratique, un fenêtrage temporel est mis en place afin de limiter le nombre maximal d'hypothèses. D'autre part, pour faire face aux changements éventuels de dynamique ou d'aspect d'un objet, plusieurs modèles de transition d'état peuvent être choisis pour représenter la propagation des caractéristiques des objets pistés (position, vitesse, apparence), tels que les modèles à interactions multiples (IMM) [162]. Une telle approche est utilisée combinant information de couleur et de profondeur [169].

Si l'ensemble des détections et pistes d'objets est suffisamment dense dans une région donnée, il est également envisageable d'en dériver un modèle de carte géospatiale pouvant par la suite servir d'*a priori* [152]. La définition de modèles peut aussi être dynamique et capturer un ensemble de comportements "normaux" et par conséquent détecter des pistes s'écartant des modèles appris [230].

Il est aussi possible de ne pas détecter les objets image par image, mais travailler directement sur le volume spatio-temporel en caractérisant la géométrie locale d'un nuage de points correspondant à un mouvement résiduel significatif dans un espace espace-vitesse [279] ou en étudiant la saillance spatio-temporelle [204, 147]. Les modèles de transition d'état peuvent également inclure d'autres attributs que la position et la vitesse. Ces attributs complémentaires décrivent l'apparence de l'objet suivi (couleur, texture, coefficients dans une base d'ondelettes...) [245, 12] voire des détails plus abstraits tels qu'un type de comportement ou trajectoire.

Techniques d'apprentissage automatique Dans le contexte d'interprétation de séquences vidéo aériennes, les approches d'apprentissage automatique supervisé sont intéressantes. Elles peuvent viser à obtenir des résultats interprétables sur le plan sémantique, à partir de données d'apprentissage étiquetées. Ces données peuvent être simplement des instances d'objets spécifiques (bâtiments, véhicules, piétons, animaux...) dans un cadre de détection, de reconnaissance ou d'identification d'objet. Une interprétation sémantique humaine de la scène tirera alors parti des objets automatiquement détectés. Mais l'apprentissage automatique peut également concerner, sur les données complexes que sont les séquences vidéo, des actions voire des scénarios mêlant actions, points de repère et interactions entre différentes entités. La modélisation d'activités ou de scénarios représente en effet une problématique en plein essor qui s'attache particulièrement au cas de la modélisation et de la reconnaissance d'actions voire d'activités humaines. La logique spatio-temporelle [208] est alors notamment utilisée. Elle permet de décrire les interactions dans le temps et l'espace de différents objets et fournit ainsi des descripteurs pertinents pour un apprentissage automatique d'activités.

Les principales méthodes d'apprentissage supervisé regroupent notamment l'apprentissage par arbre de décision, les réseaux de neurones artificiels, les machines à vecteurs de support (SVM) ainsi que les réseaux bayésiens. Il faut rajouter à cela les différentes approches d'ensemble qui combinent ou agrègent des classifieurs souvent primitifs afin de fournir des classifieurs plus élaborés. Les plus connus sont le "bagging", le "boosting" et les "random forests" ou forêts aléatoires. L'annexe B détaille ces différentes méthodes.

1.5 Discussion

Les défis associés à l'indexation de séquences vidéo et la recherche d'événements d'intérêt au sein de ces mêmes séquences sont multiples. Les données "bas niveau" sont délicates à traiter : recalage et gestion de la parallaxe, détection et pistage d'objets. La modélisation des activités en des notions sémantiques "haut niveau" se heurte à différents aspects du "fossé sémantique" qui ne sont pas encore totalement résolus : diversité et complexité de scénarios, intégration de l'utilisateur humain au sein d'une interface homme - machine efficiente et capable de tirer parti de retours de pertinence, évaluation des différentes solutions algorithmiques et logicielles.

La détection automatique d'objets ou de structures spécifiques est rendue possible par le biais d'approches d'apprentissage automatique. Toutefois, la diversité des conditions de prise de vue et des environnements est particulièrement importante dans le cadre de séquences vidéo aériennes. D'importantes bases de données et / ou un orthorecalage sont donc nécessaires afin de pouvoir détecter les différentes entités à reconnaître. La variabilité intra-classe est de plus particulièrement importante par rapport à la variabilité inter-classes (par exemple, routes et bâtiments, véhicules et structures en trois dimensions ou différentes incrustations de marquage ou introduites par les instruments du système lors de l'enregistrement) selon les primitives d'apparence ou de mouvement. Il importe donc de choisir judicieusement un ensemble de primitives afin de séparer au mieux les classes dans l'espace correspondant.

Des approches existent pour la modélisation d'activités et la reconnaissance d'activités ou de scénarios suspects ou extraordinaires. Mais ces approches concernent principalement des séquences avec caméra fixe : la zone observée étant constante, la quantité d'informations disponibles est grandement supérieure, ainsi que la redondance d'actions usuelles à intégrer dans un modèle de comportement "normal". Il n'existe de plus pas de problème de recalage ni de parallaxe (les occultations restent présentes). L'inférence de modèles de comportement et par la suite la reconnaissance d'actions ou de schémas d'actions nouveaux ou extraordinaires reste difficile dans le cas de vidéos aériennes et nécessite l'intégration de modèles sémantiques logiques (interactions entre différentes entités définies par des relations de logique spatio-temporelle) fournis par un opérateur humain, voire appris sur des bases de données (avec caméra fixe) de contextes proches (environnements montagneux, urbain ou rural par exemple, associés à des ensembles d'acteurs, d'objets et de structures ainsi que des interactions spécifiques).

L'utilisateur humain doit ainsi pouvoir pallier l'absence complète ou partielle de contexte en occupant une place prépondérante au sein de la chaîne de traitement, par l'apport de modèles logiques de comportement voire une interprétation manuelle des interactions entre les différents acteurs et entités automatiquement détectés.

1.6 Conclusion

L'analyse de séquences vidéo nécessite d'allier des approches complémentaires d'analyse "bas niveau" des pixels et de leurs déplacements d'une part ; des approches de modélisation d'activités par l'apport de connaissances *a priori* dictées par l'expérience humaine voire apprises sur de grands ensembles de données d'autre part.

La complexité des objets spatiotemporels étudiés, à savoir des séquences vidéo aériennes, limite toutefois une modélisation autonome et automatique de comportements sémantiques significatifs pour un opérateur humain.

PARTIE 1
Conditions de prise de vue

Qualité image

2

Problématique Les séquences vidéo présentent généralement des disparités dans la qualité image des différents segments. Dans des cas extrêmes, certains segments sont inexploitables. Des critères de qualité image permettent non seulement de filtrer ces segments, mais aussi d'améliorer le détail et l'aspect visuel de résumés vidéo ou spatiaux.

La qualité du flux vidéo dépend notamment des choix de compression effectués. Il est possible d'analyser directement le flux vidéo compressé, mais les formats de compression sont variés. Nous avons choisi de considérer la qualité des différentes images de la séquence prises séparément afin de nous affranchir de ce paramètre. De plus, de nombreuses méthodes permettent de caractériser différents aspects de qualité, notamment le flou, d'une image donnée.

La cohérence et la continuité temporelle du flux apportent un avantage supplémentaire : il est en effet possible d'observer les variations de qualité au cours du temps et non pas simplement une qualité absolue pour chaque image. Un calcul pertinent de ces variations nécessite cependant en théorie un recalage préalable. En effet, la zone observée sur chaque image évolue au cours du temps. Ainsi, la comparaison au cours du temps de métriques de qualité agrégées sur l'ensemble de chaque image, sans recalage, serait faussée par la variation du contenu (plus grande importance de zones non texturées par exemple). Toutefois, dans le cadre de séquences vidéo modernes pour lesquelles la fréquence d'échantillonnage est élevée, le recouvrement entre des images consécutives est particulièrement important. Cela réduit l'impact du recalage sur la pertinence des variations observées.

Contributions L'ensemble des dégradations de qualité image et vidéo fait l'objet d'un grand nombre de publications, notamment le flou sous ses diverses formes. Il ne s'agit donc pas ici de produire un algorithme et des métriques dépassant l'état de l'art, mais plutôt d'évaluer sur plusieurs exemples d'images et à partir de plusieurs algorithmes, la viabilité d'un critère de flou comme paramètre d'indexation vidéo. Après un rappel d'éléments bibliographiques en section 2.1, nous présentons le modèle de flou et les approches d'estimation du flou de mise au point retenus en section 2.2. Les résultats obtenus sont commentés en section 2.3. L'utilisation d'un tel critère de qualité image dans un but d'indexation de séquence vidéo est ensuite explorée en 2.4.

2.1 Rappels bibliographiques

Une étude de différents modèles de vision ainsi que des métriques utilisées pour évaluer la qualité de flux vidéo numériques est réalisée par S. Winkler dans [265]. Nous avons choisi d'étudier plus précisément l'importance du flou de mise au point, susceptible de varier au cours de la séquence, à la différence des effets de compression tels que l'entrelacement, le repliement de spectre ou encore les effets de bloc. Des approches spécifiques existent pour l'estimation de certaines de ces dégradations tels que les effets de bloc [254] ou de repliement [207, 56]. Ces effets relèvent toutefois principalement de la compression et

sont supposés constants au sein d'une même séquence. Il apparaît donc moins intéressant de les évaluer dans un but d'indexation.

Si la compréhension générale d'une séquence vidéo reste possible en présence de flou, une analyse plus fine du contenu est en revanche plus délicate, voire impossible. Il sera en effet difficile de distinguer les détails d'objets ou structures de grandes dimensions, et de détecter ceux de dimensions plus faibles (de l'ordre de quelques pixels d'aire). Une approche de type super-résolution est certes possible, mais plus coûteuse et incertaine que le recours unique à des images nettes. Le tableau 2.1 classe un certain nombre d'approches d'estimation du flou selon plusieurs critères : calcul de métriques locales ou globales sur l'image, primitives utilisées, classification ou non du flou, variations auxquelles l'approche est robuste.

Le flou de mise au point correspond à la présence d'objets situés hors de la profondeur de champ de la caméra. La détection du flou constitue une première étape éventuellement suivie par une tentative de restauration d'une image non floue. Le flou peut concerner une partie ou l'intégralité de l'image, suivant la mise au point et les différences de profondeur de la scène. Après avoir déterminé l'extension spatiale du flou grâce à des machines à vecteurs de support (SVM), Hsu et Cheng classent le flou comme flou de bougé ou de mise au point [105] et estiment la fonction d'étalement de point (PSF) sur la région où le flou a été détecté. Liu et al. [149] classent également le flou d'une image par apprentissage sur plusieurs primitives telles que les couleurs et les contours ainsi que le spectre de l'image.

L'évaluation du flou passe souvent par une détection préalable des contours et une estimation de la largeur de l'étalement correspondant. Harasse et al. modélisent les profils de contours par des gaussiennes amplifiées avec fond constant et proposent deux mesures du flou en estimant la largeur de ces derniers [93]. Une autre approche consiste à modéliser la perception humaine des contours et du flou par l'intermédiaire de la théorie de l'espace d'échelle [145] en détectant les pics et extensions de réponse après deux étapes de filtrage par des filtres de dérivation gaussiens [82]. Wu et al. [267] estiment la PSF à partir de la fonction d'étalement de ligne (LSF) déduite à partir des informations de contour de l'image et l'utilisation d'une transformée de Radon locale. Wang et al. [252] utilisent également les informations de contours, mais en recherchant les contours les plus nets et en estimant l'étendue du flou correspondante, à partir de l'étalement de l'intensité orthogonalement aux contours.

Le flou peut aussi être mesuré sans utilisation des contours. Hu et de Haan modélisent le flou de mise au point comme un filtre passe-bas gaussien et estiment le noyau correspondant à partir de différences entre l'image originale et deux images obtenues en floutant celle-ci par deux noyaux gaussiens de taille différente [106]. Un autre type d'approche encore consiste à estimer le flou dans le domaine spectral, à partir de la transformée de Fourier du logarithme du spectre et y détecter les motifs périodiques correspondant au flou [268].

Plusieurs méthodes tentent d'améliorer la netteté de l'image après avoir estimé l'importance du flou. Dans [107], Hu et de Haan utilisent des filtres aux moindres carrés à partir de l'estimation du rayon local du flou et de la structure de l'image. Almeida et Almeida [7] emploient un *a priori* favorisant des contours nets et recréent progressivement la structure de l'image, des traits grossiers aux détails plus fins, en minimisant un critère des moindres carrés incluant une fonction de régularisation. Un état de l'art des algorithmes de détection de flou et de restauration d'images jusqu'en 2000 est proposé par Lagendijk et Biémond [133].

Approche	Localisation	Primitives	Classification	Robustesse
Hsu et Cheng [105]	local	gradients	oui	Contenu, luminosité, bruit
Liu et al. [149]	local	gradients, couleurs, spectre	oui	Contenu, luminosité
Harasse et al. [93]	local	gradients	non	Contenu, luminosité
Georgeson et al. [82]	local	gradients	non	Contenu, luminosité
Wu et al. [267]	local	gradients	non	Contenu, luminosité, bruit
Wang et al. [252]	global	gradients	non	Contenu, luminosité
Hu et de Haan [106]	global	noyaux gaussiens	non	Contenu, luminosité, bruit
Wu et al. [268]	local	spectre	non	Contenu, luminosité, bruit

TAB. 2.1 – *Classification des approches d'estimation de flou. La troisième colonne indique les primitives ou filtres utilisés dans les approches (principalement les gradients de l'image d'intensité). La quatrième colonne indique si l'approche permet de classifier le flou : bougé, mise au point. La dernière colonne indique les éventuelles variations d'image pour lesquelles la méthode d'estimation est robuste, autrement dit ne varie pas.*

2.2 Modèle de flou et choix des méthodes d'estimation

Le cas général d'un modèle de flou spatialement variable et en présence de bruit peut être modélisé par une convolution avec un noyau, à laquelle s'ajoute un bruit additif :

$$I(x,y) = B(x,y) * I_0(x,y) + N(x,y) \tag{2.1}$$

où x et y sont les coordonnées de l'image, $I_0(x,y)$ et $I(x,y)$ respectivement l'image d'origine et l'image dégradée au pixel (x,y), N l'image de bruit, souvent considéré blanc et de moyenne nulle, et B représente le noyau correspondant au flou et convolué avec l'image I_0. Le noyau B est une fonction d'étalement de point (PSF) et peut varier spatialement. Cette fonction représente également le résultat de la convolution du noyau avec une impulsion d'intensité. Il peut s'agir d'un flou de bougé, auquel cas la PSF présentera une structure linéaire et sa direction correspondra à la direction dans laquelle l'image sera floutée. Dans le cas d'un flou de mise au point, la PSF sera plate et étalée, chaque pixel étant parasité par les valeurs de l'ensemble des pixels voisins. Dans tous les cas, la PSF prend des valeurs positives et d'intégrale ou somme égale à 1 : il n'y a ni création ni déperdition d'énergie.

Nous nous intéressons plus particulièrement à l'estimation du flou de mise au point. Le noyau B peut être modélisé par une fonction gaussienne normalisée :

$$B(x,y,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), (x,y) \in \mathbb{Z}^2 \tag{2.2}$$

où σ représente le rayon du noyau à estimer. Il s'agit là d'un modèle simplifié ne tenant pas compte de la profondeur mais communément utilisé dans la littérature. Des modèles spatialement variables existent également mais nécessitent généralement une interaction

avec l'utilisateur ou des hypothèses sur le noyau [17, 140] ou encore une approche par régions reposant sur la fusion de plusieurs paramètres [149, 59]. Nous avons choisi de suivre l'algorithme décrit dans [106] : cette méthode ne nécessite pas la détection préalable de contours et présente des résultats corrects même en présence de contours proches. En revanche, elle ne permet pas de distinguer le flou de bougé du flou de mise au point, ni l'effet de flou produit par la présence d'ombres (image moins nette) du flou de mise au point. Toutefois, nous recherchons plus une indication de dégradation de qualité visuelle qu'une caractérisation extrêmement précise de cette dégradation.

2.2.1 Estimation de l'écart-type (rayon) du flou par noyaux gaussiens

Le principe de cet algorithme repose sur le re-floutage de l'image par deux noyaux gaussiens de rayons connus σ_a et σ_b avec $\sigma_b > \sigma_a$. Le ratio $\frac{I-I_a}{I_a-I_b}$ est maximal le long des contours et est directement lié à σ, σ_a et σ_b, et cela indépendamment de l'amplitude et de l'intensité du contour. Dans le cadre d'un contour idéal en une dimension d'amplitude A et d'offset B, le signal discret $f(x)$ s'écrit :

$$f(x) = \begin{cases} A+B, & x \geq 0 \\ B, & x < 0 \end{cases}, x \in \mathbb{Z}$$

où x décrit la position dans la direction orthogonale au contour. Le noyau 1D correspondant s'écrit : $g(n,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n^2}{2\sigma^2}\right), n \in \mathbb{Z}$ et on a $g(n,\sigma_1) * g(n,\sigma_2) = g(n, \sqrt{\sigma_1^2+\sigma_2^2})$. En notant respectivement $b = f * g$ le contour idéal f flouté par le noyau à estimer g, et , $b_a = b * g_a$ et $b_b = b * g_b$ les versions re-floutées de b par $g_a = g(n,\sigma_a)$ et $g_b = g(n,\sigma_b)$, on obtient le ratio $r = \frac{b-b_a}{b_a-b_b}$. Ce ratio est maximal en $x = 0$ et $x = -1$ et vaut

$$r(x)_{\max} = \frac{\frac{1}{\sigma} - \frac{1}{\sqrt{\sigma^2+\sigma_a^2}}}{\frac{1}{\sqrt{\sigma^2+\sigma_a^2}} - \frac{1}{\sqrt{\sigma^2+\sigma_b^2}}}$$

Dans le cas où $\sigma_a, \sigma_b >> \sigma$, les termes en racine peuvent être approchés pour obtenir :

$$\sigma \approx \frac{\sigma_a \cdot \sigma_b}{(\sigma_b - \sigma_a) \cdot r(x)_{\max} + \sigma_b}$$

Pour une image donnée I et un rayon σ de flou correspondant à estimer, σ peut donc être estimé localement à partir des extrema locaux de r calculé à partir de versions floutées $I_a(x,y) = I(x,y) * g(x,y,\sigma_a)$ et $I_b(x,y) = I(x,y) * g(x,y,\sigma_b)$. Les noyaux sont ici pris en deux dimensions, cela permet de ne pas avoir à détecter la direction des contours. En effet, chaque projection d'un noyau 2D en 1D est un noyau gaussien 1D et sur les points de contour, la réponse projetée r_{1D} sera maximale quelle que soit la direction de projection. Le maximum local de r en deux dimensions correspondra bien à la position du contour. En revanche, si σ est trop important, les versions de l'image refloutées I_a et I_b seront trop homogènes pour obtenir un résultat correct. Il reste toutefois possible de sortir de l'approximation en relâchant la condition $\sigma_a, \sigma_b >> \sigma$ et en tabulant r en fonction de σ afin de trouver la valeur de σ.

Nous avons également étudié l'énergie des hautes fréquences spatiales de l'image, au travers des gradients et du domaine spectral. La PSF n'est pas explicitement calculée, en revanche il est également possible d'obtenir des critères corrélés avec l'importance du flou.

2.2.2 Énergie des gradients d'intensité

Les gradients discrets de l'image sont obtenus par simple dérivation discrète. L'image de la norme des gradients donne une indication de l'importance du flou. En effet, à partir

d'une même image, un flou plus important sur les contours originaux se traduit par un étalement et donc une diminution plus ou moins importante des gradients associés. En effet, en reprenant l'exemple du contour 1D décrit précédemment,

$$b(x) = \begin{cases} \frac{A}{2}\left(1 + \sum_{n=-x}^{x} g(n,\sigma)\right) + B, & x \geq 0 \\ \frac{A}{2}\left(1 - \sum_{n=x+1}^{-x-1} g(n,\sigma)\right) + B, & x < 0 \end{cases}, x \in \mathbb{Z} \tag{2.3}$$

Le gradient discret $b'_d(x) = b(x+1) - b(x)$ se réécrit alors :

$$b'_d(x) = \begin{cases} \frac{A}{2}\left(g(x+1,\sigma) + g(-x-1,\sigma)\right) = A * g(x+1,\sigma), & x \geq 0 \\ \frac{A}{2}\left(g(-x-1,\sigma) + g(x+1,\sigma)\right) = A * g(x+1,\sigma), & x < -1 \\ \frac{A}{2}\left(2 * g(0,\sigma)\right) = A * g(x+1,\sigma), & x = -1 \end{cases}, x \in \mathbb{Z} \tag{2.4}$$

$b'_d(x)$ est donc une fonction strictement décroissante de σ et une autre formulation du gradient discret $b'_d(x) = \frac{b(x+1)-b(x-1)}{2} = A * \frac{g(x,\sigma)+g(x+1,\sigma)}{2}$ aboutit à la même conclusion. Les images des normes des gradients peuvent être ensuite réduites en une forme plus compacte pour une comparaison plus aisée entre images, telle qu'histogramme voire même moyenne ou médian des valeurs. Il est plus significatif de ne conserver que les maxima locaux de la norme des gradients, qui correspondent aux contours, et obtenir ensuite des métriques plus robustes que celles calculées sur l'ensemble de l'image.

2.2.3 Énergie dans le domaine fréquentiel

La transformée de Fourier de l'image apporte des informations sur la répartition fréquentielle de l'énergie. Ainsi, dans le cadre d'une image possédant des contours nets, l'énergie des hautes fréquences associées apparaîtra comme des points ou lignes dans l'espace spectral, correspondant respectivement aux coins et arêtes linéaires de l'image. La valeur à l'origine correspond à une fréquence spatiale nulle en x et y et traduit la moyenne de l'intensité sur l'image. La distance à cette origine des points ou lignes observés dans l'espace spectral sera d'autant plus grande en présence de textures plus fines.
La transformée de Fourier de 2.1 s'écrit :

$$\hat{I}(f_x, f_y) = \hat{B}(f_x, f_y)\hat{I}_0(f_x, f_y) + \hat{N}(f_x, f_y) \tag{2.5}$$

où f_x et f_y représentent les fréquences spatiales associées aux dimensions x et y, et $\hat{I}$, $\hat{B}$, $\hat{I}_0$ et $\hat{N}$ les transformées de Fourier respectives de l'image floue I, le noyau de bruit B, l'image nette I_0 et le bruit additif N. Dans le cadre où le noyau de flou est un noyau gaussien de rayon σ donné par 2.2, sa transformée de Fourier s'écrit :

$$\hat{B}(f_x, f_y) = \sqrt{\frac{1}{2\pi}} \exp\left(-\frac{\sigma^2}{2}(f_x^2 + f_y^2)\right) \tag{2.6}$$

Plus σ est grand, plus l'énergie de la transformée de Fourier du signal flou sera faible, à un couple de fréquences spatiales données, et ce d'autant plus que ces fréquences ou plus précisément leur norme euclidienne, sera élevée : l'énergie des hautes fréquences sera fortement réduite alors que les basses fréquences ne seront que peu perturbées. Il est donc envisageable de n'étudier que l'énergie correspondant aux hautes fréquences pour estimer l'importance du flou. Dans le cas d'une image discrète de taille $H \times L$, cela revient à ne pas prendre en compte dans l'espace fréquentiel discret centré de taille $H \times L$ les coefficients centraux (f_x, f_y) correspondant aux basses fréquences, avec $\sqrt{f_x^2 + f_y^2} < \alpha * \sqrt{(\frac{H}{2})^2 + (\frac{L}{2})^2)}$ où α représente le pourcentage de basses fréquences coupées :

$$E_{HF} = \sum_{f_x, f_y} \mathbb{1}_{\sqrt{f_x^2+f_y^2} > \alpha f_{max}} \times \hat{I}(f_x, f_y)$$

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

FIG. 2.1 – *Images classiques utilisées pour l'estimation du flou (a) Mire (b) "boat" (c) "mandrill" (d) "lena"*

Dans le cas d'un flou dû à un mouvement horizontal de la caméra, d'amplitude l, le noyau B peut s'écrire : $B(x,y) = \mathbb{1}_{|x|\leq\frac{l}{2}}$. La transformée de Fourier correspondante est $\hat{B}(f_x,f_y) = sinc(\pi f_x l)$. Comme la TF d'une rotation dans le domaine spatial correspond à une rotation de même angle dans le domaine fréquentiel, pour un mouvement d'amplitude l et d'angle θ, $\hat{B}(f_x,f_y) = sinc\left(\pi l(f_x * \cos(\theta) + f_y * \sin(\theta))\right)$. La TF de I présentera donc des "bandes" de direction orthogonale à la direction du mouvement. L'énergie correspondant aux fréquences hors de ces bandes (en dehors des pics du sinus cardinal) sera très faible. La norme des coefficients même de ces pics diminue fortement (en $\frac{1}{\|\mathbf{f}\|}$). L'énergie E_{HF} sera donc d'autant plus faible que l est grand, autrement dit que le flou de bougé est important.
L'utilisation de la transformée de Hough [59, 151] ou de Radon [69] pourrait permettre de retrouver la direction du mouvement si celui-ci est suffisamment important, avec des bandes marquées dans l'espace fréquentiel. Il serait également possible, par l'application de filtres de Gabor sur l'image floue I, d'estimer la direction d'un éventuel flou de bougé
Pour réduire la dimension des résultats, plusieurs métriques sont envisageables, la norme L^2 utilisée dans E_{HF}, ou des statistiques sur la répartition de l'énergie en fonction des fréquences, par histogrammes sur la norme L^2 de (f_x,f_y) par exemple.

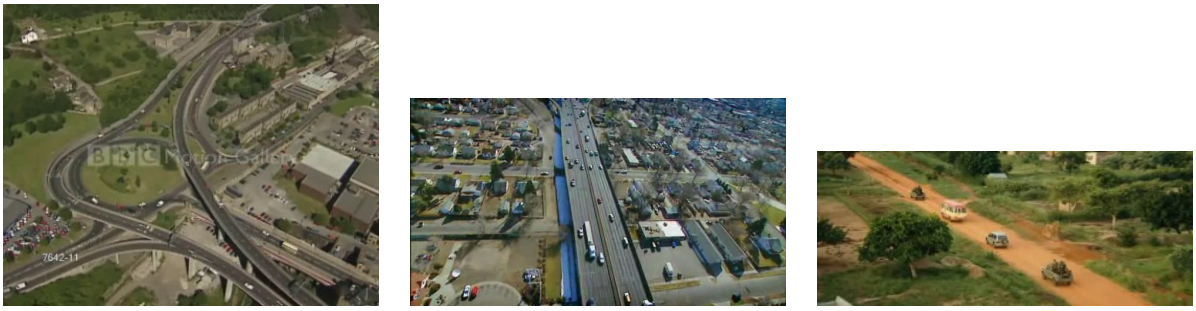
2.3 Evaluation des méthodes

Afin d'évaluer les méthodes d'estimation du flou, il est nécessaire de disposer de plusieurs images nettes (pour lequel le rayon de noyau de flou est faible voire nul). Il est alors possible d'analyser plusieurs cas de dégradation contrôlés en bruitant par exemple ces images par des flous de paramètres variables. Afin de couvrir au mieux l'ensemble des images réelles, nous avons choisi plusieurs images test synthétiques et réelles illustrées par les figures 2.1 et 2.2 :

- des mires de fréquences spatiales différentes,
- des images réelles classiques : "boat" avec des éléments fins (mâts, lettres formant le nom du bateau), "mandrill" présentant des textures fines (poils du visage) et plus grossières (marques nasales), "lena" où coexistent des régions plus ou moins floues,
- des images de vidéos aériennes présentant un contenu texturel de complexité variée et plus ou moins floues.

Différentes conditions de dégradation ont été appliquées à ces images et sont illustrées figure 2.3 :

- un bruit gaussien de rayon variable (cas "idéal" du bruit à estimer),



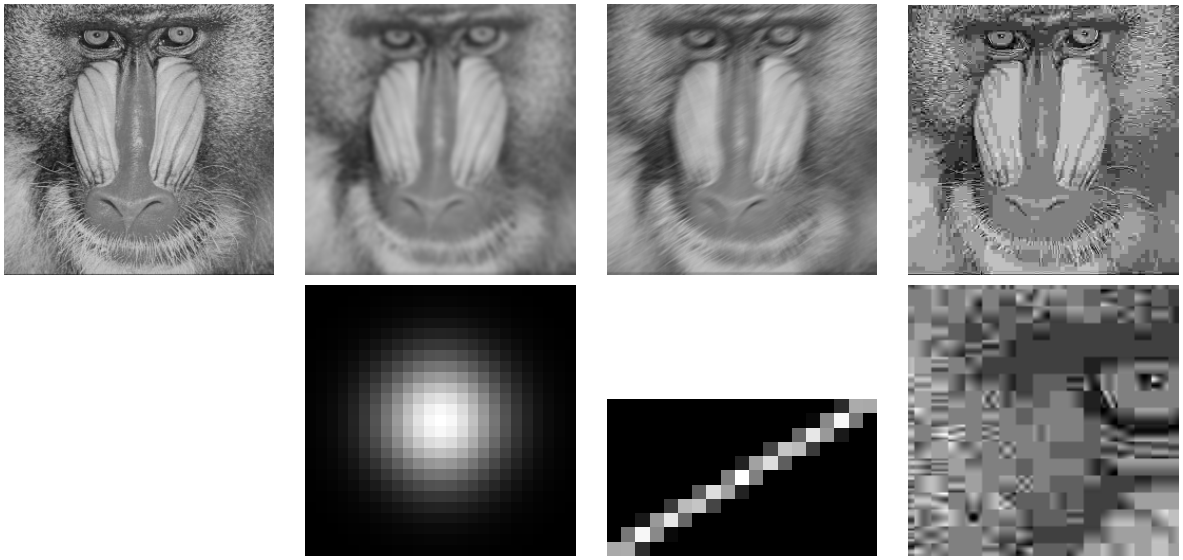
(a) (b) (c)

FIG. 2.2 – *Images aériennes utilisées pour l'estimation du flou (a) BBC (b) are we destroying Planet Earth (c) Blood Diamond*

- un flou de bougé de direction et d'intensité (nombre de pixels de déplacement) variables,
- une compression JPEG plus ou moins marquée en faisant varier le facteur de qualité. Ce facteur correspond à une quantification plus ou moins forte des coefficients obtenus après transformation en cosinus discrets (DCT) de l'image. La matrice de quantification divise pour chaque bloc 8×8 la matrice obtenue après DCT, les coefficients étant plus importants pour les fréquences élevées. Le facteur de qualité divise les valeurs de cette matrice de quantification : plus il est élevé, plus les valeurs de la matrice de quantification obtenue seront faibles et moins la quantification sera forte sur les valeurs obtenues après DCT (d'où une meilleure qualité image). Il s'agit d'une dégradation due à la compression du flux vidéo et non d'une dégradation due aux conditions de prise de vue. Cette dégradation devrait donc être constante pour une séquence donnée. En supposant l'absence de flou de bougé ou de mise au point (hypothèse toutefois peu réaliste en raison des variations de profondeur de la scène et des possibles mouvements brusques de caméra), ce défaut représente une dégradation parfois sévère de la qualité image. Si des approches spécifiques existent pour l'estimation de ce défaut [256, 246], il reste intéressant d'étudier la capacité les différentes approches détaillées ci-après à capturer cette dégradation qui pourraient être utilisées comme critères d'appréciation générale de qualité sans avoir recours aux approches spécialisées.

Dans le cas de la mire, nous avons fait varier au sein de la même image le rayon du bruit gaussien. Il s'agit non plus d'un flou isotrope, mais d'un flou unidimensionnel gaussien appliqué orthogonalement au contour (dans la direction verticale pour des bandes horizontales). Cela permet pour les algorithmes fournissant des résultats localisés, tels celui fondé sur la différence de gaussiennes, d'évaluer par régions l'exactitude de l'estimation du flou et de préciser les difficultés rencontrées (sur- ou sous-estimation du rayon du noyau, mauvaise localisation).

Enfin, suivant les méthodes, des métriques sont nécessaires afin de pouvoir comparer les résultats obtenus sur les images après différentes dégradations. Il peut s'agir simplement de la moyenne ou médian des cartes de valeurs, ou de statistiques plus robustes dérivées des histogrammes des valeurs ou calculées sur un sous-ensemble significatif (par exemple, l'espace correspondant aux hautes fréquences spatiales dans le domaine fréquentiel). Pour l'estimation par les rayons, une moyenne sur l'ensemble de l'image traduit de manière simple l'amplitude de la déformation. Pour l'estimation fréquentielle, le pourcentage de l'énergie conservée par rapport à l'image originale est calculé sur plusieurs parties du domaine spectral constituées des $X\%$ plus hautes fréquences. Pour l'énergie des gradients, nous prenons le pourcentage de l'énergie conservée par rapport à l'image originale sur les gradients "significatifs", de norme supérieure à plusieurs seuils (0 pour l'intégralité des gradients).



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

FIG. 2.3 – *Effet de différentes dégradations de qualité image.* 1*ère ligne : images (a) sans dégradation (b) avec flou gaussien, rayon* 4 *(c) avec flou de bougé, de* 21 *pixels, avec un angle de* 30 *degrés (d) avec compression jpeg avec un facteur de qualité* 3. 2*ème ligne : (b) noyau isotrope gaussien, de rayon* 4 *(c) noyau pour un déplacement d'angle* 30 *degrés et de taille* 21 *pixels (d) détail de l'image avec compression jpeg d'un facteur de qualité* 3.

Résultats L'objectif de cette évaluation est d'étudier la viabilité des méthodes comme critères discriminants de qualité. Dans l'idéal, la nature et l'intensité des dégradations présentes dans les images observées seraient automatiquement extraites et évaluées de façon à filtrer les segments temporels et / ou régions des images non exploitables. Nous partons de l'hypothèse selon laquelle les défauts de compression ne varient pas au cours d'une séquence vidéo. Le flou reste en revanche une dégradation d'intensité variable selon la mise au point ou des mouvements caméra trop brusques, c'est ce point que nous allons développer.

2.3.1 Flou gaussien

Nous avons choisi des noyaux gaussiens de rayons croissants entre 0.5 et 4 par pas de 0.5. Le noyau de rayon 0.5 représente une dégradation légère et presque invisible. Le noyau de rayon 4 en revanche correspond à une dégradation sévère, mais les structures générales de l'image restent visibles. Des noyaux de rayons supérieurs fournissent des images très plates ou homogènes, l'estimation du flou perd en précision. En effet, les deux images obtenues après application de ces noyaux sont chacune homogènes et leur différence minime, d'où un risque de divergence.

La figure 2.4 présente quelques résultats des différents algorithmes d'évaluation décrits ci-dessus. Les cartes de la deuxième ligne (colonnes (a) à (c)) représentent les rayons de flou gaussiens estimés sur des blocs 4×4 de la façon suivante : pour chacun de ces blocs, la valeur minimale est retenue (elle correspond au gradient le plus marqué du bloc et donc l'estimation la plus fiable du rayon). Les images de la troisième ligne sont les normes des coefficients complexes de la transformée de Fourier des images correspondantes (originale et floutées), ou plus précisément le logarithme de ces normes, pour une meilleure visualisation. La dernière ligne montre les normes des gradients des images, visualisés avec rehaussement de contraste. Pour chaque ligne, l'échelle des valeurs est identique pour une comparaison visuelle non biaisée. La dernière colonne montre l'évolution de métriques d'agrégation (moyenne, énergie) avec l'augmentation du rayon. Pour les gradients et le domaine de Fourier, ces métriques sont normalisées relativement à la valeur obtenue sur l'image originale. Pour l'estimation de rayon, une telle normalisation n'est pas nécessaire

puisqu'il suffit de comparer les rayons.

Les trois méthodes, à savoir estimation du rayon par refloutages gaussiens, étude de l'énergie spectrale et énergie des gradients, montrent bien la progression dans la dégradation de qualité par rapport à l'image originale. Toutefois, il apparaît que certains choix de paramètres sont plus adaptés.

- Ainsi, un seuillage trop important des gradients (quatrième ligne, colonne (d)) aplatit totalement la courbe des énergies de gradients (prises relativement à l'énergie de l'image originale pour le même seuil) pour des rayons élevés. L'énergie obtenue en conservant l'ensemble des gradients est plus robuste mais tend à se stabiliser pour un rayon supérieur à 3.
- Quelle que soit la portion du domaine de Fourier retenue, l'intégralité ou les plus hautes fréquences seulement, l'énergie correspondante, relative à l'énergie de l'image originale pour les mêmes fréquences, se stabilise au-delà d'un rayon égal à 3. Là encore, le choix de l'ensemble du domaine (ici de Fourier) pour le calcul de l'énergie semble plus discriminant.
- Ce comportement est logique, qu'il s'agisse des gradients ou de l'énergie dans le domaine spectral. En effet, les gradients les plus forts correspondant aux contours nets ou les hautes fréquences associées disparaissent prioritairement, dès l'application d'un flou modéré. Les énergies correspondantes ne diminuent ensuite que peu à l'application d'un flou plus intense.

L'algorithme de calcul explicite du rayon fournit une indication plus précise du rayon de flou gaussien et ce pour différents choix des paramètres σ_a et σ_b. En revanche, nous pouvons noter plusieurs problèmes mineurs dans l'évaluation du rayon.

- Tout d'abord, pour un flou très léger (rayon 0.5), le rayon estimé reste au même niveau que celui estimé pour l'image originale, voire diminue légèrement. Cela peut être dû à plusieurs raisons : d'une part, le paramètre global extrait de la carte des rayons est une simple moyenne. D'autres choix tels que médian, ou moyenne et médian des rayons en-dessous de divers seuils, n'apportent pas d'amélioration. Il peut être souhaitable de ne considérer les rayons estimés qu'au voisinage des contours, où l'estimation est la plus fiable.
  D'autre part, pour de faibles valeurs, l'estimation peut être moins précise, il s'agit en effet d'images en valeurs quantifiées pour lesquelles l'erreur relative de quantification pour un flou modéré sera non négligeable.
- Pour un flou important en revanche, les valeurs de rayon obtenues sont plus faibles que les valeurs réelles, les courbes de couleur correspondant à différents choix de paramètres σ_a et σ_b s'écartant de la valeur "théorique" obtenue en considérant pour l'image d'origine un flou gaussien de rayon σ_{orig} estimé par l'algorithme. Par exemple, pour le choix de paramètres $\sigma_a = 5$ et $\sigma_b = 7$, la courbe extrapolée noire creuse l'écart avec la courbe bleue obtenue en pratique. L'hypothèse d'un flou uniforme gaussien de rayon σ_{orig} n'est pas forcément exacte, mais surtout l'approximation $\sigma_a, \sigma_b >> \sigma$ n'est plus valable dès que le rayon du noyau de flou gaussien appliqué à l'image originale (et donc le rayon correspondant à la composition du flou de l'image originale et du flou rajouté) prend des valeurs supérieures à 1 ou 1.5. Un choix de σ_a et σ_b plus importants encore n'est pas souhaitable car les versions re-floutées I_a et I_b perdront trop de détails. Il reste possible d'obtenir des valeurs moins sous-estimées de σ en tabulant r en fonction de σ afin de trouver la valeur de σ (cf. 2.2) au prix d'un coût calculatoire plus élevé. Toutefois, à des fins de comparaison ou d'estimation approximative du rayon, l'approximation donne des résultats

satisfaisants.

La figure 2.7 résume les résultats obtenus sur différentes images d'origine en conservant pour chaque méthode les paramètres les plus robustes : $\sigma_a = 5, \sigma_b = 7$ pour l'estimation du rayon, l'intégralité du domaine spectral et des gradients pour les calculs des énergies respectives.

2.3.2 Flou de bougé

Nous avons fait varier deux paramètres, l'angle et la longueur du déplacement. Les angles varient de 0° (déplacement horizontal) à 90° (déplacement vertical), et la longueur de 5 à 25 par pas de 5. La figure 2.5 montre des résultats pour un mouvement de 10 et 25 pixels, avec des angles respectifs de 60° et 45°. La colonne (d) présente à angle constant (ici 0°) l'évolution des métriques avec la longueur du déplacement.

L'étude des images appelle plusieurs remarques.

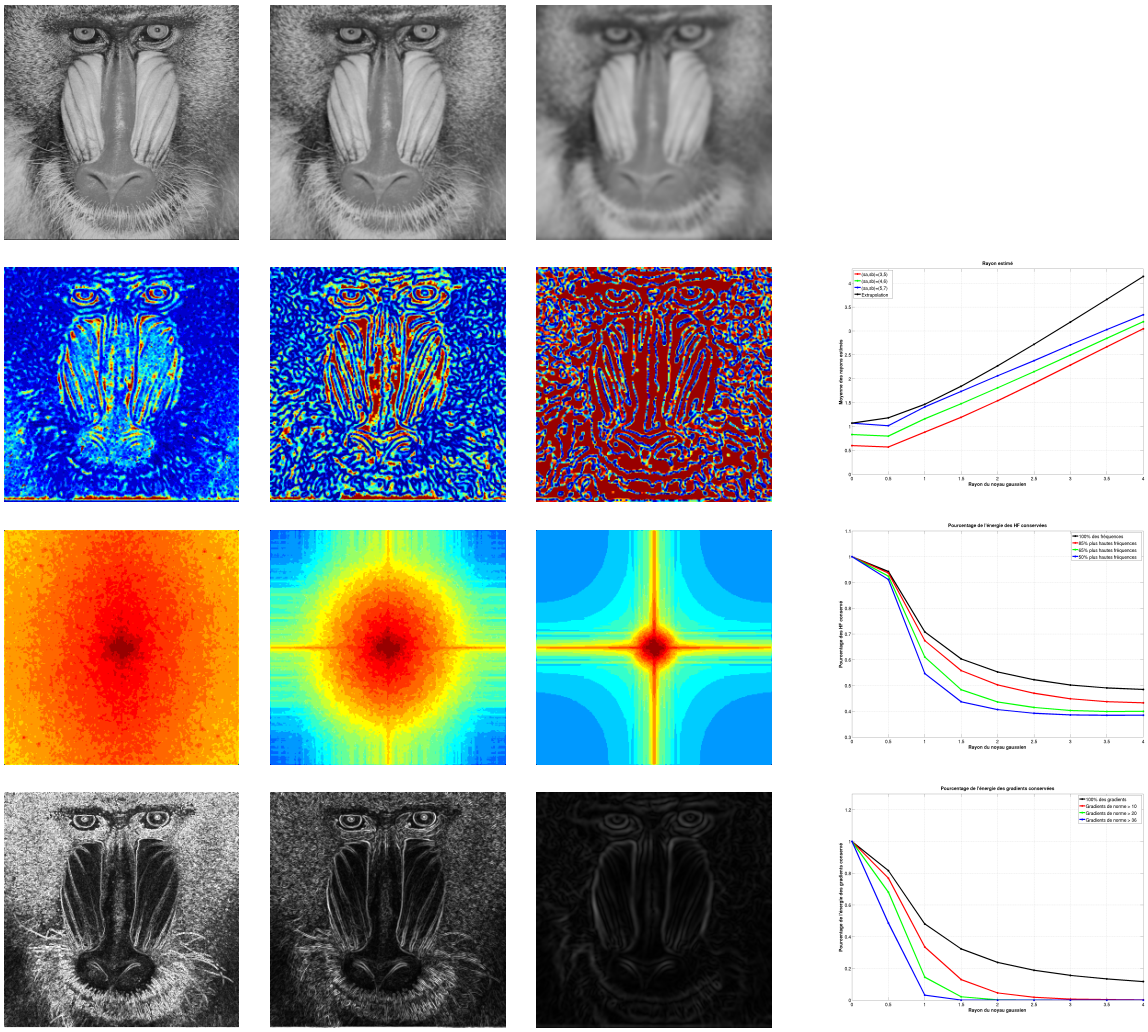
- Les images originales et après flou de bougé montrent que même un déplacement important (25 pixels) n'entraîne pas une impression de flou particulièrement importante. Cela dépend notamment du sens du mouvement par rapport à la direction des contours de l'image. En effet, ici le déplacement est horizontal, les structures horizontales sont ainsi mieux conservées. Cette impression se retrouve dans les cartes de rayons estimés : le flou gaussien isotrope n'augmente que peu avec la longueur du déplacement. En revanche, les images des gradients montrent que ces derniers disparaissent progressivement.

- La diminution de l'énergie ou norme L^2 des gradients est beaucoup plus marquée que l'augmentation du rayon de flou gaussien estimé. Cela peut s'expliquer car les gradients des structures fines du "mandrill" (i.e., fourrure et poilds du visage) sont très sensibles à un mouvement ou flou et une partie importante de ces gradients disparaît. La conservation des détails de même direction que le mouvement associé au mode de calcul par blocs du rayon, limite l'influence du mouvement sur l'estimation du rayon isotrope : le flou de bougé conserve une partie des détails, ce qui limite l'impression de flou isotrope de type "mise au point".

- Le graphe d'évolution ainsi que les images correspondant à l'analyse fréquentielle font également ressortir plusieurs points. Premièrement, l'évolution de la dégradation en fonction de la longueur du déplacement ne dépend quasiment pas des fréquences considérées : la fonction en sinus cardinal dans une direction particulière détruit beaucoup moins de hautes fréquences que la fonction gaussienne d'un flou de mise au point. La diminution de la métrique résulte de l'augmentation du nombre de bandes et diminution de la largeur des ces bandes au fur et à mesure que la longueur du déplacement augmente : la taille de l'espace fréquentiel discret étant fixée, la place relative des premières bandes diminue lorsque la longueur du déplacement, et par conséquent la fréquence du sinus cardinal, augmente.
  L'observation du spectre laisse prévoir une possible estimation du flou de bougé, à partir d'une transformée de Hough ou de Radon ou des bancs de filtres directionnels appliqués aux (normes des) coefficients de Fourier, qui ferait ressortir la direction des bandes, orthogonale à la direction du déplacement.

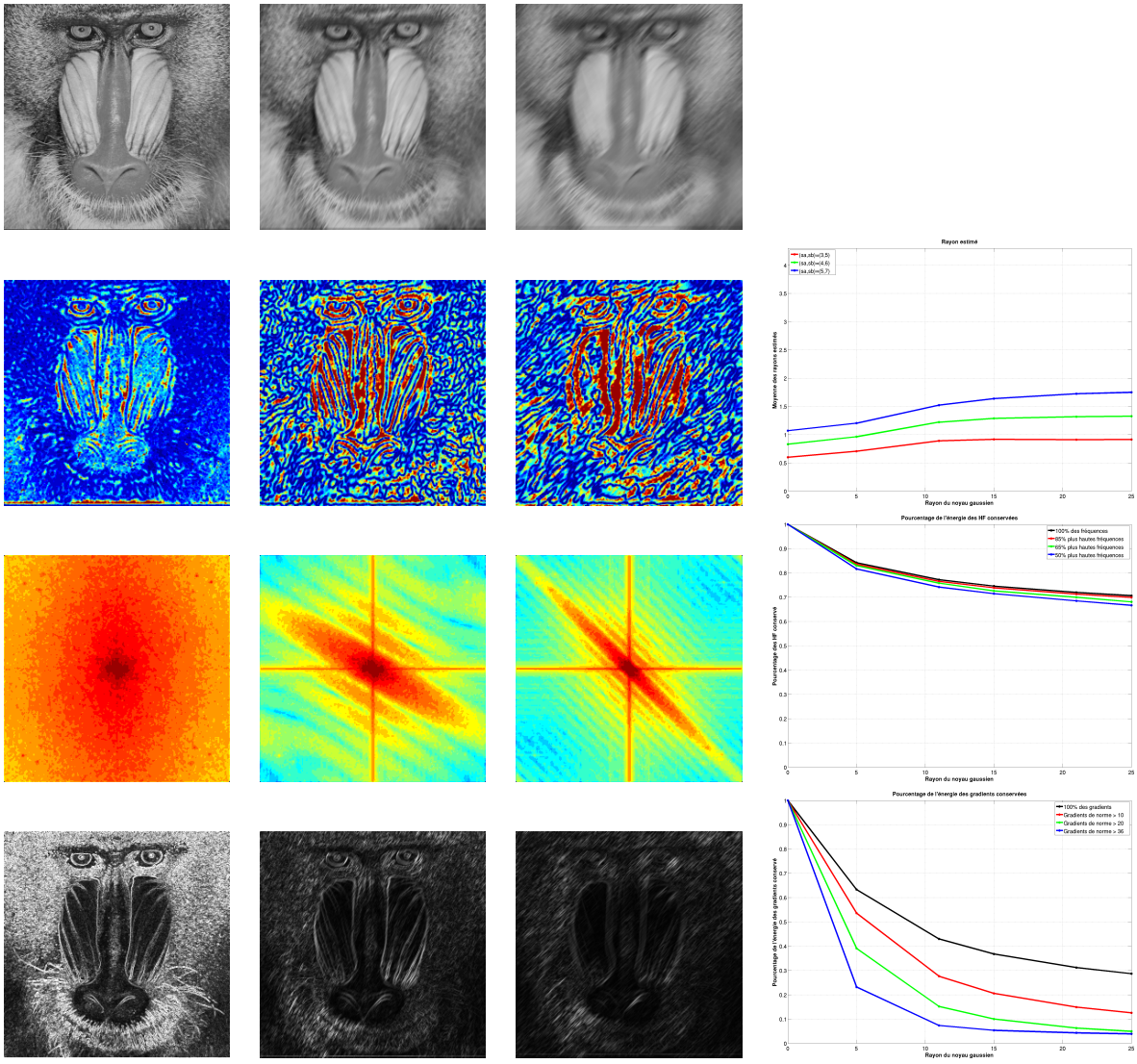
2.3.3 Bruit de compression jpeg

La compression jpeg suit plusieurs étapes, découpages en blocs, transformation de couleurs, sous-échantillonnage, transformée en cosinus discrets (DCT), quantification et



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

FIG. 2.4 – *Image "mandrill", flous gaussiens et résultats (a) Originale (b) Rayon* 1 *(c) Rayon* 4*. (d) Métriques de résumé :*
*pour l'estimation par les rayons (*2*ème ligne), avec respectivement* (σ_a, σ_b) *choisis à* $(3,5)$,$(4,6)$,$(5,7)$*(rouge, vert et bleu). La courbe noire représente l'extrapolation théorique à partir de la valeur estimée du rayon pour l'image non floutée avec* $(\sigma_a, \sigma_b) = (5,7)$*. Le rayon semble donc sous-estimé.*
*pour l'estimation fréquentielle (*3*ème ligne), pourcentage de l'énergie conservée par rapport à l'image originale sur les* 100%, 85%,65% *et* 50% *des plus hautes fréquences (resp. noir, rouge, vert et bleu).*
pour l'énergie des gradients, pourcentage de l'énergie conservée par rapport à l'image originale sur les gradients de norme supérieure à 0, 10, 20 *et* 36 *(resp. noir, rouge, vert et bleu).*
1*ère ligne : images floues.* 2*ème ligne : carte des rayons estimés par blocs* 4×4 *avec (d) moyenne des rayons estimés . 3ème ligne : logarithme décimal de la norme du spectre avec application d'un filtre médian* 5×5 *avec (d) pourcentage des hautes fréquences conservées. 4ème ligne : norme des gradients avec (d) pourcentage de l'énergie des gradients conservés avec différents seuils*



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

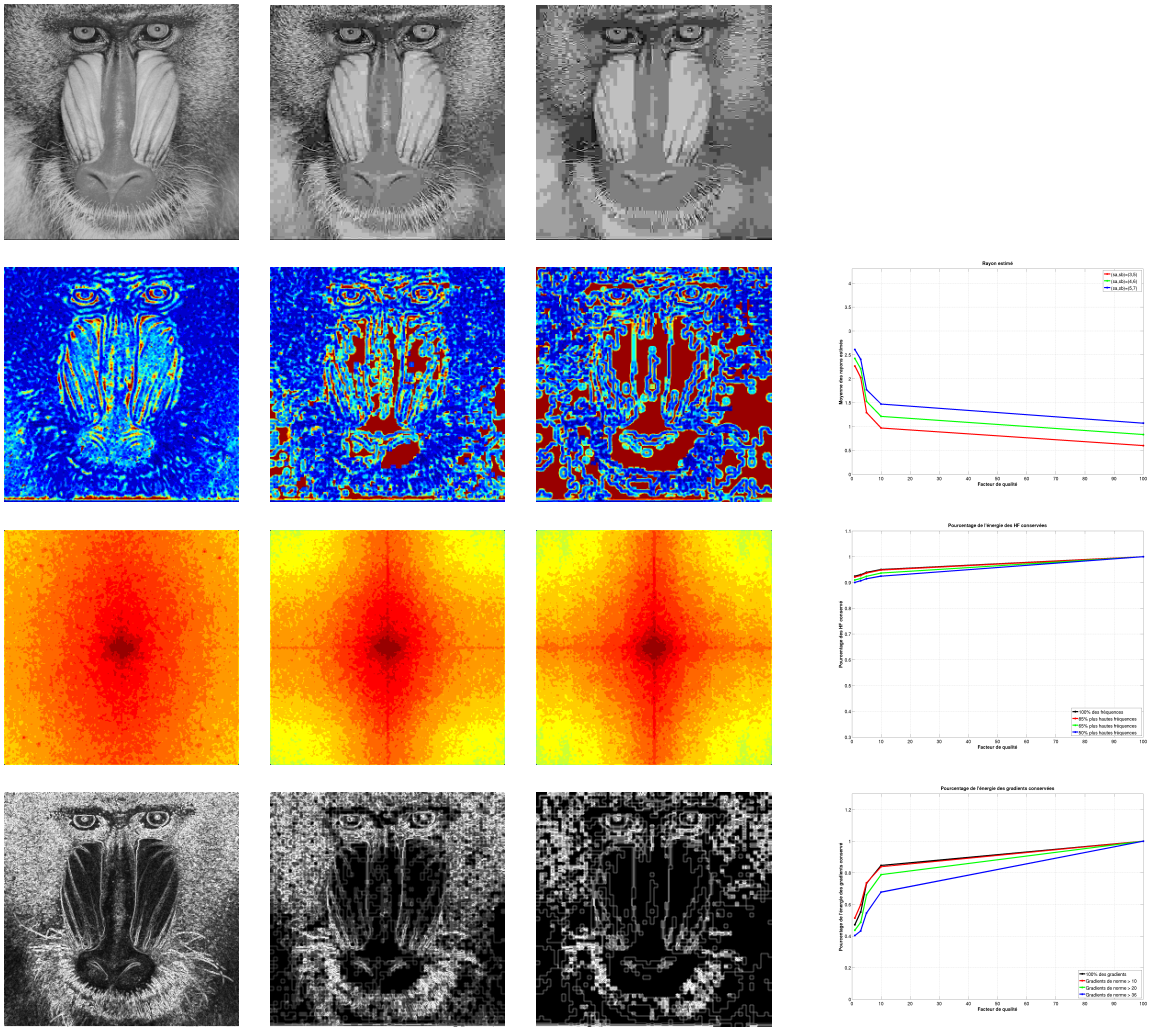
FIG. 2.5 – *Image "mandrill", flous de bougé et résultats (a) Originale (b) Déplacement de* 11 *pixels avec un angle de* 60°*. (c) Déplacement de* 25 *pixels avec un angle de* 45°*. (d) Métriques de résumé (cf. légende de la figure 4.4).* 1*ère ligne : images floues.* 2*ème ligne : carte des rayons estimés par blocs* 4×4 *avec (d) moyenne des rayons estimés .* 3*ème ligne : logarithme décimal de la norme du spectre avec application d'un filtre médian* 5×5 *avec (d) pourcentage des hautes fréquences conservées.* 4*ème ligne : norme des gradients avec (d) pourcentage de l'énergie des gradients conservés avec différents seuils*

codage. Nous avons fait varier le "facteur de qualité" q, directement lié à la matrice des poids de quantification appliqués aux valeurs des matrices blocs obtenues par DCT : plus le facteur de qualité est proche de 1, plus la compression est forte, au contraire, plus le facteur est proche de 100, plus faible elle sera. En pratique, la baisse de qualité de l'image est peu visible pour un facteur supérieur à 50 voire 20. Nous avons choisi de faire évoluer q entre 1 et 10. La figure 2.6 fournit des illustrations des différentes approches étudiées sur l'image "Mandrill" ayant subi des compressions jpeg d'intensité variable.

L'étude des résultats dans le domaine spectral montre que l'approche correspondante, l'analyse de l'énergie fréquentielle, est peu adaptée pour juger la qualité d'une image compressée sous format jpeg. En effet, la variation des énergies est minime : seules les très hautes fréquences sont filtrées véritablement, les fréquences plus faibles étant peu perturbées. La courbe des énergies montre, comme pour le flou de bougé, que la diminution relative de l'énergie par rapport à l'image originale ne dépend guère des fréquences considérées. La raison en est que la répartition spectrale varie peu : la compression jpeg conserve presque intactes les fréquences les plus basses en les favorisant lors de la quantification (cette dernière est moins brutale sur les coefficients correspondant aux basses fréquences). Sur l'image "mandrill", il est toutefois notable que des "bandes" centrées sur les axes horizontaux et verticaux ressortent pour un facteur de qualité minimal (égal à 1), ce qui montre une atténuation des fréquences supérieures. En revanche, pour "boat" et "lena", le rapport de l'énergie des hautes fréquences de l'image compressée par rapport à l'énergie des hautes fréquence de l'image originale, n'est pas minimal pour les facteurs de qualité les plus bas : la compression jpeg crée des discontinuités entre régions homogènes quantifiées fortement, ce qui rajoute des hautes fréquences. Cet effet dépend du contenu de l'image.

L'étude des gradients montre l'évolution de l'énergie avec la modification du facteur de qualité q. Les gradients disparaissent lorsque q diminue, particulièrement au sein des zones les plus homogènes de l'image d'origine. L'effet de bloc est visible, particulièrement dans les conditions les plus dégradées (facteur de qualité égal à 1). Selon l'image d'origine, cela entraîne d'ailleurs une augmentation de l'énergie des gradients pour un facteur de qualité faible (inférieur à 10). Ainsi, dans le cas de l'image "lena", l'apparition de zones homogènes est associée à la création de nouveaux gradients à la frontière de ces zones et l'énergie résultante est supérieure à l'énergie originale. Une solution possible consiste à ne compter que le nombre des gradients et non pas leur énergie. En effet, l'homogénéisation de l'image avec un facteur de qualité faible diminue le périmètre des contours. L'énergie des gradients nécessite une référence de même contenu image (hors effet de la dégradation) afin de disposer d'une mesure de qualité signifiante car les gradients dépendent fortement du contenu de l'image (par le biais des contours). Dans le contexte applicatif retenu, à savoir une mesure de qualité image de séquences vidéo aériennes, cette approche parait donc inadaptée. Il faudrait en effet recaler les images pour une évaluation relative (une image de la séquence étant choisie comme référence), avec un changement régulier de référence étant donnés les déplacements importants du capteur et la dérive correspondante de la zone observée au sol.

L'estimation du rayon de noyau gaussien suit également l'évolution du facteur de qualité. En effet, lorsque q diminue, les régions homogènes de l'image sont remplacées par des blocs constants de taille de plus en plus grande. Le rayon estimé pour ces régions augmente donc visiblement (cf. figure 2.6, deuxième ligne, (c)). Il s'agit des régions où les hautes fréquences, ou détails de textures, sont particulièrement filtrés voire supprimés. Ces régions coïncident d'ailleurs avec les zones "sans gradients" (cf. figure 2.6, quatrième ligne, (c)). La courbe des rayons estimés peut donc être utilisée comme une indication de la qualité de



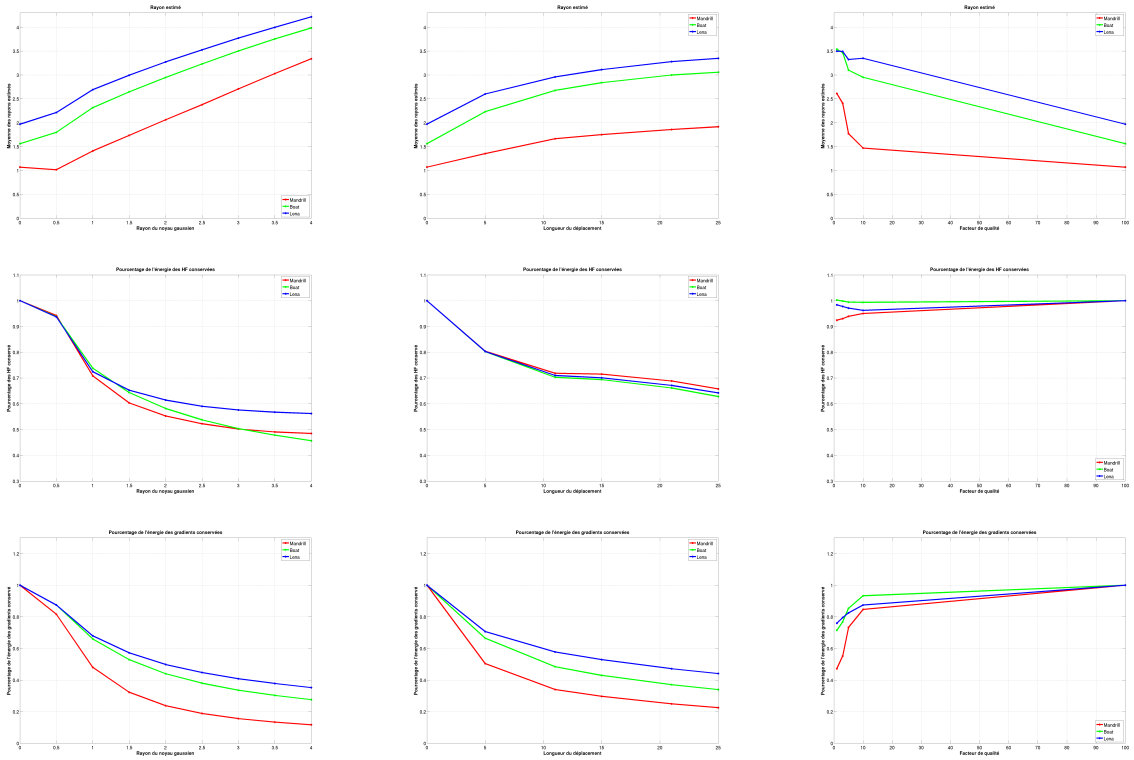
(a) (b) (c) (d)

FIG. 2.6 – *Image "mandrill", compression jpeg et résultats (a) Originale (b) Facteur de qualité* 5 *(c) Facteur de qualité* 1*. (d) Métriques de résumé.* 1*ère ligne : images floues.* 2*ème ligne : carte des rayons estimés par blocs* 4×4 *avec (d) moyenne des rayons estimés .* 3*ème ligne : logarithme décimal de la norme du spectre avec application d'un filtre médian* 5×5 *avec (d) pourcentage des hautes fréquences conservées.* 4*ème ligne : norme des gradients avec (d) pourcentage de l'énergie des gradients conservés avec différents seuils*

l'image, mais elle ne traduit pas réellement l'intensité d'un quelconque flou de bougé ou de mise au point.

2.3.4 Conclusion

Les différents critères d'évaluation présentés ci-dessus permettent donc dans une certaine mesure d'estimer la sévérité des dégradations de qualité, par rapport à une référence ou dans l'absolu. La méthode fondée sur la différence de gaussiennes a le mérite de fournir un résultat interprétable directement, sans ambiguïté due à une méconnaissance du contenu de l'image (ce résultat apparaît toutefois peu discriminant en présence de flou de bougé, mais l'indicateur obtenu fournit tout de même une indication acceptable du "flou ressenti"). Ainsi, une image très complexe présentant de nombreux contours aura une énergie de gradients ou spectrale plus importante qu'une image moins complexe ou avec moins de détails et contours, même si elle est plus floue que cette deuxième image. Il est alors nécessaire de disposer d'une référence, qui peut être une image de la séquence vidéo analysée. En revanche, dans le cadre d'une séquence prise par une caméra mobile, le contenu sera progressivement modifié. Il faudra donc convenir d'une stratégie de com-



(a) (b) (c)

FIG. 2.7 – *Résultats des différentes méthodes d'estimation du flou sur plusieurs images.* 1*ère ligne : estimation du rayon de noyau gaussien par différence d'images refloutées.* 2*ème ligne : énergie spectrale.* 3*ème ligne : énergie ou norme* L^2 *des gradients. (a) Avec flou gaussien (b) Avec flou de bougé d'angle* 0° *(c) Avec compression Jpeg. Rouge : mandrill, vert : boat, bleu : lena*

paraison des images (cf. paragraphe 2.4).

L'analyse spectrale est particulièrement intéressante pour repérer la présence d'un flou de bougé ainsi que la direction et la longueur du déplacement correspondant. La détection sera toutefois plus délicate en présence de plusieurs défauts complémentaires (flou de mise au point ou compression jpeg par exemple) qui brouilleront les bandes créées par les normes des coefficients dans l'espace de Fourier. Elle est en revanche moins indiquée pour la caractérisation du flou de mise au point, particulièrement lorsque ce dernier est important. Quant à l'estimation du facteur de qualité en présence de compression jpeg, elle apparaît inadaptée.
L'étude des gradients permet d'apprécier l'importance des trois défauts cités. Elle rend également la localisation des régions fortement texturées possible, ainsi qu'*a contrario* celle des régions homogènes. Elle nécessite en revanche une référence dans un but de comparaison et dépend fortement du contenu de l'image et apparait à ce titre peu adaptée au cadre de séquences vidéo aériennes.

2.4 Intensité du flou comme critère d'indexation vidéo

Afin de pouvoir comparer les résultats obtenus pour différentes images, qu'il s'agisse des cartes de rayon local de PSF, d'images de normes de gradients ou encore de transformées de Fourier, il semble nécessaire tout d'abord de recaler les images de la séquence d'origine dans un référentiel commun, afin de compenser le déplacement de la caméra. Cela permet de sélectionner sur chaque image un masque correspondant à une zone (ou contenu de la scène) observée identique. Ce recalage présente toutefois un double

inconvénient, l'augmentation du temps de calcul et l'introduction d'erreurs de recalage (imprécision et occultations).

Dans le cadre de vidéos aéroportées avec un déplacement important du porteur, cela n'est donc pas envisageable. En effet, si le recouvrement entre des images successives est important et permet d'effectuer une telle comparaison, cela n'est plus valable pour des séquences comportant des dizaines voire des centaines d'images. Il reste possible d'étudier l'évolution du rayon du noyau sur les couples d'images successives mais un compromis entre temps de calcul et précision doit être trouvé. Le recalage permet de comparer les images à contenu constant aux occultations près (ces dernières peuvent être prises en compte en n'effectuant la comparaison que sur les zones observables dans les deux images) mais son coût temporel est important car il nécessite le calcul du flot optique (ou autre méthode permettant d'évaluer le déplacement image) et d'une interpolation.

2.4.1 Sans recalage

Une dernière solution consiste à n'effectuer aucun recalage et à évaluer la qualité de la séquence image par image selon une ou plusieurs des méthodes présentées en 2.3. En effet, l'intervalle temporel pour une séquence vidéo est très court (pour une fréquence classique de 25 Hz, 40 millisecondes). Le recouvrement entre deux images consécutives de la séquence est donc important et dans un but d'évaluation de la qualité, le recalage ne modifiera que très peu les résultats. L'évolution du rayon et de l'inverse des énergies, spectrale ou des gradients, relativement aux valeurs prises pour la première image de la séquence, sont illustrées sur la première ligne de la figure 2.8 (a). Une augmentation correspond donc selon le critère choisi à une augmentation du rayon estimé, ou à une diminution de l'énergie, ce qui traduit un flou plus marqué. Un premier point notable concerne l'utilisation de la transformée de Fourier. En effet, malgré des variations importantes de contenu et de sévérité du flou, les variations de l'énergie spectrale sont minimes. En particulier, la dégradation due à un zoom rapide entre les images 121 et 161 n'est pas marquée par une augmentation du critère, ou cette dernière est à peine visible. Un changement d'échelle permet d'observer une faible augmentation, mais dont l'intensité (quelques pourcents) n'est pas significative en regard de la sévérité des dégradations associées, entre la deuxième et troisième image de la figure 2.8.

L'énergie des gradients en revanche semble fournir un critère intéressant corrélé aux variations de qualité. La dégradation entre les images 121 et 161 apparaît en effet clairement, de même que l'amélioration plus progressive entre les images 161 et 201, respectivement par une augmentation importante puis une diminution plus marquée du critère, i.e., l'inverse de l'énergie des gradients. Toutefois, l'amélioration progressive de la qualité entre les images 161 et 201 est traduite approximativement, avec des valeurs équivalentes pour les images 171 et 181, 191 et 201 respectivement. Cela s'explique par la variation du contenu et de la résolution, en effet sur ces dernières images, le déplacement et changement de focale de la caméra fait apparaître un contenu plus texturé et donc porteur d'énergie de gradients (à conditions de dégradations égales).

L'estimation des rayons ne fournit sur cette séquence pas de bons résultats. Les variations du critère ne correspondent en effet pas aux variations observées du flou, avec notamment une augmentation du rayon estimé entre les images 161 et 201. Cela est dû à deux phénomènes. D'une part, la variation de contenu et notamment la disparition du contour circulaire de l'objectif qui correspondait à une zone où le rayon estimé était important, compense l'augmentation du rayon estimé sur le centre de l'image entre les images 121 et 161. D'autre part, les watermarks incrustées sur l'image et peu texturées créent artificiellement des régions importantes où le rayon estimé est élevé. Ces régions pèsent fortement dans le calcul du critère. Une solution est envisageable à partir de métriques plus élaborées telles

que comparaison d'histogrammes ou des quantiles correspondant aux valeurs basses du rayon estimé. En effet, les régions présentant de faibles valeurs seront plus présentes si le flou est moins marqué. Elles correspondent de plus aux régions les plus finement texturées et sont donc plus fiables.

2.4.2 Recalage

La comparaison des énergies spectrales est mal adaptée au cadre du recalage. En effet, les masques correspondant à la région observée commune aux deux images successives à comparer sont de taille différente et peuvent présenter des "trous" dus aux occultations ou à la présence d'objets mobiles en bord d'image (présents sur l'une des deux images et absents sur l'autre). L'utilisation des coefficients de Fourier suppose donc de faire intervenir des métriques robustes ou de combler les "trous", par un algorithme de type "inpainting" par exemple. Toutefois, les résultats obtenus en l'absence de recalage montrent que l'analyse de l'énergie spectrale est peu discriminante malgré des variations visibles de flou et les différences de contenu dues aux occultations et dérive globale de la caméra.

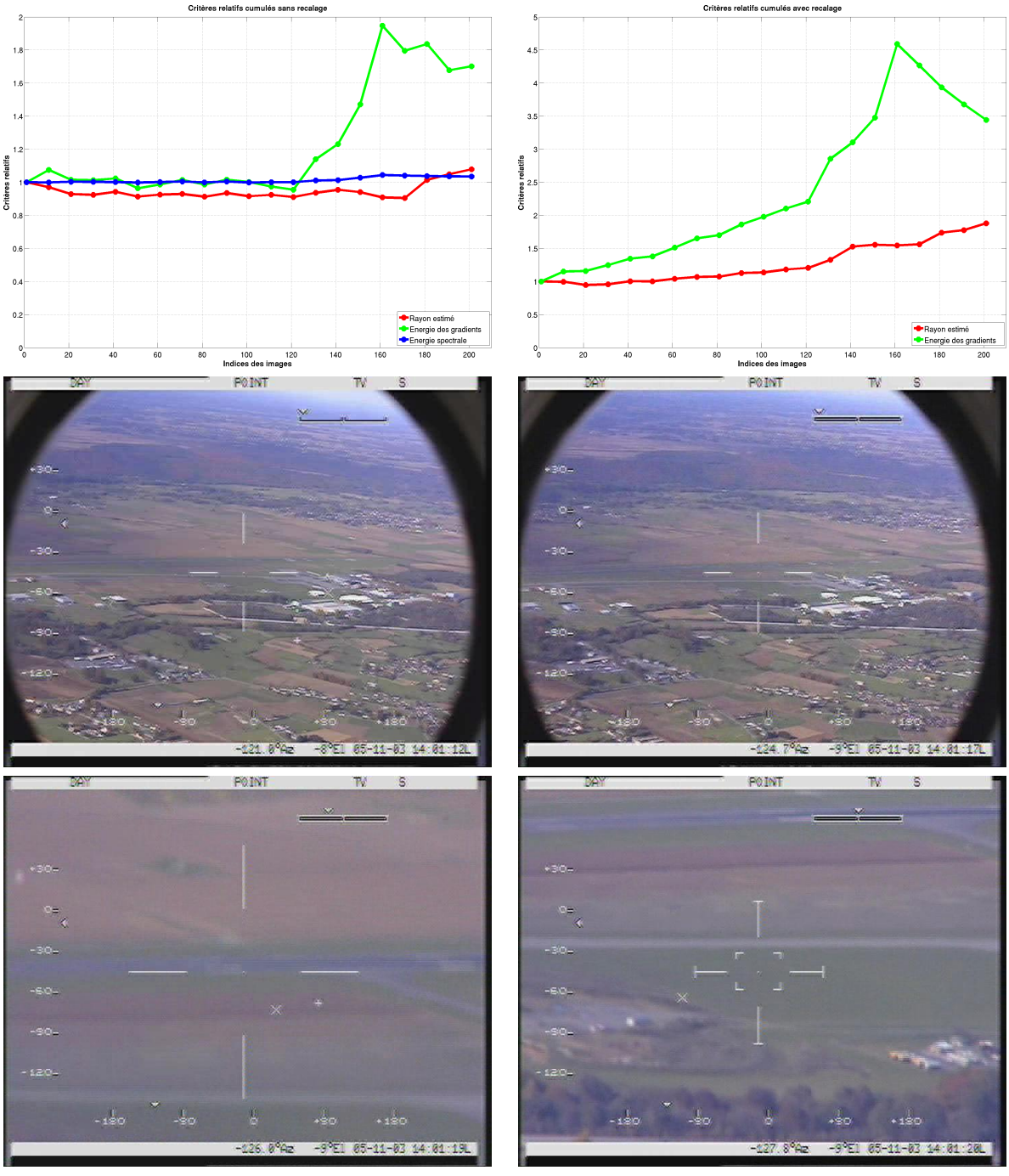
En revanche, l'énergie des gradients ou la moyenne des rayons peuvent être comparées facilement sur la région commune (dans les référentiels respectifs de chacune des deux images). La question se pose d'annuler la transformation affine voire le flot complet avant de calculer l'énergie ou la moyenne suivant la méthode. Un recalage complet à partir d'un flot optique calculé en chaque point nuirait fortement au but affiché de la méthode : en effet, l'image recalée (la deuxième) présenterait des gradients ou structures locales très similaires à ceux de l'image de référence (la première) et l'évolution des critères, quasi nulle, ne donnerait aucune indication sur les modifications de sévérité du flou. Il apparaît donc plus pertinent de n'utiliser le recalage que pour sélectionner la région commune.

Pour les gradients, nous avons choisi de représenter l'évolution cumulée du rapport des énergies calculées pour chaque couple d'images sur les gradients restreints à la région commune aux deux images considérées. Pour la méthode fondée sur les différences de gradients, nous conservons également l'évolution cumulée du rapport des rayons estimés, ainsi que sur les régions communes à chaque couple. Les graphes d'évolution de ces deux critères ne sont donc pas associés à une unique région ni à un repère de référence. La figure 2.8 (b) trace l'évolution des deux critères, rayon relatif et énergie des gradients relative, au cours de la séquence.

Le recalage permet de supprimer l'influence de la modification de contenu en ne sélectionnant pour chacune des deux images que la région commune (et en normalisant l'énergie par le nombre de pixels des masques correspondants). Ainsi, les critères fondés sur les gradients et les rayons estimés sont plus cohérents avec l'évolution observée manuellement de la sévérité du flou. L'amélioration entre les images 161 et 201 est ici traduite par une diminution continue de l'inverse de l'énergie des gradients. En revanche, les rayons estimés sont toujours décorrélés du flou observé empiriquement. Le choix du critère, discuté au paragraphe 2.4.1, doit donc être plus robuste aux variations de contenu afin de ne prendre en compte que les régions les plus texturées et ignorer les éventuelles watermarks.

2.4.3 Indexation vidéo

Une fois choisi le mode d'évaluation du flou au cours du temps, il est possible d'en dériver un critère d'indexation puis éventuellement de résumé de la séquence vidéo. Il peut s'agir d'une simple sélection des images les plus nettes. La visualisation de ces images dans un but de détection voire d'identification, manuelle ou automatique, d'objets ou de structures en trois dimensions, sera d'autant plus efficace que le flou sera faible. Toutefois, une sélection des images nettes sur ce seul critère représentera la contenu de la scène de



(a) (b)

FIG. 2.8 – *Évolution des critères relatifs par rapport à la première image.* 1ère *ligne : critères (a) sans recalage, énergie ou moyenne relative à la valeur correspondante pour la* 1ère *image ; (b) avec recalage, produit cumulé des rapports d'énergies ou moyennes pour chaque paire de images consécutives.* 2ème *ligne : images* 1 *et* 121 *(début et fin de la première "période" de dégradation progressive.* 3ème *ligne : images* 161 *(fin du zoom et dégradation abrupts) et* 201 *(fin de stabilisation et amélioration de la qualité)*

façon aléatoire dans le temps. Le résumé obtenu pourrait alors ne pas représenter des segments entiers car trop flous et entraînerait donc une perte d'informations. Il est donc nécessaire de disposer d'une vue d'ensemble qui permettra dans un premier temps de laisser un interprète humain restreindre la zone d'étude pour la sélection d'images nettes.

Mosaïques Cela suppose par exemple la création de mosaïques. Dans l'optique d'une analyse statique, avec détection et identification d'objets fixes ou structures en trois dimensions telles que des bâtiments, une mosaïque du fond, sans information dynamique, est suffisante. Si la fréquence des images nettes disponibles est suffisamment élevée, il est possible d'améliorer la qualité de la mosaïque en la fabriquant à partir de ces images sélectionnées [36, 72].

Super - résolution Enfin, dans une recherche de détail, une approche de super-résolution pourra fournir des détails supplémentaires sur la forme d'objets ou structures présents dans la scène, tels que des véhicules ou bâtiments, ou encore piétons, selon la résolution d'origine. Une estimation du flou peut alors être utilisée comme information [202] ou comme critère de performance de l'algorithme de super-résolution utilisé [189]. L'ensemble des diverses opérations de recalage, super-résolution et estimation du flou peut aussi être réalisée de façon conjointe et itérativement [282]. Létienne et al. s'attachent en particulier à la super-résolution d'objets en mouvement et à l'importance de l'étape de recalage [155]. Rochefort et al. proposent un nouveau modèle d'observation dans le cadre de vidéos aériennes [210] afin de traiter des mouvements affines plus généraux.

Résumé dynamique Si en revanche, l'objectif de l'interprète est de disposer d'un résumé dynamique, le flou sera moins important mais permettra tout de même d'obtenir un résumé plus net, dans le cas par exemple d'un échantillonnage de la séquence. Il faudra alors trouver un compromis entre régularité temporelle et netteté des images retenues, particulièrement sur la base d'un échantillonnage uniforme. Cela peut passer par la formulation d'une énergie à minimiser prenant en compte ces différents critères (régularité temporelle, critère de qualité) sur le modèle de [198].

2.5 Discussion

Un flux ou séquence vidéo est un objet spatiotemporel riche et complexe qui peut être étudié sous différents critères complémentaires. Outre l'analyse de son contenu statique et dynamique, une estimation de la qualité visuelle complète la caractérisation de la séquence. Simple mesure évoluant au cours du temps, elle peut représenter un index temporel afin de sélectionner directement les segments de la séquence de bonne qualité, plus à même de contenir des détails nets. Des structures plus élaborées telles que des mosaïques ou des résumés vidéo peuvent également intégrer une telle mesure de qualité afin de faire ressortir les détails de la scène.

Parmi les différents artefacts et dégradations possibles sur une séquence vidéo, nous nous sommes plus particulièrement attachés à l'étude du flou, notamment du flou de mise au point, susceptible d'évoluer avec le déplacement et changement de focale de la caméra ainsi que les modifications du terrain observé. Le développement de trois approches a permis de mettre en exergue plusieurs points.

- Une référence est nécessaire afin d'obtenir une évolution significative du critère retenu. Cette référence peut prendre la forme d'un recalage entre couples d'images consécutives (voire sur des segments temporels plus longs). Le critère est alors calculé

Caractéristiques	Estimation de noyau	Énergie spectrale	Énergie des gradients
Référence nécessaire	non	oui	oui
Flou gaussien	bonne	moyenne	bonne
Flou de bougé	moyenne	bonne	bonne
Jpeg	bonne	inadaptée	moyenne
Sensibilité au contenu	non	oui	oui
Sensibilité à la métrique	oui	oui	oui

TAB. 2.2 – *Caractéristiques des approches d'estimation du flou détaillées au chapitre 2*

sur un contenu image équivalent. Les méthodes utilisant respectivement les énergies spectrales ou de gradients dépendent particulièrement du contenu image (les énergies correspondantes sont beaucoup plus élevées si la scène observée présente des détails nombreux ou des textures fines), cette étape de recalage est importante. Etant donnée la fréquence temporelle d'échantillonnage élevée dans la plupart des séquences vidéo récentes, une approximation par couples d'images consiste à ne pas effectuer de recalage et observer l'évolution relative du critère sur le couple, en considérant un recouvrement quasi-total de la scène observée entre les deux images. Cette approximation pourra faire défaut lors de l'apparition ou disparition sur l'image de contenu haute fréquence. L'estimation du rayon gaussien de flou fournit une mesure moins dépendante du contenu, mais un calcul par blocs ne donne des résultats pertinents qu'aux abords des contours.

Il importe donc également de choisir une métrique réduisant la dépendance au contenu de l'image. Si une moyenne simple de la carte de rayons estimés aboutit à une première idée de l'évolution du flou, elle n'est pas totalement robuste à l'apparition ou disparition de contenu à hautes fréquences spatiales. Ce choix de métrique peut également aider à limiter la dépendance aux données des deux autres approches (énergie spectrale et énergie des gradients).

– Les approches peuvent être complémentaires. Ainsi, l'approche fréquentielle semble plus adaptée à l'estimation d'un éventuel flou de bougé que les deux autres, plus sensibles à un flou de mise au point. Les séquences vidéo aériennes étant des objets riches présentant des perturbations complexes, il est plausible que les déformations comprennent à la fois un flou de bougé et un flou de mise au point, voire des artefacts de compression tels qu'entrelacement, effets de bloc... Une combinaison de méthodes d'estimation de défauts spécifiques, ou des méthodes plus génériques visant à estimer conjointement l'ensemble des déformations contenues dans le flux vidéo, sont donc nécessaires pour caractériser au mieux la qualité de ce flux.

2.6 Conclusion et perspectives

Dans le cadre de l'indexation de séquences vidéo, la qualité image apparaît comme une modalité pertinente pour filtrer des segments temporels jugés peu informatifs car trop dégradés ou, au contraire, pour déterminer les images les plus susceptibles de contenir des détails nets. Cette modalité est donc particulièrement intéressante pour des applications requérant une grande précision telles que l'identification d'objets ou structures spécifiques.

Nous avons dans cette partie étudié plus particulièrement des méthodes d'estimation du flou, dégradation due au mouvement de la caméra voire d'objets mobiles et à des changements de mise au point. L'utilité de ces méthodes pour l'évaluation d'un défaut de

compression par blocs de type jpeg a également été mesurée. La table 2.2 résume plusieurs caractéristiques de trois méthodes d'estimation : estimation de noyau, énergie spectrale et énergie des gradients. Les méthodes décrites sont complémentaires, avec des performances variables selon le type de défaut rencontré. Des critères fondés sur l'énergie des gradients ou sur l'estimation du rayon fournissent de bonnes estimations de l'importance du flou pour un contenu donné. En revanche, les gradients dépendent fortement du contenu de la scène et l'estimation du rayon n'est précise qu'aux abords des contours. Dans un but d'indexation temporelle, un critère robuste au contenu et résumant les cartes de résultats en un paramètre unidimensionnel apparaît ainsi nécessaire afin de fournir une "note" ou critère de qualité unidimensionnel.

Les approches étudiées montrent plusieurs limitations et ouvrent plusieurs perspectives. Afin d'obtenir des critères robustes au contenu et des représentations compactes correctes et visuellement agréables, il importe d'améliorer divers points :

- Construire une ou plusieurs métriques limitant la dépendance au contenu image, par exemple ne conserver que le maximum des rayons de flou gaussien sur l'image ou utiliser des mesures relatives pour les énergies de gradient ou de spectre (en étudiant par exemple l'évolution de la distribution d'énergie spectrale plutôt que l'évolution de l'énergie elle-même)

- Intégrer les mesures de qualité dans des résumés compacts. L'idée implicite est de maximiser le ratio entre quantité d'information et temps nécessaire à la consultation. Ainsi, un résumé dynamique sur lequel les différents objets ou autres détails de bâtiments apparaissent nettement évitera une relecture de la séquence originale et l'utilisation d'algorithmes de super-résolution. Il existe par exemple un compromis entre quantité de données disponibles (nombre d'images utilisées pour la constitution d'une mosaïque) et la qualité de ces dernières, une seule image choisie parmi une dizaine d'images étant insuffisante pour afficher l'intégralité de la scène.

Etude du mouvement global

3

Caractériser précisément la qualité d'une séquence vidéo permet d'orienter une recherche axée sur la précision et l'identification du contenu ou encore d'obtenir un résumé statique (mosaïque) ou dynamique (résumé vidéo) de meilleure qualité visuelle. Le flou de mise au point, notamment, est un critère de qualité susceptible de varier au cours de la séquence, au contraire de dégradations liées à la compression. La qualité n'apporte en revanche aucune information sur la dynamique de la séquence. Le mouvement global image peut être utilisé dans un but de recalage, pour la détection de changement ou mouvement, ou pour la réalisation de mosaïques. Mais il représente également la projection sur le plan image du mouvement caméra par rapport à la scène observée. La caractérisation du mouvement caméra fournit une autre modalité d'indexation de la séquence vidéo. Ainsi, les segments temporels présentant des mouvements caméra parasites et une image inexploitable sont ainsi directement filtrés, les segments associés à des changements de focale signifient un agrandissement ou rétrécissement du champ de vue.

3.1 Rappels bibliographiques

L'analyse d'une séquence vidéo par le biais du mouvement global passe par plusieurs étapes. Lorsque cela est pertinent, dans le cas de documentaires, de films ou d'autres séquences présentant des coupures ou des transitions entre scènes, une première étape consiste à trouver les limites de chaque scène. Chaque scène peut ensuite être analysée plus en détail, en caractérisant le mouvement global selon plusieurs critères tels que zoom, translation horizontale ou verticale, rotation... Ces informations peuvent être ensuite utilisées dans un but de résumé ou d'indexation temporelle de la séquence vidéo. Il est également possible d'interprèter le mouvement caméra comme un signe d'intention particulière de la part de l'opérateur lors de la prise de la vidéo. Le schéma de la figure 3.1 résume l'enchaînement de ces étapes successives.

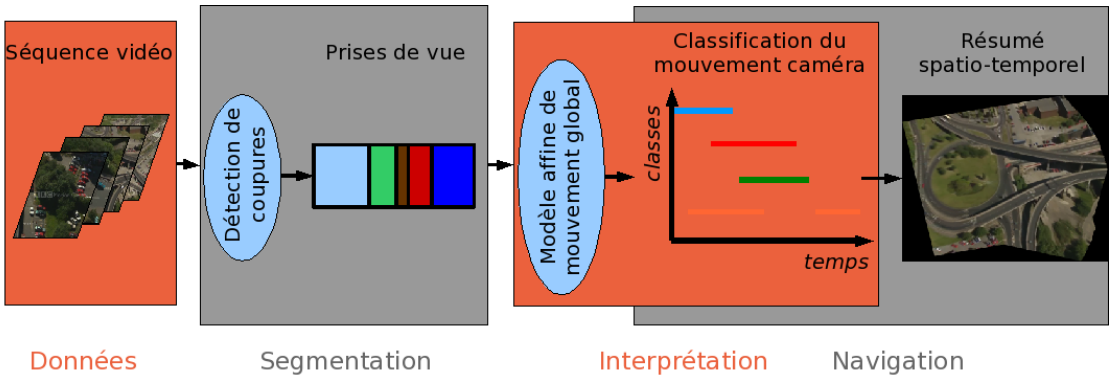


Fig. 3.1 – *Exemple de représentation de chaîne d'indexation d'une séquence vidéo par le mouvement global.*

Découpage d'une séquence vidéo L'analyse structurelle de séquences vidéo est une étape préliminaire nécessaire à une analyse automatique du contenu vidéo. Une séquence peut être décrite selon plusieurs niveaux de structure tels que l'image, la prise de vue, la scène, etc. Dans une optique de navigation et de recherche vidéo, le niveau de la prise de vue est considéré comme pertinent dans [226]. Ce niveau correspond à un mouvement caméra continu dans le temps. Les transitions entre prises de vue peuvent être abruptes, il s'agit alors de "coupures", ou progressives (dissolution, balayage, damier, fondu...).

Un grand nombre de méthodes visent à détecter ces transitions. Ces méthodes peuvent être fondées sur l'analyse du contenu. Ainsi, Huang et al. apparient des points d'intérêt [108] pour une certaine robustesse par rapport aux mouvements d'objets ou de caméra. Aoki propose d'analyser l'erreur des mouvements image estimés afin de détecter et caractériser les changements de prise de vue [9]. Une modélisation statistique du mouvement image par une distribution de Laplace permet de détecter les changements de prise de vue tout en restant robuste aux changements de luminosité [11]. Plusieurs méthodes opèrent directement dans le domaine de compression vidéo [26, 166, 143]. Un état de l'art des méthodes pour la détection de changement de prise de vue a été réalisé par Lefèvre et al. dans [137] puis par Yuan et al. dans [281]. Ce problème constitue l'une des tâches de la campagne TRECVID [Tre, 180] organisée par l'Institut national des standards et technologies (NIST). Un large ensemble de séquences avec les vérités terrains associées est fourni dans le cadre de cette campagne à des fins d'évaluation.

Estimation du mouvement de la caméra La caractérisation du mouvement caméra au cours du temps de chaque prise de vue apporte des précisions supplémentaires dans le but d'une indexation plus précise, d'une aide au résumé ou la détection de régions d'intérêt spatio-temporelles. Ainsi, Bouthemy et al. [37] utilisent-ils une approche statistique fondée sur des tests de rapport de vraisemblance pour interpréter le mouvement caméra en 6 classes selon la présence de zoom, rotation autour de l'axe optique, translations et rotations autour des deux axes restants. Dans [134], Lan et al. proposent un modèle à trois paramètres, deux de translation horizontale et verticale, et un facteur radial. Le mouvement dominant est alors associé à l'un de ces trois paramètres, soit 4 classes de mouvement (la dernière classe correspondant à l'absence de mouvement). D'autres approches analysent directement le mouvement dans le domaine compressé. Ainsi, Tiburzi et Bescos sélectionnent parmi les vecteurs de mouvement locaux dans le domaine compressé (format MPEG) les plus représentatifs du mouvement global : après un découpage de chaque image en régions, seuls les ensembles de vecteurs d'une région cohérents en direction et majoritaires dans la région sont conservés. Ces vecteurs sont ensuite utilisés afin de déterminer le mouvement global selon un modèle affine à 6 paramètres [235]. Su et al. forment des flux de mouvement en suivant les vecteurs de mouvement locaux dans [228]. Ces différentes approches détectent également les changements de prise de vue, à partir de variations des modèles obtenus ou directement à partir des variations de mouvement (dans le domaine compressé ou non).

Classification et interprétation du mouvement caméra Les résultats obtenus sur l'estimation de mouvement global, sous la forme de modèle paramétrique (affine ou homographique par exemple) ou directement de classes de mouvement telles que zoom, rotation autour de l'axe optique, tilt ou pan etc., peuvent ensuite être interprétés manuellement ou automatiquement. Cela passe par une traduction sémantique de haut niveau en combinant et en seuillant les différents mouvements élémentaires de caméra tels que zoom, translation ou rotation autour des différents axes de la caméra. Cette analyse peut être fondée sur une estimation en deux [122] ou en trois dimension [248] du mouvement de caméra. Des a priori sont toutefois nécessaires afin de définir les classes ou combinaisons de classes de mouvement susceptibles de représenter un intérêt particulier pour l'interprète.

Résumé de séquence vidéo à partir du mouvement caméra La caractérisation du mouvement caméra peut servir de base, avec ou sans interprétation sémantique de haut niveau, à un résumé de la séquence vidéo analysée. En effet, il est possible de regrouper les segments temporels présentant un mouvement global homogène, par exemple par une approche de classification hiérarchique ascendante [129] en fusionnant itérativement les groupes de séquences semblables. La création d'un résumé vidéo peut ensuite être guidée par ces segments temporels, en sélectionnant des images représentatives ou "keyframes" pour chacun de ces segments [2]. Cela rejoint les travaux de Nam et Tewfik [172] qui extraient ces "keyframes" en fonction d'un indicateur de mouvement calculé à chaque image. Une autre solution consiste à orienter le choix des "keyframes" ou segments temporels pertinents par une mesure d'intérêt, déduite par exemple de la quantité de mouvement associée et mesurée à l'aide d'un modèle probabiliste en considérant le problème de sélection comme un problème d'inférence bayésienne [129].

Autres emplois du mouvement caméra D'autres applications encore découlent de l'étude du mouvement global. Ainsi, la détection de régions d'intérêt peut être conduite par les paramètres de mouvement global en association avec des cartes de saillance dérivant d'informations d'apparence telles que des cartes de contraste [3]. En effet, selon les auteurs, l'attention visuelle est attirée vers les nouvelles régions susceptibles de contenir de nouvelles informations d'intérêt lors de mouvements de translation ou rotation tels que pan / track et tilt / boom, ou vers le centre de l'image en cas de zoom ou d'absence de mouvement. Une autre application consiste en la détection directe d'objets mobiles à partir des modèles de mouvement global 2D obtenus [57].

3.2 Choix du modèle et méthode de calcul du mouvement global

Le mouvement caméra peut être décomposé en un mouvement de rotation $\mathbf{w}$ et un mouvement de translation $\mathbf{v}$ en trois dimensions. Le mouvement relatif d'une scène fixe par rapport à la caméra correspond en trois dimensions à l'opposé du mouvement de la caméra par rapport à la scène. Il est traduit en deux dimensions par une projection sur le plan image. La fonction de projection correspondante relève d'une famille de fonctions différente selon le modèle de caméra choisi. Un modèle standard est le modèle sténopé ou projectif linéaire, défini par le centre optique (placé à l'endroit de la caméra) et le plan rétinien ou image sur lequel est projetée la scène 3D. Le modèle correspond à une projection centrale de l'espace euclidien sur ce plan. Dans ce modèle, le déplacement image associé au mouvement de rotation $\mathbf{w}$ et mouvement de translation $\mathbf{v}$ de la caméra s'écrit, en un point (x, y) du plan image :

$$u(x,y,t) = \frac{1}{Z}\begin{pmatrix} -1 & 0 & x \\ 0 & -1 & y \end{pmatrix}\mathbf{v}(t) + \begin{pmatrix} xy & -(1+x^2) & y \\ 1+y^2 & -xy & -x \end{pmatrix}\mathbf{w}(t) \tag{3.1}$$

où Z est la profondeur du point de l'espace réel dans le repère caméra projeté en (x, y) dans le repère associé au plan image. En faisant l'hypothèse d'une surface plane (dont le cas particulier d'une profondeur fixe), cela peut se réécrire sous l'approximation paramétrique quadratique suivante [176], en omettant pour des raisons de clarté la variable t :

$$u(x,y) = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} q_1 & q_2 & 0 \\ 0 & q_1 & q_2 \end{pmatrix}\begin{pmatrix} x^2 \\ xy \\ y^2 \end{pmatrix} \tag{3.2}$$

La différence d'intensité associée à ce modèle entre deux images consécutives f et g en suivant le déplacement, soit la différence d'intensité entre la première image f au point (x, y) et la seconde g au point associé après déplacement $(x, y) + u_\Phi(x, y)$ s'écrit alors ,

en notant $\Phi = (c_1, c_2, a_1, \cdots, a_4, q_1, \cdots, q_6)$ et u_Φ le champ de déplacement paramétrique associé :

$$DF_{\Phi,\xi}(x,y) = g((x,y) + u_\Phi(x,y)) - f(x,y) + \xi$$

où ξ correspond à un éventuel changement global d'illumination entre f et g. Les paramètres peuvent être estimées par la méthode des moindres carrés consistant à minimiser les carrés des différences sur (Φ, ξ) :

$$\min_{\Phi,\xi} \sum_{(x,y)\in S} DF_{\Phi,\xi}(x,y)^2$$

avec S le support choisi au sein de l'image. Une méthode plus robuste aux valeurs aberrantes (outliers) consiste à utiliser un M-estimateur, comme dans [176]. Le nombre de paramètres est souvent réduit, par exemple un modèle quadratique réduit avec par exemple $q_3 = q_4 = 0$ et $q_5 = q_1, q_6 = q_2$ ou un modèle affine avec $q_1 = \cdots = q_6 = 0$ et $a_3 = -a2, a_4 = a_1$. Ces différents modèles réduits prennent en compte diverses composantes du mouvement telles que divergence ou rotationnel du flot. Un modèle homographique à huit paramètres peut aussi être utilisé, associant à un plan principal de f un plan correspondant dans g :

$$f \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \tag{3.3}$$

où (x', y') représente les coordonnées du point dans g correspondant à (x, y).

Choix de l'algorithme de calcul du modèle global Nous avons utilisé le logiciel Motion 2D fondé sur la minimisation d'un critère d'erreur robuste de modèles paramétrés en deux dimensions [176]. Il s'agit de minimiser un M-estimateur de la différence d'images d'intensité (l'image d'origine et la seconde image après recalage) sur un support R_t compris dans l'image de référence. En notant $\rho(x, C)$ une fonction de pondération de la variable x et de paramètre d'échelle C, prenant des valeurs finies pour des valeurs élevées de x, et en intégrant dans les notations pour plus de clarté le paramètre d'illumination ξ dans Φ (ce qui revient à "écrire" $\Phi = (\Phi, \xi)$), le modèle de mouvement est alors donné par :

$$\hat{\Phi} = \arg\min_{\Phi} E(\Phi) = \arg\min_{\Phi} \sum_{(x_i,yi)\in R_t} \rho\left(DF_\Phi(x_i, y_i), C\right) \tag{3.4}$$

L'optimisation est réalisée au sein d'un schéma multi-résolution incrémental selon le schéma de Gauss-Newton et n'utilise que les dérivées spatio-temporelles de l'intensité. A chaque étape incrémentale k, qu'il s'agisse d'un passage à une échelle plus fine ou l'itération suivante à une même échelle, le modèle est mis à jour selon :

$$\Phi = \hat{\Phi}_k + \Delta\Phi_k$$

avec $\hat{\Phi}_k$ l'estimation en cours du vecteur de paramètres Φ. Une linéarisation de DF_Φ en $\hat{\Phi}_k$ fournit un résidu $r_{\Delta\Phi_k}$ linéaire en $\Delta\Phi_k$ et qui s'exprime en fonction du gradient spatial de l'intensité au temps $t + 1$. Dans le schéma multi-résolution incrémental, la minimisation de $E(\Phi)$ revient donc à minimiser :

$$E(\Delta\Phi_k) = \sum_{(x_i,y_i)\in R_t} \rho\left(r_{\Delta\Phi_k}(x_i, y_i), C\right)$$

Cette fonction d'erreur est minimisée selon un schéma des moindres carrés pondérés itératifs avec une initialisation nulle de $\Delta\Phi_k$.

Une autre possibilité consiste à calculer d'abord un flot optique sur l'ensemble de la première image et d'en dériver un modèle global par régression robuste en utilisant là encore un M-estimateur, par exemple fonction de Leclerc ($\rho(u, \sigma) = 1 - \exp[-(u^2/\sigma^2)]$) ou

Geman-McClure ($\rho(u,\sigma) = u^2/(\sigma^2+u^2)$) [81] sur le flot résiduel (ou flot après compensation du mouvement dominant). Les modèles paramétriques obtenus (affines, quadratiques) sont similaires à ceux de Motion 2D (il est possible sur le même principe d'estimer un modèle homographique plus général), et il peut être utile de conserver les flots optiques, à des fins de mosaïcage ou d'extraction de flot résiduel par exemple. En revanche le coût de calcul est plus élevé.

Choix du modèle global Le modèle affine présente les avantages d'être plus robuste qu'un modèle quadratique complet. De plus, la composition des affinités est simple et permet de recaler simplement des séquences d'images dans un repère commun. La figure 3.2 montre sur l'exemple choisi qu'un recalage affine semble plus robuste à de fortes perturbations de parallaxe. Seuls les véhicules en mouvement ainsi que les bâtiments de grande hauteur ressortent sur l'image de la norme du flot résiduel. Des modèles quadratique ou homographique ne recalent pas totalement le fond, particulièrement le modèle quadratique qui intègre le mouvement de parallaxe au détriment du recalage du fond. La présence d'un masque des bâtiments permettrait d'obtenir un meilleur modèle. Cela nécessite toutefois l'obtention de cartes de profondeur ou l'utilisation d'autres algorithmes de détection de structures en trois dimensions et un surcoût temporel important.

Toutefois, ce modèle est insuffisant en présence de distortions radiales ou de différents plans principaux dans la scène. Les distortions radiales et l'erreur obtenue en conséquence dans l'estimation du flot résiduel présentent un obstacle à la détection d'objets mobiles ou structures en trois dimensions par le biais du flot résiduel. La création de mosaïques ou d'images super-résolues est également impossible dans les régions mal recalées. Toutefois, pour l'objectif de classification du mouvement global, ces erreurs ne sont pas suffisantes pour perturber l'estimation du modèle. Ce dernier, estimé sur la partie centrale de l'image, traduit en effet correctement le déplacement du capteur par rapport à la scène (projeté sur le plan image), par exemple les composantes de translation et de rotation pour un modèle affine.

En revanche, si la scène présente plusieurs plans différents sans plan dominant marqué (un plan occupant une surface majoritaire ou du moins nettement supérieure aux autres plans sur l'image et servant ainsi de base pour une estimation robuste de modèle paramétrique de mouvement global), l'estimation du modèle sera fortement perturbée. La figure 3.3 montre un exemple d'une telle situation, avec deux plans occupant chacun presque la moitié de l'image. Une sélection manuelle approximative d'un unique plan peut remédier à cette difficulté au détriment du caractère automatique de l'estimation. Si la scène observée varie peu par rapport au capteur, il sera tout de même possible de propager le masque grâce au modèle paramétrique estimé.

3.3 Classification et interprétation du mouvement global

Les séries de paramètres des modèles obtenus pour les couples d'images consécutives d'une séquence vidéo ne sont pas tous directement interprétables, c'est-à-dire traduisant l'intensité de mouvements caméra classiques tels que zoom ou rotation.

Dans le cadre du modèle affine retenu toutefois, les deux paramètres de translation horizontale et verticale (c_1 et c_2 dans l'équation (3.2)) donnent le déplacement de l'intersection de l'axe optique avec le plan image. Il s'agit plus simplement du déplacement du centre de l'image en pixels. Ces deux paramètres permettent donc de tracer l'évolution de la zone observée, au facteur de zoom près. Cela peut être particulièrement utile dans le cadre d'une indexation spatiale, avec recherche des segments temporels associés à une région donnée : cette région peut être sélectionnée sur une mosaïque réalisée à partir de la séquence, ou directement par coordonnées d'une carte si un recalage géographique est disponible. En

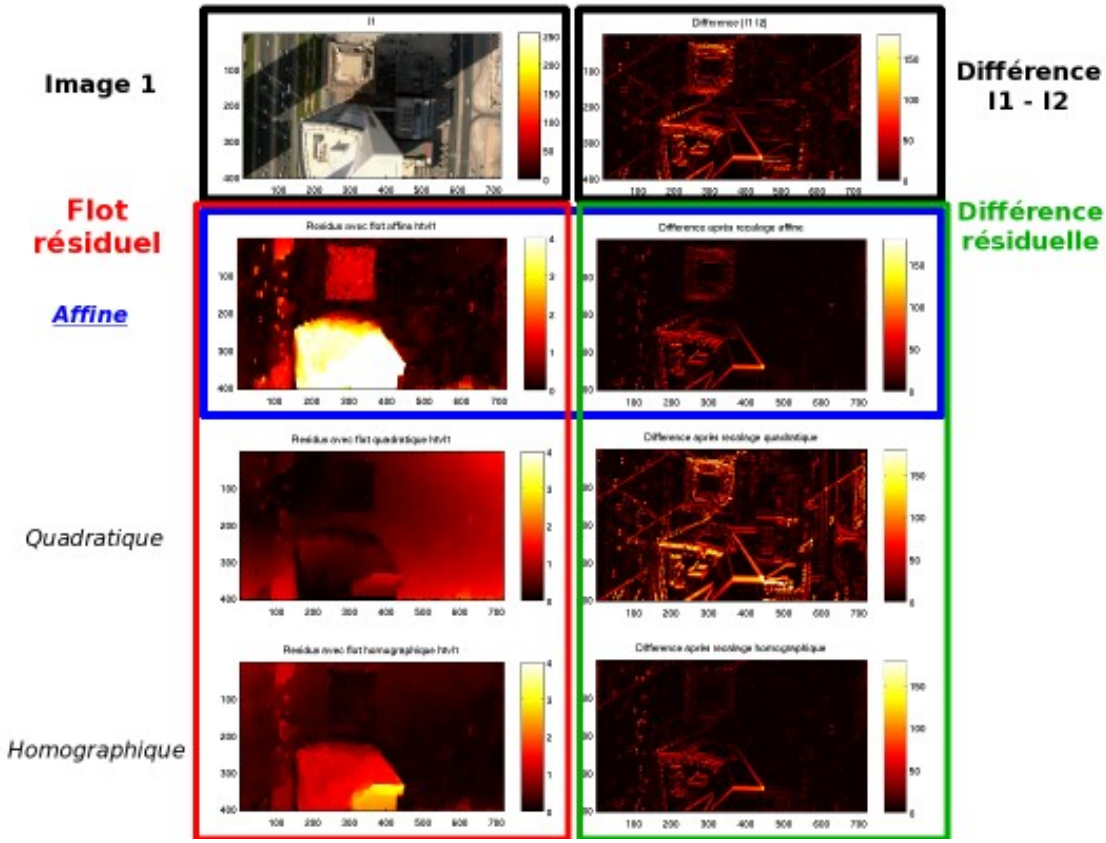
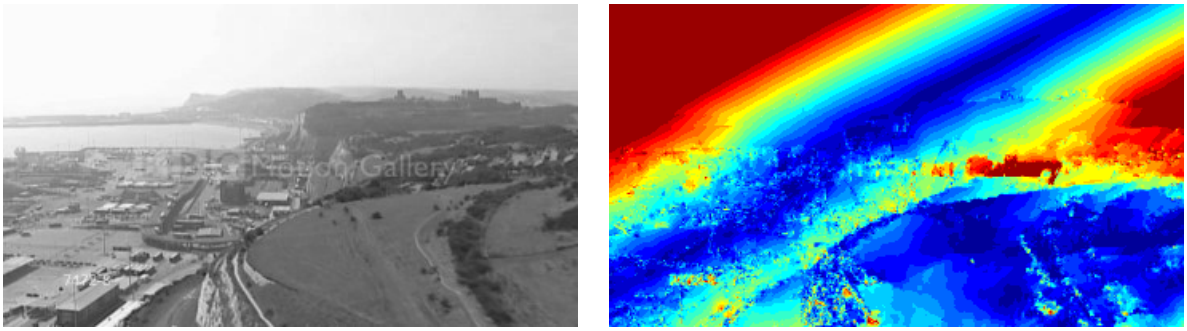


FIG. 3.2 – *Résidus flot et image selon le modèle de mouvement global paramétrique.* 1*ère ligne : première des deux images et différence image absolue (intensité) avec l'image suivante.* 2*ème,* 3*ème et* 4*ème lignes : recalages affine, quadratique et homographique. La première colonne illustre les normes de flots résiduels, la seconde les différences image absolues (intensité) après recalage. Le recalage global devrait être effectué sur le plan du sol, n'affichant ainsi des résidus de flot que sur les éléments mobiles ainsi que 3D.*



(a) (b)

FIG. 3.3 – *(a) Image comportant plusieurs plans dominants (b) norme du flot résiduel après recalage affine : le modèle affine estimé est un compromis entre les modèles correspondant à chaque plan*

reprenant les notations de (3.2), le nouvel ensemble de paramètres est donné par :

$$\begin{aligned} t_x &= c_1 \\ t_y &= c_2 \\ s &= \frac{a_1 + a_4}{2} \\ \theta &= \frac{a_3 - a_2}{2} \end{aligned} \tag{3.5}$$

Deux paramètres sont inchangés. Il s'agit des paramètres de translation t_x et t_y, correspondant respectivement à un déplacement latéral ou vertical dans le repère image.

Outre ces deux paramètres, le paramètre s est directement lié à l'amplitude d'un zoom éventuel : positif, il marque un zoom "in" ou diminution de focale ; négatif, il signifie au contraire une augmentation de la focale ou zoom "out".

Le paramètre θ, ici exprimé en radians, traduit une rotation de la caméra autour de son axe optique dans le sens trigonométrique, ce qui correspond à une rotation de l'image autour de la projection du centre optique sur le plan image (souvent situé au centre de l'image).

Ces paramètres pris isolément pour chaque couple d'images peuvent être bruités, en raison notamment d'erreurs d'alignement ou de scènes complexes présentant différents plans par exemple. Une étape préalable avant interprétation consiste donc à lisser chaque série de paramètres. En effet, le mouvement caméra est continu, à la fréquence d'échantillonnage d'une séquence vidéo classique (environ 40 ms entre chaque image). Même un décrochage (suite à une rotation rapide de la caméra par l'opérateur lors de la prise de vue) est traduit par une continuité de paramètres sur des images consécutives.

3.3.1 Translation

Les deux paramètres de translation fournissent plusieurs informations. Tout d'abord, leur cumul informe sur la dérive progressive par rapport au point de départ, dans le plan image. L'évolution de ce cumul résume de manière concise le trajet de la région observée au sol dans le repère caméra. Toutefois, en l'absence d'orthorecalage, les dérives calculées en pixels ne peuvent être directement traduites en trajet sur une carte, il faut en effet tenir compte de l'incidence (et des variations de celle-ci au cours de la séquence). La détection de boucles, ou retours sur des régions déjà observées, est une application directe de l'estimation des translations. Si un orthorecalage n'est pas nécessaire, il faut cependant cumuler les transitions dans un repère commun, afin d'annuler les rotations et les changements de focale. La connaissance des modèles affines de transition entre les couples d'images successives permet de prendre en compte ces changements de repère. L'équation suivante traduit la modification des coordonnées de translation dans le repère d'une image j au repère de l'image suivante $j+1$:

$$\begin{pmatrix} t_x^{j+1} \\ t_y^{j+1} \end{pmatrix} = A_j \begin{pmatrix} t_x^{j} \\ t_y^{j} \end{pmatrix} \tag{3.6}$$

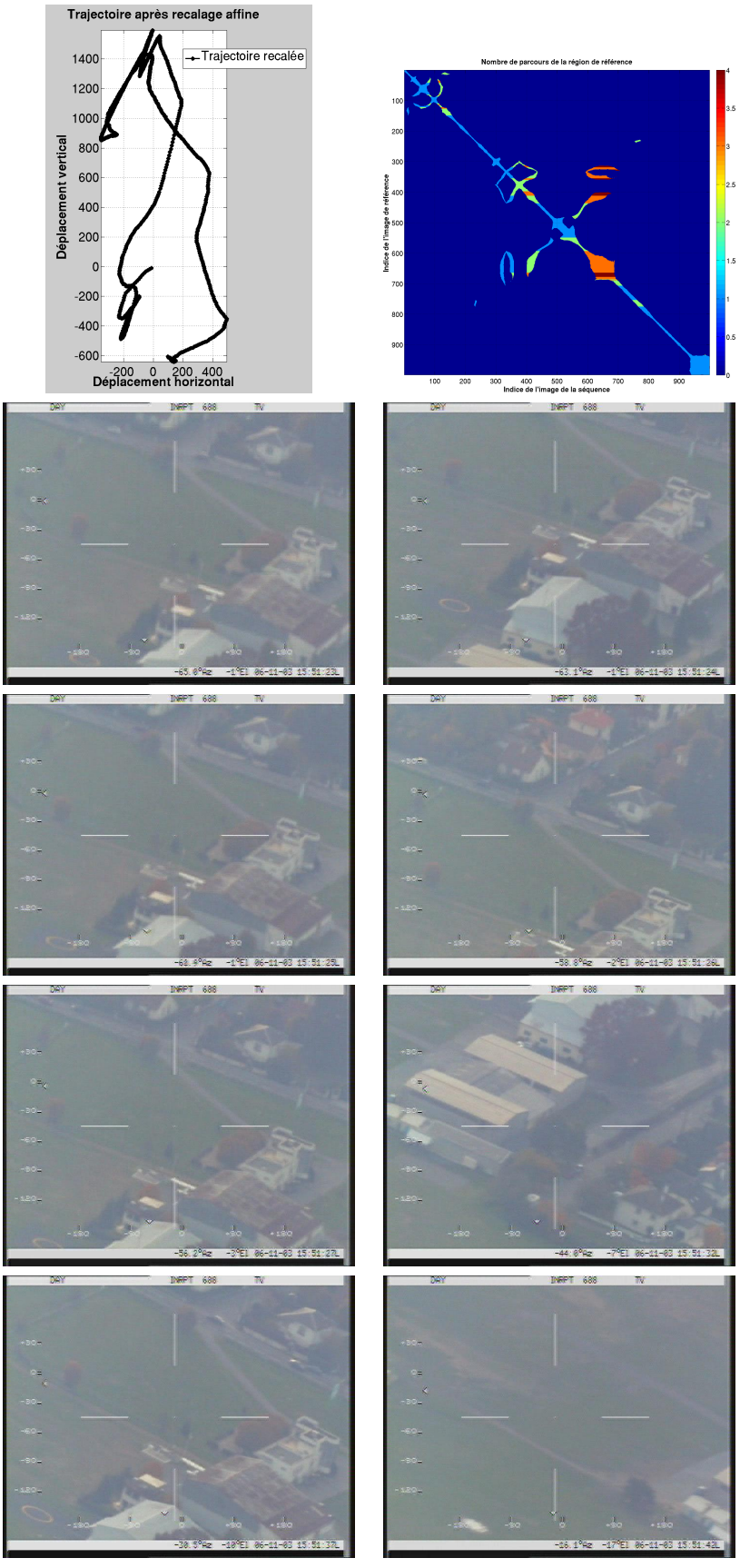
où (t_x^{j+1}, t_y^{j+1}) et (t_x^j, t_y^j) représentent les coordonnées de la translation en pixels de l'image j à $j+1$, respectivement dans les repères des images j et $j+1$, et A_j la matrice d'affinité permettant de passer de j à $j+1$ (plus spécifiquement, les deux premières lignes et colonnes de cette matrice, la translation n'intervenant pas ici). En reprenant ces notations, on peut écrire en composant les affinités :

$$\begin{pmatrix} t_x^{j} \\ t_y^{j} \end{pmatrix} = \prod_{i=0}^{K-1} (A_{j+i})^{-1} \begin{pmatrix} t_x^{j+K} \\ t_y^{j+K} \end{pmatrix} \tag{3.7}$$

Le risque est de cumuler les erreurs au fur et à mesure de la séquence lors du produit des affinités. Une correction périodique en calculant par exemple l'affinité pour un écart temporel de 10 images, à supposer que le recouvrement entre les deux images ainsi séparées reste important, peut permettre de limiter cette dérive. La figure 3.5 montre l'évolution des paramètres cumulés de translation selon les deux dimensions, respectivement sans (en rouge) et avec (en bleu) recalage. Les images dont les ordonnées sont associées à plusieurs segments temporels distincts dans les deux dimensions à la fois correspondent à des régions parcourues plusieurs fois. La détection de ces segments temporels de "retour sur région" dépend de plusieurs paramètres de sensibilité. En effet, il faut définir un critère permettant d'affirmer qu'il y a, ou non, un tel retour. Il peut s'agir de l'aire du recouvrement entre deux images considérées, mais un tel critère est peu adapté car il intègre difficilement les changements de focale ou les rotations. Une simple comparaison des paramètres cumulés de translation, qui revient à comparer les positions des centres des images, est plus simple et plus robuste. Le seuil de distance (et le choix de la norme, L^1, L^2 ou L^∞) peut alors être ajusté suivant la sensibilité de l'interprète, le processus de recherche étant quasi immédiat. La figure 3.4 montre la trajectoire du point central après recalage affine dans le repère de la première image de la séquence (première ligne (a)). Une analyse graphique avec sélection manuelle des points de retour est également envisageable : les retours correspondent aux croisements de trajectoire et sont aisés à détecter pour un interprète humain. La matrice (première ligne (b)) traduit sous une autre forme les nombres de passages (en bleu clair, 1, en vert, 2, en orange, 3 et en rouge foncé, 4) : chaque numéro de ligne de la matrice correspond à l'indice d'une image de référence et les numéros de colonnes aux indices de premier passage ou de retour sur la zone correspondant à cette image de référence (au critère de distance près). Ainsi, la zone de l'image 321 est associée à 4 segments temporels différents (lignes 2 à 5, colonne (a)). Un autre paramètre est nécessaire, de sensibilité temporelle cette fois : il s'agit du nombre d'images consécutives maximal d'images de dérive (ne satisfaisant pas le critère de distance) autorisées pour ne pas considérer qu'il y a eu abandon de la région, ou encore du nombre minimal d'images de dérive entre deux passages sur une même zone. Ce paramètre temporel permet d'être flexible en regard d'oscillations rapides de la caméra, lors de stabilisation par exemple. En effet, ces oscillations créeraient artificiellement des "retours sur zone" sinon, notamment avec un paramètre de sensibilité spatial réduit. Les paramètres pris dans l'exemple montré figure 3.4 sont respectivement 50 pixels en norme L^∞ et 25 images (soit une seconde) en temps.

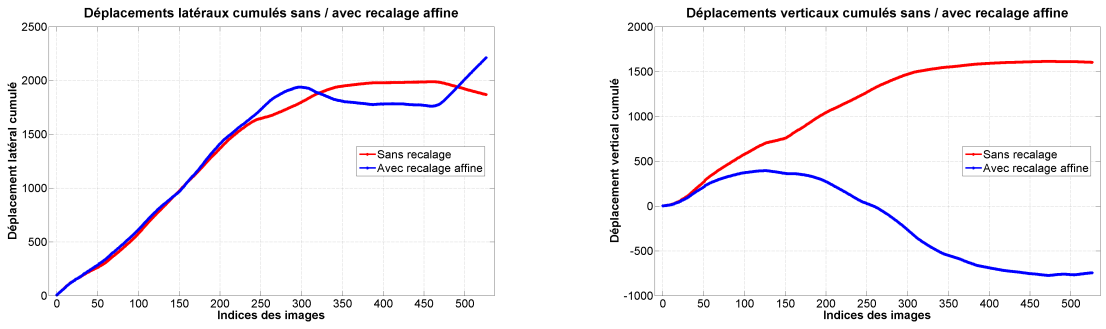
Les translations cumulées permettent également d'extraire les segments temporels "fixes", pendant lesquels la région observée au sol évolue peu. Ces segments correspondent à une suite continue d'images dont l'écart translationnel à un centre de référence ne dépasse pas un certain seuil (par exemple le quart voire une portion inférieure de la dimension minimale de l'image). Le centre de référence peut être pris sur l'image centrale du segment recherché, auquel cas la mesure de stabilité est calculée sur chaque dimension comme le nombre maximal d'images antérieures et postérieures vérifiant la contrainte sur l'écart défini ci-avant. La mesure peut aussi ne pas être centrée, auquel cas le nombre d'images considérées est la somme du nombre d'images antérieures et du nombre d'images postérieures satisfaisant la contrainte. Les courbes rouge et noire de la première figure de 3.6 donnent pour chaque dimension spatiale ces deux mesures de stabilité. Il suffit ensuite d'agréger les deux dimensions en considérant le minimum des deux mesures. Nous avons choisi d'utiliser les mesures centrées (courbe en turquoise), le choix de la mesure non symétrique étant également possible et conduisant à des segments "fixes" détectés plus importants.

La norme et l'angle de la translation fournissent plusieurs critères de classification du mouvement caméra. Tout d'abord, des mouvements rapides ou au contraire l'absence de mouvement translationnel correspondent à une norme importante ou au contraire faible.



(a) (b)

FIG. 3.4 – 1*ère ligne : déplacements horizontaux et verticaux après recalage affine dans le repère de la* 1ère *image (a) déplacements (b) retours sur zone déjà explorée : pour chaque image de référence (ligne) et région associée, la couleur donne l'indice du passage sur cette région.* 2*ème à* 5*ème ligne : (a) images de plusieurs passages sur une même région à moins de* 50 *pixels (images* 321,355,404 *et* 656*) (b) images intermédiaires montrant un déplacement de caméra avant le retour (après pour la dernière) (images* 338,380,530 *et* 800*)*



(a) (b)

FIG. 3.5 – *Paramètres cumulés de translation. Rouge, sans recalage. Bleu, après recalage affine avec comme repère de référence celui de la première image. (a) horizontale (b) verticale*

Plusieurs précisions sont alors nécessaires, le choix du repère et la détermination de seuils haut et bas.

L'amplitude apparente du mouvement lors de la visualisation est associée au repère image qui évolue au cours du temps. Il apparaît donc pertinent de retenir ce dernier plutôt qu'un repère associé à une image et échelle de référence pour l'interprétation de la norme. En effet, si la séquence présente un zoom important par rapport à l'image de référence, un mouvement rapide après zoom ne sera plus considéré comme tel après recalage dans le repère de référence. Les seuils sont subjectifs et ne peuvent être choisis arbitrairement. Ces valeurs dépendent également de la taille de l'image, un déplacement global de 2 pixels n'ayant pas le même ressenti visuel sur une image de dimension représentative 100 ou 2000. Une norme faible sur un segment temporel peut aussi être un signe de stabilité, avec une observation prolongée d'une région au sol. Il faut alors vérifier que la dérive cumulée reste marginale.

3.3.2 Autres types de mouvements

Le paramètre θ (équation (3.5)) traduit le sens et l'intensité d'une rotation de l'image. Il marque donc une évolution de l'orientation en modifiant un repère de direction (le nord par exemple). Si une rotation inverse des images afin de compenser ces changements d'orientation est aisée à effectuer, cette information est toutefois utile car elle fournit directement les temps auxquels la direction d'observation change, afin par exemple de suivre une route ou un véhicule en translation rectiligne.

Le paramètre s traduit une évolution de la focale ou la présence de zoom "in" (valeurs positives) ou "out" (valeurs négatives). La résolution est un élément important lors de l'analyse vidéo, en particulier lors de la recherche de détails, car elle permet de sélectionner rapidement les segments temporels susceptibles de fournir des images de haute résolution. Cela est particulièrement utile dans un but de caractérisation voire l'identification d'objets d'intérêt de petite taille (ou de détails de structures plus importantes). En contrepartie, une grande focale entraîne également un angle de vue réduit. Le parcours d'une région à observer donnée est plus long et les éventuelles informations dynamiques de la scène hors du champ caméra sont alors perdues.

3.3.3 Classes

D'autres classes de mouvement peuvent être déduites de ces quatre séries de paramètres. Ainsi, un mouvement de transition dont la direction varie rapidement, conséquence possible de mouvements brusques de la caméra, est difficile à suivre manuellement et représente une qualité de visualisation amoindrie. Là encore, il est nécessaire de définir un seuil (dans le but d'une classification binaire) ou de laisser à l'interprète le choix de

sélectionner des segments plus ou moins stables (par un curseur par exemple).

Les déplacements dans une direction continue et en l'absence de zoom ou de rotation peuvent être regroupés en une classe de "translations stables". L'échelle temporelle d'intégration des paramètres doit être suffisante pour correspondre à des mouvements de durée significative, de quelques secondes au minimum, mais également suffisamment courte pour ne pas rater ces mouvements : un zoom important peut être effectué en quelques secondes à peine, de même qu'un mouvement rapide ou chaotique, ou encore une rotation de plusieurs dizaines de degrés. Il est de toute façon simple de laisser plus de latitude à l'interprète dans le choix des seuils de durée ou d'intensité pour les différent paramètres ou combinaisons de paramètres.

3.3.4 Mouvement global et intention de l'opérateur

Nous avons choisi d'ajouter à ces classes élémentaires de mouvement deux types de classes "composées" :

- d'une part, des mouvements plus lents, plus difficiles à discerner à l'œil nu sur un nombre réduit d'images, mais qui conduisent sur une durée plus longue à un changement significatif : ainsi, un zoom progressif d'un facteur 2 ou une rotation de 90° sur plusieurs dizaines d'images ;

- d'autre part, des mouvements susceptibles de traduire une intention de l'opérateur ayant manipulé la caméra lors de l'enregistrement de la séquence : une compensation du mouvement du porteur afin de verrouiller une région d'observation particulière, un zoom suivi d'une stabilisation de la caméra sont des signes potentiels d'un intérêt, dans la recherche d'informations dynamiques (arrêt sur région) ou de précision spatiale (zoom et stabilisation).
  Un zoom "out" en revanche peut témoigner d'une préférence pour une couverture élargie dans la recherche de contexte plus exhaustif et / ou une mobilité accrue.
  Enfin, un panorama, pas forcément latéral, pourrait signifier la volonté de couvrir une région étendue sans sacrifier à la précision, ou encore de suivre un véhicule ou autre objet d'intérêt en translation rectiligne.

3.3.5 Résumé des classes choisies

L'ensemble des classes choisies peut être résumé comme suit, chaque classe étant subdivisée en plusieurs types de mouvements :

- zoom - dezoom : il s'agit d'un zoom "in" ou "out" visible, qu'il s'agisse du seul mouvement remarquable ou non ;
- roll trigo - antitrigo : il s'agit de rotations dans le sens trigonométrique horaire de l'image, qu'il s'agisse du seul mouvement remarquable ou non ;
- parasites (translation rapide) - parasites (cahots) : il s'agit de mouvements "rejet", correspondant a priori à des images difficiles à analyser : des images en translation rapide ou de "cahots", changements rapides de direction avec mouvement d'une certaine amplitude ;
- Intention_panorama ; Intention_zoom ; Intention_dezoom ; Intention_fixe : il s'agit de mouvements traduisant peut-être une "intention" de l'opérateur, à savoir, respectivement :
  - panorama : il s'agit de translation globale de direction lentement variable (ce mouvement peut être dû au porteur), l'effet obtenu étant celui d'un panorama,

Paramètre	Classe	Critère	Intervalle temporel	Observations
θ_1	roll	$\|\theta\| > \theta_1$	2 images	sens direct ou indirect (trigonométrique)
θ_2	roll (lent)	$\|\theta\| > \theta_2$	2 images	visuellement apparent sur un intervalle de temps > 1 seconde
s_1	zoom in / out	$\|s-1\| > s_1$	2 images	
s_2	zoom in / out (lent)	$\|s-1\| > s_2$	2 images	visuellement apparent sur un intervalle de temps > 1 seconde
D	mouvements parasites (rapides)	$\sqrt{t_x^2 + t_y^2} > D$	2 images	
σ	mouvements parasites (chaotiques)	$var\left(atan(\frac{t_y}{t_x})\right) > \sigma$	1 seconde	variations brusques de la direction du mouvement de translation
d	fixe (suivi de zone)	$\forall t \in [t_{min}, t_{max}], \|\sum_{t'=t_0}^{t} t_x\| + \|\sum_{t'=t_0}^{t} t_y\| < d$	$>$ 2 secondes	cf. paragraphe 3.3.1 et figure 3.6, 1^{re} image
α	panorama	$var\left(atan(\frac{t_y}{t_x})\right) < \alpha$	$>$ 2 secondes	critères supplémentaires : pas de changement de zoom in/out ni de rotation

TAB. 3.1 – *Ensemble des paramètres utilisés lors de l'établissement de la classification. La quatrième colonne précise l'intervalle temporel (nombre d'images consécutives) considéré pour le calcul du critère (troisième colonne).*

- zoom - dezoom : il s'agit d'un zoom comme mouvement dominant, les autres mouvements (roll, translation) étant absents ou faibles. Ces mouvements peuvent traduire un intérêt de l'opérateur pour une zone particulière : recherche de détails (zoom "in") ou revenir à un champ plus large pour avoir une vue d'ensemble (zoom "out"),
- fixe : l'opérateur tente de fixer une scène particulière, afin d'observer dans le temps cette zone.

La figure 3.6 montre un exemple de telle classification automatique sur la séquence "BBC". Les différents paramètres utilisés sont résumés dans le tableau 3.1.

Les différentes classes correspondent à des échelles temporelles différentes et ne sont pas exclusives. Il est nécessaire de régulariser les résultats de classification obtenus dans le temps. En effet, dans un objectif d'indexation par segments temporels, la multiplication de segments très courts (de l'ordre d'une seconde ou moins) n'est pas souhaitable. Une requête sur une classe particulière de mouvement renverrait en effet un nombre conséquent de segments difficilement et fastidieusement interprétables, de quelques secondes voire de durée inférieure. La réelle difficulté de la classification réside dans le choix des seuils : un compromis doit être trouvé entre intervention de l'interprète d'une part et vitesse d'une interprétation automatique d'autre part, avec paramètres fixés mais moins flexible en regard du contexte de la séquence vidéo et de la sensibilité de l'interprète aux différents mouvements.

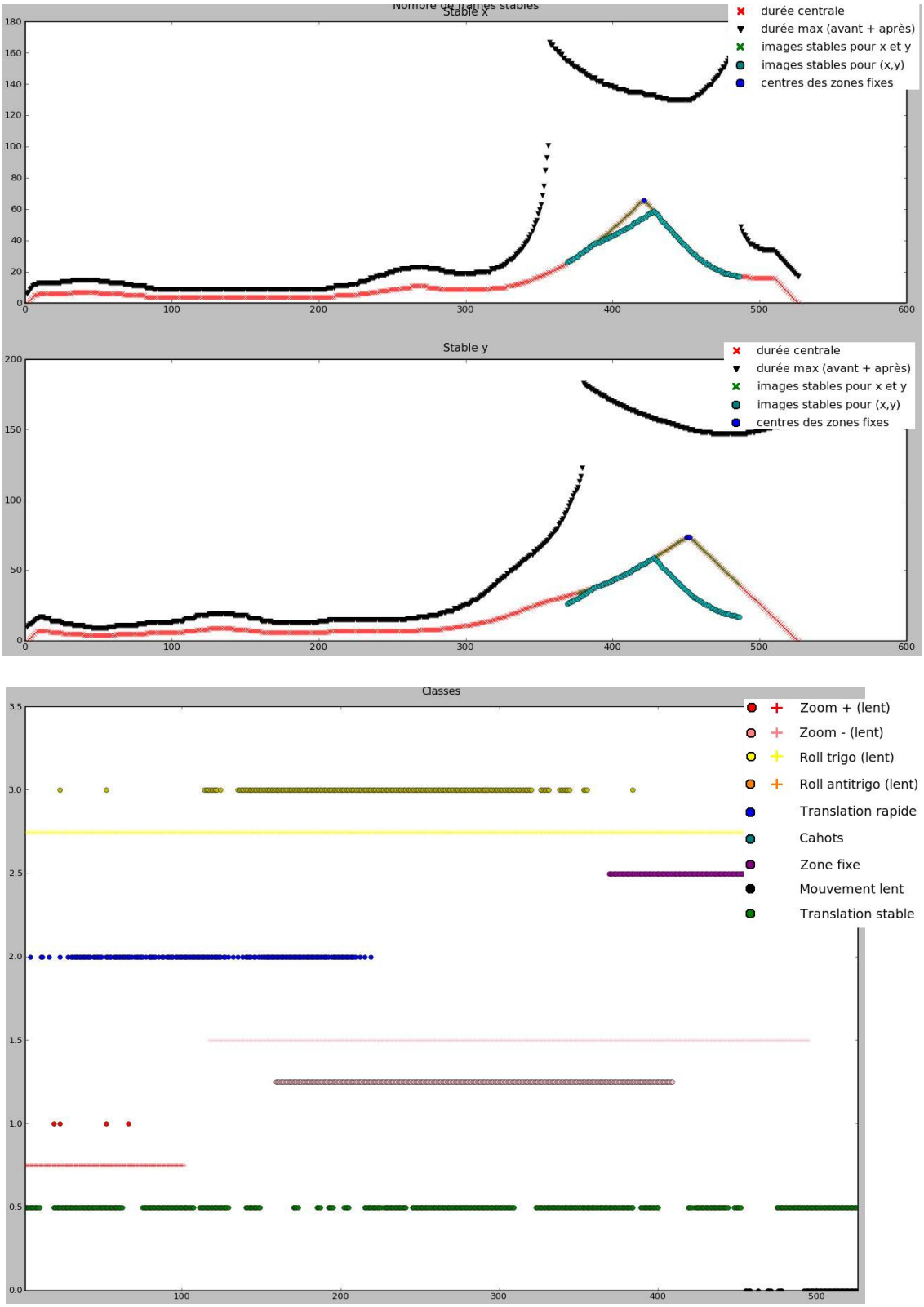


FIG. 3.6 – *Séquence "BBC". 1ère image : détection des segments temporels "fixes" avec peu ou pas de mouvement. Haut : mouvement horizontal, bas : mouvement vertical. Noir : nombre d'images consécutives comprenant l'image actuelle et de mouvement cumulé inférieur à un seuil. Rouge : nombre maximal d'images précédentes ET consécutives à l'image actuelle vérifiant simultanément cette condition sur le mouvement cumulé. Turquoise : minimum des courbes rouges selon les deux dimensions. 2ème image : classes de mouvement. Dehaut en bas : rotation (jaune), violet (fixe), bleu (mouvement rapide), rose et rouge (zoom out et in), vert (translation stable) et noir (mouvement lent).*

3.4 Évaluation

L'évaluation de la classification obtenue est délicate car elle relève plus d'une interprétation subjective du mouvement global que de valeurs numériques 'objectives', par exemple les valeurs de flou optique. La "vérité terrain" va donc dépendre pour la classification de l'interprète humain concerné. Il est préférable de disposer d'un panel large d'interprètes afin d'établir une "vérité terrain" moyenne ainsi que l'incertitude correspondante, et ce pour chaque classe. Plusieurs options souhaitables comprennent :

- une caractérisation la plus complète possible de chaque classe afin de minimiser la part d'interprétation subjective,
- un ensemble de séquences présentant des mouvements variés et de complexités différentes,
- un nombre important d'évaluateurs,
- une interface d'annotation.

Les classes à annoter correspondent aux classes retenues pour la classification automatique et détaillées au paragraphe 3.3.5.

Des exemples de séquences ainsi qu'une interface d'annotation et visualisation ont été fournis à plusieurs expérimentateurs afin d'établir une vérité terrain. La figure 3.7 montre les différentes composantes de l'interface (avec un exemple fictif d'annotations). La comparaison des différentes annotations sur plusieurs séquences vidéo fait apparaître plusieurs points :

- certaines classes sont peu sujettes à confusion : zoom ("in" ou "out"), rotations et translations stables (direction constante ou légèrement variable),
- les mouvements "rapides" sont annotés avec plus de variation, ainsi que les mouvements "interprétés",
- des classes sont ambigües lors de mouvements complexes,
- les seuils de mouvement significatif (zoom ou rotation notamment), en-dessous desquels le mouvement n'est pas annoté comme tel, varient suivant l'interprète.

Les mêmes points reviennent lors de la comparaison avec les résulats obtenus automatiquement à partir des paramètres affines comme décrit ci-dessus. Il est envisageable de laisser alors à l'interprète le choix du seuil, en observant directement par l'intermédiaire de l'interface visuelle les segments correspondants. Cela s'applique également pour le seuil "faible" correspondant à un mouvement (zoom "in" par exemple) progressif visible au bout de plusieurs secondes.

La granularité temporelle est encore un autre paramètre subjectif : à partir de quelle durée un segment doit-il être retenu ou au contraire éventuellement absorbé dans un segment plus important ? Cela dépend entre autres du type de mouvement, un zoom "in" rapide devant être affiché même si de durée très courte, une stabilisation de quelques images pouvant être au contraire ignorée.

Enfin, des ambiguïtés subsistent toutefois en présence de mouvements complexes, divisant même les interprètes humains. Les principales sources d'incertitude sont les effets en trois dimensions d'une part, une incidence rasante d'autre part. Le mouvement du porteur peut alors donner l'impression de zoom. Dans les faits, il s'agit de mouvement le long de l'axe optique, mais le rapprochement vis-à-vis des bâtiments et autres structures en trois dimensions présente un effet similaire au zoom. De même, le rapprochement du porteur d'un bâtiment associé à un "tilt" (rotation autour de l'axe latéral de la caméra traduit par un déplacement vertical de l'image) agrandit le bâtiment mais il n'y a pas de zoom. Des exemples de ces séquences ambigues sont donnés figure 3.8 (respectivement première et

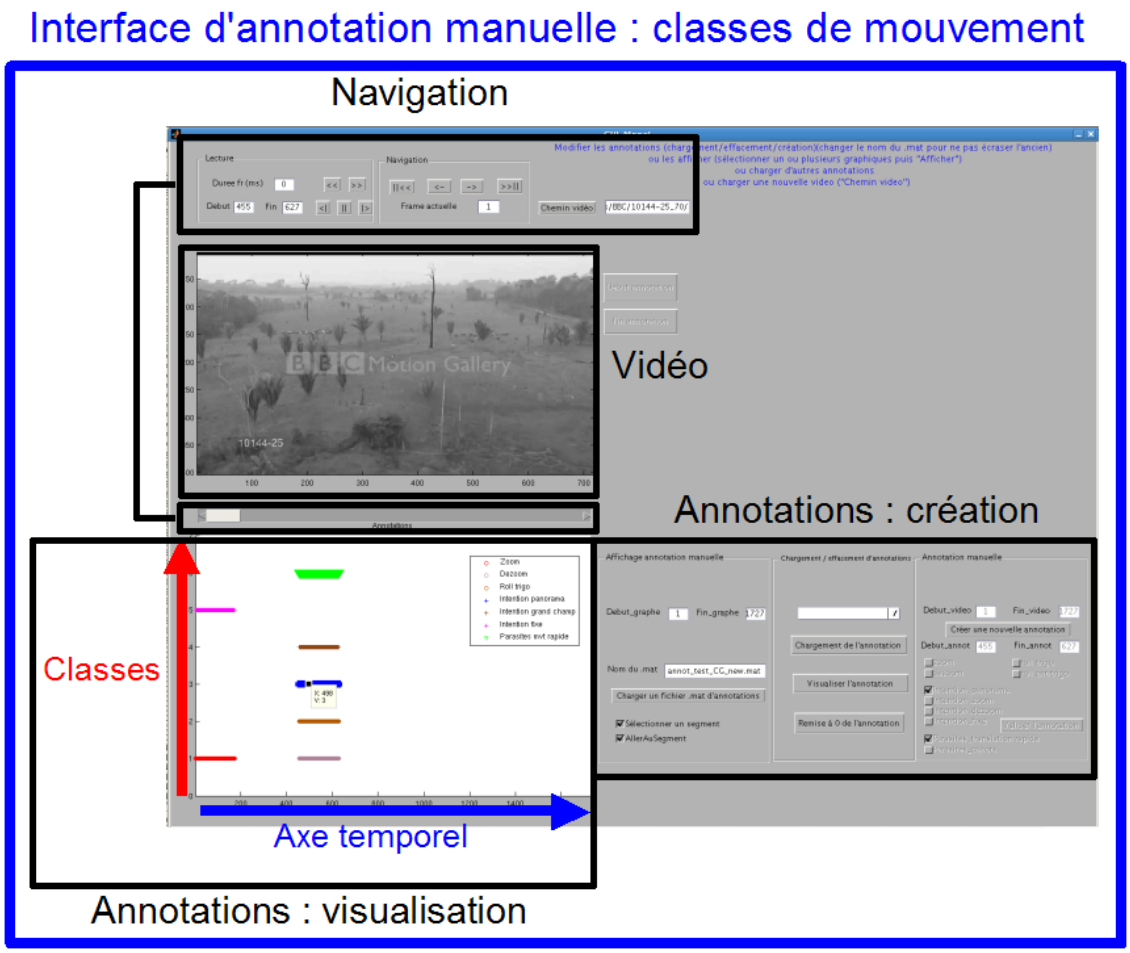


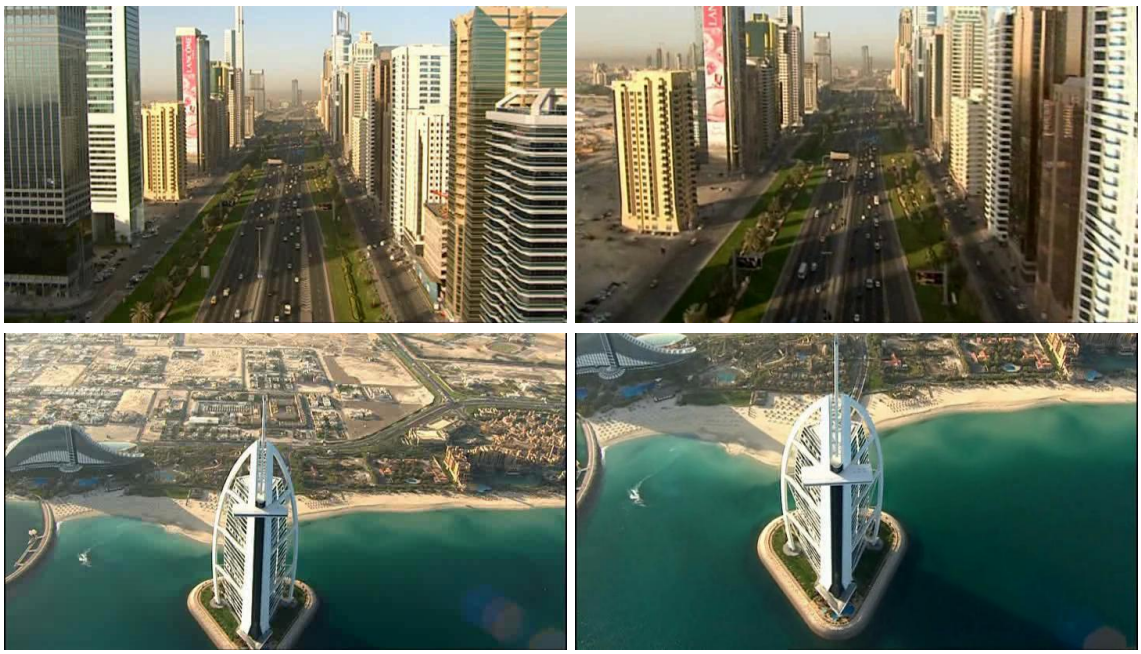
FIG. 3.7 – *Interface d'annotation et de visualisation des segments. La partie haute permet de naviguer et visualiser la vidéo. La partie basse permet de créer ou modifier (à droite) les annotations de classes de mouvements et de visualiser ou sélectionner (à gauche) les segments et classes associées.*

deuxième ligne).

La classification de séquences vidéo dépend donc de plusieurs paramètres de sensibilité qui peuvent être ajustés selon l'interprète. Une analyse automatique plus sémantique des classes obtenues est possible mais ne peut fournir de solution unique. Il s'agit donc d'une modalité d'indexation à utiliser en complément d'autres algorithmes ou comme aide à la navigation.

3.5 Utilisation des classes de mouvement global pour la création de résumés vidéos

Dans un but d'efficacité temporelle, il est possible d'utiliser les classes de mouvement global comme index de navigation et des paramètres intermédiaires dans la création de résumés vidéo sous des formes diverses. La méthode la plus simple consiste à naviguer dans la séquence en sélectionnant les segments au sein d'une interface homme-machine telle que celle présentée figure 3.7. Outre la simplicité de cette représentation, la fidélité visuelle est conservée par rapport à la séquence originale. En revanche, la compression temporelle est réduite. Une première idée pour améliorer la compression est de résumer chaque segment sous une forme adaptée aux classes correspondantes. Ainsi, une vue d'ensemble sera disponible pour chaque segment vidéo, ce qui évitera son parcours intégral. Le choix de la forme du résumé est inspiré de [172] :



(a) (b)

FIG. 3.8 – *Exemples de mouvements ambigus.* 1*ère ligne : pan, tilt ou zoom ?* 2*ème ligne : tilt ou zoom ?*

- pour les segments de zoom ("in" ou "out"), la première et la dernière image du segment (voire des images intermédiaires si la variation d'échelle est importante) ;
- pour les segments stables, une mosaïque ou une image choisie selon des critères de qualité ;
- pour les translations continues, des panoramas ou mosaïques sont un choix logique, il faut choisir alors le repère de référence, soit une image de la séquence, soit un repère orthorecalé ;
- les mouvements parasites sont a priori filtrés mais s'il s'agit des seuls segments couvrant la zone associée, ils restent porteurs de contenu même dégradé. Là encore, le segment peut être échantillonné suivant un critère de qualité ou compressé en une mosaïque, notamment pour les mouvements à forte oscillation : la source d'inconfort à la visualisation disparaît avec une représentation statique (le contenu dynamique en revanche est perdu et doit être recouvré autrement) ;
- les mouvements complexes sont plus délicats à compresser. Un échantillonnage uniforme peut être suffisant, avec une compression temporelle plus faible.

Une autre possibilité pour résumer l'intégralité d'une séquence vidéo consiste à créer un objet spatio-temporel global, dans lequel il faut intégrer les différentes formes de résumé. Une possibilité simple consiste à fabriquer un résumé vidéo à partir d'un échantillonnage irrégulier sans recalage dont la fréquence dépend des classes de mouvement. Le résumé peut aussi être réalisé à partir d'images recalées et devient une mosaïque dynamique, ce qui améliore la visualisation en supprimant le recalage visuel humain. Suivant le principe de [206], il est envisageable d'adapter les différents termes afin de prendre en compte non plus l'activité mais le type de mouvement, pour créer un résumé équilibrant compression temporelle, clarté de la visualisation et cohérence temporelle.

3.6 Conclusion et perspectives

L'efficacité de la classification du mouvement global comme modalité d'indexation de séquences vidéo dépend de plusieurs facteurs. Le choix du modèle paramétrique revient à un compromis entre précision et robustesse (à des erreurs de flot optique ou des dis-

torsions radiales par exemple). Hormis des cas exceptionnels (des scènes sans plan de référence majoritaire dans l'image), le modèle affine représente un compromis efficace qui peut être simplement transformé en des variables décrivant des mouvements élémentaires (translation, changement d'échelle, rotation image).

Afin de présenter un intérêt opérationnel, la classification doit toutefois fournir des segments temporels de durée "significative" (de quelques secondes au minimum, pour la plupart des mouvements). Il est ainsi nécessaire de régulariser dans le temps les mouvements élémentaires obtenus pour chaque image. Certaines classes de mouvements, tels que des changements d'échelle lents ou des mouvement "chaotiques" (changements brusques et répétés de direction) ne peuvent d'ailleurs être définis que sur des segments temporels et non localement.

Une comparaison des résultats obtenus automatiquement avec une classification manuelle fait apparaître deux ensembles de mouvements (ou classes de mouvements). Les mouvements élémentaires de translation, de changement d'échelle et de rotation sont bien reconnus. Les mouvements plus complexes, combinaisons de mouvements élémentaires, sont plus ambigus. Une solution consiste à les regrouper en une classe de "mouvements complexes" sans distinction.

Enfin, la subjectivité intervient en différents points et ne peut être négligée. En effet, les amplitudes des mouvements élémentaires étant des variables continues, comment fixer les seuils au-delà desquels ces mouvements seront considérés significatifs ? À partir de quelle fréquence temporelle et quelle amplitude d'écarts dans les directions de translation peut-on considérer un mouvement comme "chaotique" ? La définition de classes de mouvement traduisant une intention de l'opérateur caméra au moment de la prise de vue est également très subjective.

Les ambiguïtés de la classification du mouvement global ainsi que les difficultés d'établir une classification unique étant donnés les différents éléments de subjectivité mentionnés (sensibilité, définition de mouvements complexes) conduisent à définir plusieurs perspectives.

- Certains mouvements élémentaires tels qu'illustrés par la figure 3.8 sont parfois ambigus, selon le champ de vue et l'incidence. L'utilisation de primitives supplémentaires (profondeur, lignes de fuite...) pourrait aider à réduire ces ambiguïtés.

- La subjectivité peut être traitée selon deux visions opposées. D'une part, imposer des seuils (pour l'amplitude des mouvements élémentaires) et des définitions de classes de mouvement traduisant une intention de l'opérateur caméra, évite l'introduction de nouveaux paramètres à ajuster, mais contraint un interprète à suivre les *a priori* du créateur de l'algorithme. Une approche plus proche de l'interprète consisterait à laisser à ce dernier plus de liberté dans la définition des seuils et de classes "interprétées" à partir des mouvements élémentaires, en ajustant des seuils numériques ou par apprentissage.

- Outre l'indexation en segments temporels, l'analyse du mouvement global pourrait être intégrée à la création de résumés spatio-temporels (résumés vidéo, mosaïques dynamiques, compositions statiques d'images-clefs...) adaptés aux classes de mouvements obtenues.

PARTIE 2
Détection d'activité

Détection d'activité

4

Le "contenu" vidéo peut se révéler particulièrement complexe : une zone urbaine dense comporte ainsi un réseau routier, des bâtiments ou des structures en trois dimensions de formes et de tailles variées, des objets mobiles (principalement véhicules ou piétons), des éléments naturels (végétation, points d'eau par exemple)...

Le contenu statique peut être extrait des images prises séparément. Ainsi, les méthodes de détection, de reconnaissance et d'identification d'objets sur des images sont nombreuses et permettent de décomposer la scène observée en objets et structures d'intérêt. Toutefois, la dimension temporelle de la séquence vidéo permet de réaliser au préalable des mosaïques ou d'autres résumés spatiaux, et d'extraire par la suite le contenu statique sur ces résumés. Cela évite une analyse de contenu effectuée à chaque image et redondante. En revanche, il faut pour cela disposer de l'intégralité de la séquence et une telle approche n'est donc pas adaptée à une analyse "en ligne" au fur et à mesure de la prise de vue. La constitution du résumé doit également ne pas entraîner de diminution notable de qualité image, ce qui réduirait la richesse d'exploitation du contenu statique.

L'interprétation sémantique de la scène découle alors des relations spatiales reliant les différents éléments détectés ou reconnus ainsi que de règles de connaissance introduites manuellement et spécifiques au contexte d'application. Des alertes peuvent alors être déclenchées suivant le respect ou non de certaines règles prédéfinies. Des modèles de comportements d'objets ou de personnes peuvent être aussi appris, directement à partir des données ou comme combinaisons logiques d'actions élémentaires. Ces modèles, ou des écarts aux modèles, signes potentiels des comportements inhabituels, peuvent être alors reconnus.

Détection d'activité : signification et intérêt La dimension temporelle des données étudiées, en l'occurrence des séquences vidéo, permet de dépasser le stade de la détection d'objets sur images fixes. En effet, la fréquence temporelle élevée entraîne une redondance élevée du contenu, en particulier du contenu statique. Il est alors souhaitable d'établir un modèle de ce contenu statique ou fond avant d'en extraire les objets d'intérêt.

Cette dimension temporelle permet en outre un nouveau mode d'analyse de flux vidéo, l'extraction des segments d'activité, c'est-à-dire ici comportant des objets mobiles. Elle apporte ainsi un éclairage supplémentaire sur la séquence en soulignant la dynamique de celle-ci. Les informations recueillies peuvent être regroupées à plusieurs niveaux sémantiques : présence de mouvement ; trajectoires ; scénarios reliant plusieurs trajectoires entre elles ou à des structures fixes ou encore des scnéarios définissant des comportements anormaux par rapport à des modèles de référence.

La détection automatique d'activité est associée à plusieurs enjeux. Il s'agit d'une part de faciliter l'indexation d'une séquence vidéo par son contenu, en l'occurrence le contenu dynamique.

D'autre part, la détection automatique d'activité permet de réduire le temps nécessaire à l'analyse de séquences vidéo. Si l'une des tâches dévolues à un interprète humain concerne par exemple la détection de comportements suspects d'individus ou de véhicules, une

première restriction de la visualisation aux seuls segments contenant de l'activité apporte un gain de temps d'autant plus élevé que les segments d'activité sont rares. Cela favorise ainsi une concentration accrue pour l'analyse des segments d'intérêt extraits.

L'intérêt peut sembler moindre pour des environnements particulièrement fréquentés, tels des halls de gares. Dans ce type d'application néanmoins, les zones spatiales et les plages horaires d'activité évoluent également, et la détection de ces volumes spatio-temporels constitue une étape préliminaire pour une analyse plus fine de l'activité. Les capteurs étant fixes, il est en effet possible de construire des modèles d'activité habituelle ou "normale" afin de détecter par contraste des activités non usuelles.

Notre étude concerne plus particulièrement des séquences vidéo aériennes comportant des déplacements conséquents de la caméra. En dehors de quelques cas précis, tels que la surveillance de la circulation d'une portion définie de réseau routier ou de ville, les segments d'activité seront donc plus limités en nombre. La détection d'activité apparaît alors particulièrement intéressante comme modalité d'indexation.

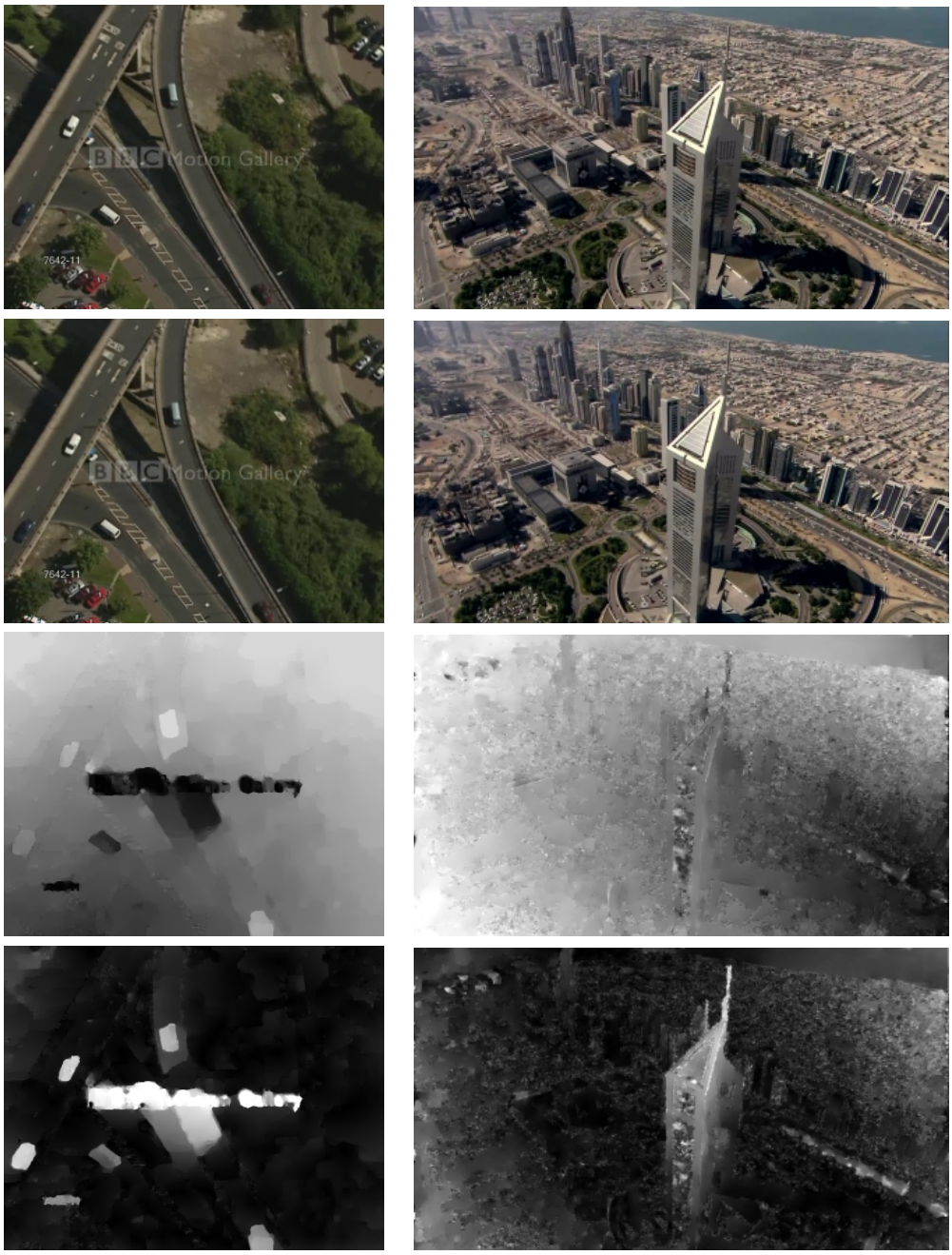
Problématique : données aériennes, recalage et associations Les difficultés d'une détection automatique dans notre cadre applicatif, le traitement de séquences vidéo aériennes, sont multiples. Les données étudiées seront présentées plus en détail dans la section 4.3. Le caractère mobile de la caméra permet d'observer une zone plus étendue et apporte une certaine flexibilité lors de la prise de vue : il est par exemple possible de revenir sur une zone d'intérêt, changer la focale afin d'agrandir le champ de vue ou au contraire améliorer la visibilité de détails. Il introduit en revanche un mouvement global qui doit être compensé, ou qui complique sinon singulièrement la détection d'activité. Ainsi, les exemples des séquences "BBC" et "Dubai Palace" (figure 4.1, respectivement (a) et (b) sur la 3ème ligne) montrent un flot optique de norme importante et variable sur le fond de l'image. Les zones de protection digitale ou "watermarks" sur la séquence "BBC" correspondent à un flot nul : les "watermarks" sont en effet immobiles dans le repère image.

Un recalage affine permet de compenser ce mouvement global. Le résultat est particulièrement propre pour la séquence "BBC", sur la quatrième ligne (a) de la figure 4.1. Toutefois, la région située dans la zone centrale de l'image, en-dessous de la "watermark", est associée à un flot résiduel non nul car le flot original a été ici mal estimé.

Les primitives associées au mouvement jouent un rôle important dans la détection de régions ou objets en mouvement. Une telle erreur est ainsi susceptible d'entraîner des fausses alarmes. Inversement, un lissage du flot supprimant les variations locales dues à un mouvement d'objet entraîne des défauts de détection.

Le recalage global crée également, en présence de structures en trois dimensions de grande hauteur par rapport à l'altitude du porteur de la caméra (avion, hélicoptère ou drone), des effets de parallaxe qui peuvent être confondus avec des éléments en mouvement. Cela est typiquement le cas lors du survol par hélicoptère ou drone à basse / moyenne altitude de zones citadines comportant des bâtiments de grande hauteur ou plus généralement la présence de relief. La norme du flot résiduel après compensation affine du mouvement global sur la séquence "Dubai Palace", cf. figure 4.1, 4ème ligne (b) est particulièrement importante sur la plus haute tour (en profondeur relative à la caméra). Cet exemple montre également les limitations d'un tel recalage avec un flot résiduel important sur les bords de l'image, signe erroné d'activité.

Approche locale ou globale en temps La détection des images d'une séquence vidéo présentant de l'activité peut être effectuée pour chaque image, à partir des informations de couleur, de texture et de mouvement (obtenues par calcul de flot optique par exemple) ou dans le domaine fréquentiel. L'information peut être propagée au cours du temps par pistage, ce qui nécessite une initialisation et une mise à jour régulière afin de repérer les



(a) (b)

FIG. 4.1 – *Exemples de complications dues au mouvement caméra. (a) déplacement et changement de focale, séquence "BBC". (b) effet de parallaxe sur la tour, séquence "Dubai Palace". 1ère et 2ème lignes : couple d'images successives, 3ème ligne : norme du flot optique estimé, 4ème ligne : norme du flot résiduel après recalage affine.*

éventuels objets mobiles auparavant occultés ou hors du champ de vision.

La séquence peut être enfin analysée comme un volume spatio-temporel sur lequel les "tubes" d'activité sont directement extraits. La présence de structures 3D de profondeur variable (bâtiments, végétation, variations de terrain) brouille cependant la recherche des objets mobiles. En effet, le mouvement image de ces derniers peut être confondu avec les effets de parallaxe induits par le relief. Différentes approches permettent de filtrer la parallaxe, notamment des approches géométriques ou la caractérisation des régions spatio-temporelles dans des espaces particuliers. Les sections 4.1 et 4.2 développent différentes approches de détection d'activité, respectivement locales en temps et faisant intervenir la cohérence temporelle.

Quelle que soit la méthode choisie (locale en temps, avec ou sans propagation, ou globale sur le volume spatio-temporel), des étapes de traitement sont communes :

- le choix des primitives : il peut s'agir directement de l'intensité ou des valeurs RGB des pixels, de descripteurs plus complexes tels que SIFT ou GLOH, de flot optique, de représentation fréquentielle ou de profils en deux dimensions du volume spatio-temporel ;
- l'obtention de "tubes" d'activité correspondant aux trajectoires d'objets ou d'ensembles d'objets avec filtrage de parallaxe ;
- l'interprétation de ces "tubes" : établissement de modèles de trafic, validation de scénarios ;
- l'indexation et compression : l'activité est représentée sous forme réduite afin d'accélérer la consultation.

Devant la diversité des contextes, un apprentissage peut être réalisé pour éviter l'utilisation d'heuristiques trop spécialisées. Cela inclut la nécessité de bases de données adaptées aux données recherchées et la définition d'une méthodologie d'apprentissage (caractéristiques sélectionnées, classifieurs, classes recherchées...)

4.1 Etat de l'art sur les méthodes de détection d'objets mobiles locales en temps

Détection d'objets mobiles avec caméra fixe et cas d'une caméra mobile La détection d'objets ou d'entités mobiles dans le cadre de caméras fixes est traditionnellement obtenue grâce à des méthodes de soustraction de fond [193, 24]. Un autre ensemble de méthodes associe des primitives d'apparence locale (couleur, texture) et de mouvement [46, 148].

Les performances de ces méthodes sont toutefois dégradées lorsque la caméra est mobile. Le mouvement et les variations de focale de la caméra rendent en effet délicates les méthodes évoquées ci-dessus. Notre cadre d'étude concerne plus particulièrement les séquences vidéo aériennes. Dans ce contexte, l'enregistrement des séquences sert souvent différents objectifs complémentaires simultanément : imagerie grand angle, surveillance de région particulière, suivi d'entités mobiles, observation d'une même cible sous divers angles de vue à des fins de caractérisation voire d'identification... Cette polyvalence se traduit souvent par la succession de conditions de prise de vue variées au sein d'une même séquence vidéo et complique la détection robuste d'activité. Les variations de profondeur des structures en trois dimensions présentes dans la scène créent de plus des effets de parallaxe qui doivent être aussi filtrés.

Modélisation géométrique [217, 275, 272] utilisent une compensation du mouvement

global et une modélisation géométrique de la scène afin de filtrer les occultations et la parallaxe.

[217] compare deux méthodes de détection de mouvement avec caméra mobile. Le premier algorithme opère un recalage homographique classique et seuille la différence d'images après recalage. Cet algorithme est efficace pour détecter des objets mobiles de petite taille sur une surface plane ou lorsque le mouvement de la caméra est une simple rotation. En revanche, il n'est pas adapté au cas d'une composante translationnelle importante de la caméra et de structures en trois dimensions qui seront détectées en tant qu'objets.

La deuxième méthode proposée consiste dans un premier temps à estimer le mouvement de la caméra en trois dimensions par un algorithme d'odométrie visuelle. Le mouvement de la caméra est compensé afin d'obtenir le flot optique résiduel. Le mouvement caméra ayant été éliminé, ce flot doit satisfaire la contrainte épipolaire. Les points de l'image dont le flot résiduel s'éloigne au-delà d'un certain seuil de la direction épipolaire sont considérés comme appartenant à des objets mobiles.

Yalcin et al. [275] estiment un flot optique dense au cours du temps dans un cadre bayésien et construisent un modèle d'apparence du fond. Chaque nouvelle image de la séquence aérienne est découpée en une couche de fond et une couche d'activité à partir du flot résiduel dense et du modèle d'apparence de l'image précédente par un algorithme de type espérance-maximisation. Les probabilités de mélange ainsi que les poids d'appartenance aux classes correspondants sont mis à jour à chaque image selon un *a priori* favorisant un mouvement résiduel lent pour le fond par rapport aux objets mobiles. L'ensemble des modèles de fond sert également à créer une mosaïque de fond au fur et à mesure de la séquence.

Xiao et Shah [272] supposent un découpage de la scène en plusieurs régions planes et proposent d'extraire les différentes couches de mouvement. Des points d'intérêt tels que des coins de Harris sont suivis sur un faible nombre d'images afin d'initialiser un ensemble de correspondances.

Les régions ainsi initialisées sont étendues en utilisant un graph-cut sur la représentation en ensembles de niveaux de l'intensité. Les régions obtenues sont ensuite fusionnées en deux étapes, en regroupant dans un premier temps les régions partageant une intersection, puis les régions isolées partageant un même modèle de mouvement affine.

Par le biais de contraintes d'occultation exprimant la continuité des occultations au cours du temps, la segmentation finale en couches de l'image est déterminée par un algorithme graph-cut ainsi que les occultations entre couches superposées. Les modèles de mouvement correspondant à chaque couche ainsi que les informations d'occultation peuvent ensuite être utilisées pour la détection d'entités mobiles.

[277, 279, 220] ont également pour objectif de détecter les objets en mouvement, et traitent simultanément le pistage multi-objets. Toutefois il s'agit de méthodes globales en temps qui nécessitent des trajectoires dans l'espace image ou dans un espace de plus grande dimension afin d'établir un modèle de fond ou un ensemble de pistes distinctes.

Classification Des classifieurs particuliers sont souvent employés afin de détecter ou reconnaître différents objets sur des images fixes. Les primitives choisies dépendent du mode de représentation adopté pour les objets, pour lequel différentes possibilités sont décrites par Yilmaz et al. dans [276]. Sur le même principe, les objets mobiles peuvent être repérés par classification à partir de séquences ou de couples d'images vidéo.

Le choix des primitives dépend également de la catégorie des objets à détecter. Par exemple, l'information de couleur des pixels est peu pertinente pour la détection de véhicules étant donné la grande diversité des couleurs rencontrées. Certains éléments tels

que les vitres ou pare-brise n'ont pas ce problème mais sont transparents ou présentent une réflexion spéculaire, leur teinte peut donc également varier. Outre la couleur, d'autres primitives peuvent inclure des informations de texture, des statistiques de région ou, dans le cas de séquences vidéo, des informations temporelles telles que le flot optique [46], des tranches d'espace-temps XT, YT voire des volumes spatio-temporels XYT [49, 198].

La disponibilité, la diversité et l'étendue des données d'apprentissage peuvent également se révéler problématiques. Liu et al. [146] transfèrent par exemple directement les étiquettes des images les plus semblables d'un ensemble d'apprentissage de grande taille, après alignement. Or l'étiquetage manuel est un processus chronophage et fastidieux. Si plusieurs bases de données existent pour des images fixes, telles que Caltech ou Pascal, les séquences vidéo aériennes de référence ainsi annotées sont beaucoup plus rares voire inexistantes. Une telle base spécifique pourrait toutefois être constituée dans le cadre d'un projet d'évaluation tel que Trecvid [Tre] regroupant différentes tâches d'analyse et de résumé sur des séquences vidéo, ou le site et base d'évaluation de flots optiques Middlebury [15].

Dans [273], Xiao et al. présentent deux approches pour la détection d'objets mobiles. La première consiste en un pistage rapide et adapté à des architectures embarquées utilisant la composante normale du flot pour plus de robustesse aux changements de luminosité et aux zones végétales.

La seconde est plus lente et regroupe une segmentation en couches de la scène et une compensation de mouvement global du fond avant de raffiner les pistes obtenues. Les entités détectées sont ensuite classées en "personnes" et "véhicules" grâce à un ensemble de classifieurs de types "histogrammes de gradients" spécialisés chacun pour la reconnaissance d'une pose (par exemple, une vue arrière de personnes) et taille précises de l'une des deux classes.

Ommer et al. [179] incluent la reconnaissance de différentes classes d'objets comme dernière étape de leur méthode après un pistage et une segmentation d'objets. Il s'agit toutefois d'une approche globale nécessitant de construire des compositions temporelles de points. De plus, les descriptions d'objets par le biais de points d'intérêt requièrent que les objets soient suffisamment résolus. Cette condition n'est souvent pas satisfaite dans le cas de séquences vidéo aériennes.

Informations de contexte Outre les informations de mouvement (flot résiduel, trajectoires obtenues par pistage, modèle paramétrique...) et d'apparence (couleur, texture, gradients), il est possible de faire intervenir des informations de contexte, directement déduites de la séquence ou imposées extérieurement comme modèles de connaissance sémantique de plus "haut niveau". L'idée est de filtrer les fausses détections par l'intermédiaire d'informations externes et complémentaires à l'apparence et mouvement propres du patch ou de la région. Ce contexte est appris sur des bases de données ou inféré directement à partir des données de la séquence par des critères de similarité.

Des méthodes d'apprentissage ont été utilisées afin d'obtenir une segmentation sémantique avec des classes telles que "route", "champ", "végétation"... dans le cadre d'images fixes ou d'étiquetage d'images spatiales. Shotton et al. [224] proposent l'utilisation de forêts de textons sémantiques, soit des ensembles d'arbres de décision agissant directement au niveau pixellique et rapides lors des phases d'apprentissage et de test. Des *a priori* de régions sont pris en compte par le biais de sacs de textons sémantiques, et un *a priori* calculé au niveau de l'image et non de pixels ou de régions sur la base d'apprentissage peut également être inclus. Kluckner et al. [124] intègrent également des *a priori* de contexte sémantique à travers des champs conditionnels de Markov (CRF) en supplément d'informations pixelliques classiques (couleur, orientation et intensité de gradients) pour classifier

des images aériennes ou satellitaires.

Plusieurs études récentes tirent parti de la disponibilité d'une base de données étiquetée représentant des vues au sol de rues [43] afin de pouvoir fournir une classification sémantique de ce type de scènes obtenues à partir de caméras fixes ou embarquées sur véhicule [44, 131]. Toutefois, ces approches sont fondées sur l'utilisation d'indices en trois dimensions peu marqués voire indisponibles dans des séquences vidéo aériennes.

Modèles graphiques et champs de Markov Le contexte spatial correspond aux dépendances entre pixels dans un voisinage. Il est utile de définir un cadre probabiliste capable de prendre en compte les dépendances au sein d'un voisinage. Les modèles graphiques fournissent un mode de représentation qui permet d'intégrer des connaissances *a priori* et un cadre de travail efficace pour l'apprentissage, la définition et l'application de règles de raisonnement, l'interprétation et la prédiction de données. Ces modèles regroupent la théorie des graphes et celle des probabilités. Ils apportent à la fois un outil de visualisation des dépendances entre variables au sein des données mais expriment également les notions de dépendance conditionnelle et de cohérence probabiliste sous forme mathématique.

Les modèles graphiques peuvent être séparés en deux branches. Les modèles orientés, ou réseaux bayésiens ou encore réseaux de croyance, sont utiles pour représenter des relations de causalité ou influence entre des variables aléatoires. Les modèles non orientés ou champs de Markov aléatoires, relient plus naturellement ces variables par des contraintes relâchées et encodent leurs corrélations.

Ces derniers constituent un cadre probabiliste fréquemment utilisé afin de modéliser les dépendances spatiales dans des données vidéo naturelles. Les champs de Markov aléatoires (MRFs) modélisent la probabilité jointe des données observées (l'image) et du champ de données cachées (les étiquettes associées aux pixels) et intègrent ainsi le processus de génération des observations. Contrairement à cette approche générative, une approche discriminative modélise directement la distribution *a posteriori* sur les étiquettes étant donnés les pixels de l'image. Dans un but de classification ou de reconnaissance, ce type d'approche est plus adapté car il ne nécessite pas la recherche du modèle génératif complet. Il évite ainsi de poser des hypothèses simplificatrices sur la nature des données dans un but de réduction de la complexité combinatoire lors des étapes d'apprentissage et d'inférence, telles que leur indépendance spatiale conditionnelle rarement vérifiée pour des images naturelles. Ces modèles conditionnels peuvent être appris de manière discriminante.

Les premiers modèles discriminants proposés ont été les modèles de Markov à entropie maximale (Maximum Entropy Markov Models ou MEMM) [163]. Ces modèles sont associés à une structure de graphe orienté et estiment des probabilités conditionnelles connaissant l'ensemble des observations ainsi que l'état précédent. Aucune hypothèse sur les observations n'est nécessaire, mais la normalisation des probabilités de transition incluse dans les MEMM introduit un "biais d'étiquetage" par rapport aux états qui ont peu de successeurs. En effet, la densité de probabilité d'un état étant répartie localement sur l'ensemble de ses successeurs, les probabilités de transition sont plus élevées pour chaque transition lorsque le nombre de successeurs est plus faible. En particulier, lorsque des états n'ont qu'un unique successeur, l'influence des observations associées est nulle car l'intégralité de la densité de probabilité est transmise à l'unique successeur. Les champs aléatoires conditionnels (Conditional Random Fields ou CRFs) remédient à ce biais en suivant une structure de graphe non orienté. Ils sont d'abord apparus sous la forme de champs unidimensionnels pour la segmentation et l'annotation de séquences [132].

Kumar et Hebert ont ensuite étendu ces modèles au cas bidimensionnel avec les champs aléatoires discriminants (Discriminative Random Fields ou DRFs) [128], notamment afin de

détecter des structures régulières telles que des constructions ou des bâtiments dans des images réelles. L'association entre classes et données est réalisée par des modèles discriminants locaux en chaque pixel de l'image et les interactions spatiales sont considérées entre pixels voisins, soit sur une structure de grille régulière bidimensionnelle.

Des approches plus récentes utilisant les CRFs comprennent des CRFs semi-supervisés [136] permettant d'inclure des données étiquetées et non étiquetées, les CRFs multi-échelle [98] faisant intervenir des primitives et les dépendances à une échelle locale, régionale et globale.

Auto-Context [241, 116] propose une variation par rapport à l'utilisation classique de champs aléatoires. Elle n'impose pas de définir des *a priori* sur les relations entre régions ou pixels voisins, ou plus généralement de concevoir la méthode indépendamment des données. L'algorithme décrit intègre ces dernières en apprenant les paramètres de fonctions d'interaction au sein d'un schéma itératif. Un premier ensemble de classifieurs locaux est appris sur des images d'apprentissage et les cartes d'étiquettes manuelles associées. Les cartes de probabilités d'appartenance aux classes sont ensuite intégrées en tant que données complémentaires de contexte pour une nouvelle étape d'apprentissage. L'algorithme itère ensuite en mettant à jour à chaque itération les cartes de probabilités et en calculant un nouvel ensemble de classifieurs à partir des nouvelles données. Ces données comprennent donc une partie fixe, les primitives issues des données image originales, et une partie variable, constituée par les cartes de probabilités. Cette approche sera privilégiée dans la chaîne de traitement proposée au chapitre 5.

Autres définitions de contexte et leurs utilisations Cette utilisation du contexte dans le domaine des images, à savoir l'intégration de relations spatiales et sémantiques afin d'améliorer notamment les tâches de classification, se retrouve également dans d'autres approches. Ainsi, Rabinovich et al. [201] incorporent un contexte sémantique, ici des matrices de cooccurrence, apprises à partir d'une base de données ou provenant de sources extérieures telles que Google Sets par le biais d'un champ aléatoire conditionnel. Cela permet d'améliorer les performances de classification d'objets.

Heitz et Koller [101] ne fournissent pas directement un ensemble d'apprentissage explicite associé à des étiquettes mais regroupent automatiquement les pixels en régions de contexte probables selon leur apparence propre ainsi que leurs interactions avec les objets détectés dans l'image. Ces régions permettent ensuite de préciser la nature des objets détectés. L'aspect non supervisé du regroupement des régions permet de capturer une grande diversité de concepts parfois non immédiats pour un interprète humain.

Les informations de contexte peuvent prendre d'autres formes encore. Wu définit dans [269] une extension du flot optique fondée non pas sur la contrainte classique de conservation de l'intensité, mais sur la "conservation de contexte". Un contexte est ici défini comme un ensemble de contextes d'apparence (intensité, couleur, gradients...) Cette contrainte tolère ainsi des déformations locales. Le système obtenu par le regroupement des contraintes associées à chaque contexte d'apparence est sur-déterminé et une solution unique peut être fournie par minimisation des moindres carrés par exemple.

En présence d'instances multiples ou proches de même type dans une image ou une séquence vidéo, il est possible de situer chacun de ces objets par rapport à ses "semblables". Le terme "contextes" fait alors référence à des groupes d'objets partageant les mêmes primitives d'apparence ou de mouvement [71]. Cette approche est pertinente lorsque de tels comportements ou apparences collectifs sont présents (par exemple dans des séquences de circulation dense, de surveillance de foule ou encore des défilés). En revanche, elle n'est pas adaptée à la détection d'objets isolés.

4.2 Etat de l'art sur les méthodes de détection d'objets mobiles globales en temps

Les approches de détection d'activité globales en temps, dans le cadre de séquences vidéo aériennes, peuvent être généralement découpées en trois étapes récurrentes : prise en compte du mouvement global, détection et pistage. Une autre séparation est possible : après compensation du mouvement global, des structures sont directement extraites du volume spatio-temporel (la séquence vidéo) ou un espace de dimension supérieure (par l'introduction de la vitesse par exemple), avant d'être classées selon des critères géométriques (forme locale ou globale de la structure) ou de cohérence (d'apparence et de mouvement).

Ces deux découpages présentent des similarités : outre la première phase commune de stabilisation, l'étape suivante est une étape de détection (image par image dans le premier cas, de structures étalées sur des segments temporels dans le deuxième) avant un dernier processus d'interprétation (associé au pistage dans le premier cas par l'étude des pistes obtenues, caractérisation de la géométrie ou de la cohérence des structures dans le second).

Une différence majeure les distingue cependant. Le premier ensemble d'approches repose sur une détection par image, donc sur des données de dimension inférieure, et la performance de l'algorithme de détection influe directement sur celle du pistage consécutif. Les autres approches considèrent des volumes spatio-temporels plus riches mais la complexité en termes de mémoire et de temps de calcul est plus élevée.

Recalage Le mouvement d'entités mobiles doit être distingué du mouvement apparent global créé par le mouvement relatif du capteur par rapport à la scène. Un ajustement préalable à toute détection est donc nécessaire afin de compenser ce mouvement global : il s'agit de l'étape de "recalage".

Diverses méthodes s'attellent à cette tâche et utilisent des primitives ou espaces de représentation variés, tels que les ondelettes [157, 27], le flot optique [112, 264] ou KLT (Kanade, Lucas et Tomasi [221]). Un état de l'art des méthodes de calcul du flot optique est présenté en annexe A.

La mesure de corrélation entre régions ou images entières [199] fournit une autre méthode encore. Une approche hiérarchique plus récente de correspondance utilisant l'outil GPU [53] est de plus robuste à la parallaxe et permet d'enrichir l'analyse du contenu, notamment par l'augmentation du champ visuel.

Dans le cadre d'approches globales pour la détection d'activité au sein de séquences vidéo aériennes, plusieurs difficultés découlent directement de cette première phase de recalage.

Tout d'abord, la dérive inhérente au trajet du capteur et translation de la scène observée rend délicate l'estimation de mouvements sur des segments temporels importants. En effet, un recouvrement est nécessaire entre les images sur lesquelles on veut apparier des points d'intérêt ou calculer un flot optique, ce qui limite l'intervalle maximal d'analyse.

De plus, certains algorithmes d'estimation de mouvement (le calcul de flot optique par exemple) gèrent difficilement de grands déplacements. Si la propagation de mouvements plus faibles, entre des images séparées par un bref intervalle temporel, est possible, elle comporte néanmois un risque d'accumulation d'erreurs.

Les effets de parallaxe apparaissant en présence de structures en trois dimensions de grande hauteur perturbent également le recalage, notamment lorsque les variations de profondeur sont peu continues et occupent une part importante de l'image : il peut alors être souhaitable d'estimer au préalable une carte de profondeurs.

Enfin, la présence éventuelle d'effets de distorsion radiale doit pouvoir être compensée.

Régularisation temporelle Parmi l'ensemble des méthodes de détection d'activité exploitant la cohérence temporelle des données vidéo, un premier groupe englobe celles de pistage. L'intervention de la dimension temporelle sur des résultats obtenus indépendamment pour chaque image est d'intérêt à plusieurs titres. En premier lieu, l'ajout de contraintes de cohérence temporelle (continuité ou comportement localement affine en temps par exemple) permet de filtrer une partie des artefacts tels que des bruits de classification, non stables dans le temps. En contrepartie, ces contraintes peuvent conduire à la suppression d'objets de faible taille ou partiellement détectés : un compromis entre précision et rappel doit être trouvé.

La cohérence temporelle permet également de générer un ensemble de pistes puis de caractériser ces dernières par leur apparence, leur position, leur vitesse voire leur accélération, leur contexte spatial voire leur positions et leur vitesses relatives par rapport à d'autres objets fixes ou mobiles... Cette caractérisation aide à établir une interprétation plus sémantique des objets considérés : s'agit-il de bâtiments de grande hauteur, de véhicules ou de piétons, de comportement typique ou anormal (vitesse excessive, changements de direction désordonnés...) ?

Le pistage est un domaine à part entière et un grand nombre d'algorithmes ont été proposés puis améliorés, dont le filtre de Kalman [117] ou le filtre à particules [58]. Si le filtre de Kalman original est adapté aux cas d'un unique objet à pister, des méthodes peuvent être construites sur son principe afin de traiter le cas d'objets multiples en intégrant des composantes d'association de données et de gestion de pistes. Bazzani et al. comparent dans [22] une telle approche, le filtre de Kalman à hypothèses multiples, à une approche fondée sur l'utilisation de filtres à particules, également étendue au cas multi-objets. Différents modèles permettent de décrire les transitions entre états (apparence, position, vitesse...) et la constitution des pistes, notamment en présence d'objets multiples. Pollard et al. associent dans [194] les avantages des filtres à densité d'hypothèse de probabilité (PHD) adaptés au pistage multi-cibles (algorithme GM-CPHD, ou PHD cardinalisé avec mélange de gaussiennes) à ceux des modèles à interactions multiples avec pistage multi-hypothèses (IMM-MHT). Les premiers fournissent une bonne estimation du nombre de cibles et de l'état de chaque cible, les seconds apportent une précision plus élevée sur l'estimation des vitesses des objets.

Une autre approche consiste à utiliser les Graph Cut à étiquettes multiples [63], ici en 3 dimensions afin de produire des étiquettes spatio-temporelles à partir de résultats de classification obtenus pour chaque image d'un segment temporel. Ces résultats peuvent provenir de différences d'images, de champs de flot résiduel (l'apparence et contexte ne sont alors pas pris en compte) ou encore de cartes de probabilités objet résultant d'un algorithme de classification tel que celui décrit à la section 5. Cette approche sera détaillée à la section 6.

Approches "directes" : volumes spatio-temporels Ces approches présentent plusieurs avantages sur les approches de régularisation. Outre des données plus riches, elles visent en effet à extraire en une seule étape, dans un espace en trois dimensions (position image et temps) ou plus, des structures cohérentes, que l'on peut associer à des trajectoires d'objets ou de groupes d'objets. Intuitivement, l'accumulation temporelle devrait filtrer automatiquement les artefacts et faire ressortir les structures voulues.

De même que pour les approches de régularisation, une étape préliminaire de recalage est nécessaire et peut introduire un bruit supplémentaire par rapport au cas d'une caméra fixe, sans mouvement à compenser ni phénomène de parallaxe. Cette étape de recalage est plus délicate ici car elle nécessite de compenser le mouvement global pour un nombre plus élevé d'images et des mouvements de plus grande amplitude, dans un ré-

férentiel unique. Par contraste, les méthodes de régularisation peuvent se contenter d'un recalage pour chaque couple d'images consécutives, sans risque de cumul d'erreur ou de mouvements de grande amplitude à estimer.

La nécessité de recaler un segment temporel dans un repère commun apporte toutefois un avantage supplémentaire : il est possible de créer un modèle de fond, par exemple par une distribution gaussienne sous condition de fond statique [266], un mélange de gaussiennes si le fond a une distribution plurimodale [286] ou encore un filtrage médian [48]. Parks et Fels présentent un état de de l'art et une comparaison de méthodes de construction de fond dans [188]. La constitution d'un tel modèle avec un recalage limité à des couples d'images consécutives est moins évidente, par exemple en utilisant un modèle probabiliste filtrant peu à peu les objets en mouvement, tout en risquant de supprimer également les structures en trois dimensions. Un modèle précis de fond fournit par différence (simple différence d'images d'intensité ou couleur, ou différence plus élaborée modélisant de possibles évolutions d'illumination et de bruit) un masque d'activité complétant d'autres primitives issues du flot optique ou de détecteurs d'apparence.

Après l'étape de recalage, l'étape suivante consiste à extraire les régions d'activité. La recherche de points d'intérêt en trois dimensions est utile sur plusieurs plans. Elle permet de réduire le volume spatio-temporel associé en un segment temporel donné à un ensemble de primitives plus léger, et ainsi de réduire les coûts de mémoire et de temps de calcul. Ces points "saillants" [205, 147] à la fois dans l'image et dans le temps, sont également fortement susceptibles d'appartenir à des objets mobiles (voire à des bâtiments de grande hauteur).

Cette approche est "directe" ou de "régularisation" suivant l'utilisation des cartes de saillance obtenues (cartes qui nécessitent un volume et non une simple image ou un couple d'images consécutives). En effet, si ces cartes sont traitées comme des primitives à associer par pistage, il faut utiliser une approche directe. Il est également possible d'en extraire dans une approche "directe" des surfaces ou volumes continus, par seuillage, par contours actifs ou par graph cuts. En supposant les objets suffisamment contrastés par rapport au fond environnant, la récupération des contours peut être guidée par les informations d'apparence.

Une approche différente consiste à accumuler les points susceptibles d'appartenir à des objets mobiles dans un espace de dimension supérieure à 3 afin de faire apparaître des structures caractéristiques d'objets mobiles ou des ensembles d'objets de comportement cohérent tels que des voies de circulation qui forment des "nappes" dans l'espace (x, y, v_x, v_y) [279]. La sélection des points provient ici des primitives de flot résiduel indiquant (après compensation du mouvement global) une activité potentielle (ou un bâtiment). Cette approche demande toutefois des données en quantité suffisante pour une caractérisation fiable des structures géométriques.

Dans le cadre de séquences vidéo aériennes avec déplacements importants, les données sont peu redondantes quant à la zone couverte : ainsi, une "nappe" d'activité ne pourrait être représentée que par le mouvement d'un unique véhicule présent sur quelques dizaines d'images consécutives voire moins. En revanche, dans un contexte opérationnel de surveillance ou de contrôle de la circulation routière, avec peu de mouvement global et une zone observée stable, les différentes voies de circulation contenant un nombre élevé de véhicules apparaissent clairement.

Afin de séparer les "nappes" d'activité des artefacts (erreurs de flot ou bruit créé par des variations d'illumination ou de qualité image par exemple) et des effets de parallaxe créés par des structures de grande hauteur, Yu et Medioni utilisent l'outil de vote tensoriel [232]. Cet outil caractérise la géométrie locale d'un nuage de points et permet de filtrer respectivement, suivant l'échelle considérée, le bruit et les effets de parallaxe. L'ensemble des points restants après ce filtrage est ensuite séparé en des "flots" ou ensembles distincts de mouvement cohérent à l'aide d'un algorithme de remplissage par diffusion. L'approche

est complétée par une phase de regroupement de flots susceptibles d'appartenir à un même groupe de trajectoires séparées par occultation, et une phase d'extraction et de pistage indépendants de chaque véhicule.

Un type d'approche similaire consiste à extraire des tubes spatio-temporels correspondant au mouvement des différents objets (et des bâtiments en présence d'effets de parallaxe) : les structures sont ici obtenues dans l'espace temporel original (x, y, t) et non dans un espace augmenté tel que dans [279]. Sun et al. utilisent un modèle de fond avec un algorithme de coupure minimale [229] afin d'extraire la couche d'avant-plan. Des tubes d'activité peuvent ensuite en être extraits comme composantes connexes dans l'espace spatio-temporel [198].

Yuan et al. proposent une approche géométrique [280]. Des applications de contraintes géométriques consécutives permettent de séparer les pixels appartenant à des régions statiques (fond sans effet de profondeur), des bâtiments (effet de parallaxe) ou des objets en mouvement. Les pixels vérifiant une contrainte d'homographie sont ainsi classés comme pixels de la partie plane du fond. La contrainte épipolaire reliant les points associés à un même point de l'espace physique réel dans deux vues différentes permet ensuite de filtrer les pixels de bâtiments ou d'autres structures en trois dimensions, qui vérifient cette contrainte.
Cependant une partie des pixels d'objets en mouvement vérifie également cette contrainte, lorsque la caméra les suit selon un mouvement parallèle au leur par exemple. Une contrainte supplémentaire est donc nécessaire afin de dissiper cette ambiguïté. La contrainte de "cohérence de structure" proposée, une relation bilinéaire entre un couple de structures projectives tridimensionnelles associées au même point physique à partir de trois vues, permet de lever l'ambiguïté. Elle ne suppose pas un plan de référence constant pour les trois images considérées ni de mouvement de faible amplitude entre les images et ne nécessite de point de référence statique.

Le choix de descripteurs spatio-temporels dépend de l'approche choisie, dense ou par points d'intérêt, et conditionne les performances des algorithmes de détection voire de recherche ou de comparaison d'événements. Il peut s'agir d'une simple fusion de descripteurs spatiaux ou d'apparence et de descripteurs temporels (flot résiduel après compensation de mouvement dominant par exemple). Les descripteurs peuvent être également en trois dimensions. Wang et al. proposent un état de l'art de tels descripteurs [250]. Citons également les extensions en $3D$ de SIFT et SURF [161].

Applications Le chapitre 1.1 détaille les différentes utilisations des pistes d'objets ou tubes d'activité dans un but d'interprétation. Les différentes étapes de valorisation de ces informations peuvent être résumées comme suit :

- *Caractérisation*. Une fois les flots d'activités et / ou pistes d'objets disponibles, un certain nombre de primitives doivent être calculées afin de décrire le comportement des objets ou des activités. Il peut s'agir de primitives de mouvement (norme et orientation de la vitesse dans un repère image de référence), d'apparence (dans un but de classification ou de reconnaissance d'objet par exemple), de géométrie (la forme d'un flot d'activité ou tube de mouvement indique s'il s'agit d'un mouvement stable ou chaotique, d'une voie de circulation rectiligne ou courbe...)

- *Indexation et reconnaissance*. Une fois les différents flots ou pistes décrits, l'indexation peut prendre diverses formes. Une transformation des primitives ayant servi à la caractérisation en éléments sémantiquement significatifs permet des requêtes sémantiques sur, par exemple, la forme, la vitesse ou l'accélération de la trajectoire.

L'indexation peut également être implicite, par le biais d'un apprentissage associant des configurations particulières dans l'espace des primitives à des exemples fournis de situations types. Ainsi, Filipovych et Ribeiro présentent des modèles adaptatifs de tubes [74], pour lesquels la forme locale des tubes sert de primitive pour l'indexation et la reconnaissance d'actions. La reconnaissance se fait ici dans un espace de paramètres où les séquences de test sont comparées à des séquences de référence, espace difficile à traduire sur le plan sémantique.

– *Représentation*. Outre l'indexation, la création de modes de représentation compacte des activités présentes dans une séquence vidéo permet de gagner un temps précieux pour la consultation de ces dernières.
Pritch et al. proposent ainsi une représentation non chronologique mêlant différentes activités lors de la visualisation [198]. Il est alors utile de pouvoir gérer le compromis entre exhaustivité du contenu dynamique, compression temporelle et clarté de la visualisation. Une forte compression ainsi qu'une demande d'exhaustivité conduiraient par exemple en cas de forte activité à une visualisation fastidieuse voire non interprétable car l'ensemble des événements serait affiché simultanément.
Parmi d'autres approches de représentation compacte, le regroupement d'activités similaires aboutissant à plusieurs résumés vidéo spécifiques à chaque type d'activité [197] permet une lisibilité accrue.

4.3 Conditions d'évaluation

4.3.1 Données étudiées

Nous avons choisi plusieurs séquences vidéo aériennes de difficultés et environnements variables sur lesquelles évaluer la classification locale. Ces séquences, illustrées sur la figure 4.2 serviront également à évaluer les étapes suivantes de raffinement et de connaissance *a priori* afin d'étudier l'apport de ces traitements.

– La séquence "Blood Diamond", de quelques secondes (100 images à résolution $624x256$), est assez simple, avec peu de véhicules et une route, bien que présentant une incidence élevée : la taille des véhicules est bien plus petite dans la partie supérieure de l'image que dans la partie inférieure des images
– La séquence "Are we changing Planet Earth", de quelques secondes également (150 images à résolution $640x352$), est plus complexe. Elle comporte plusieurs routes et un grand nombre de bâtiments dont l'apparence est similaire à celle des routes. L'incidence est également importante et le mouvement global comprend une translation selon l'axe optique, ce qui modifie l'échelle locale des objets et structures.
– La séquence "BBC", d'une vingtaine de secondes (527 images à résolution $720x576$), présente de nombreuses difficultés. Les changements importants de focale au cours du temps sont accompagnés de watermarks et d'effets de codage marqués. L'incidence est également importante, particulièrement à la fin de la séquence. Les véhicules à l'arrière-plan sont de très petite taille et leur apparence est floutée par les artefacts de compression.

4.3.2 Métriques d'évaluation

Les métriques d'évaluation de la classification dépendent de l'objectif recherché. Généralement, une matrice de confusion traduit la capacité d'un algorithme de classification à caractériser chaque classe avec peu d'erreur, ou peu de "confusion" avec les autres classes. Après normalisation, une matrice quasi diagonale indiquera ainsi une classification presque parfaite. Au contraire, des coefficients extra-diagonaux significatifs seront

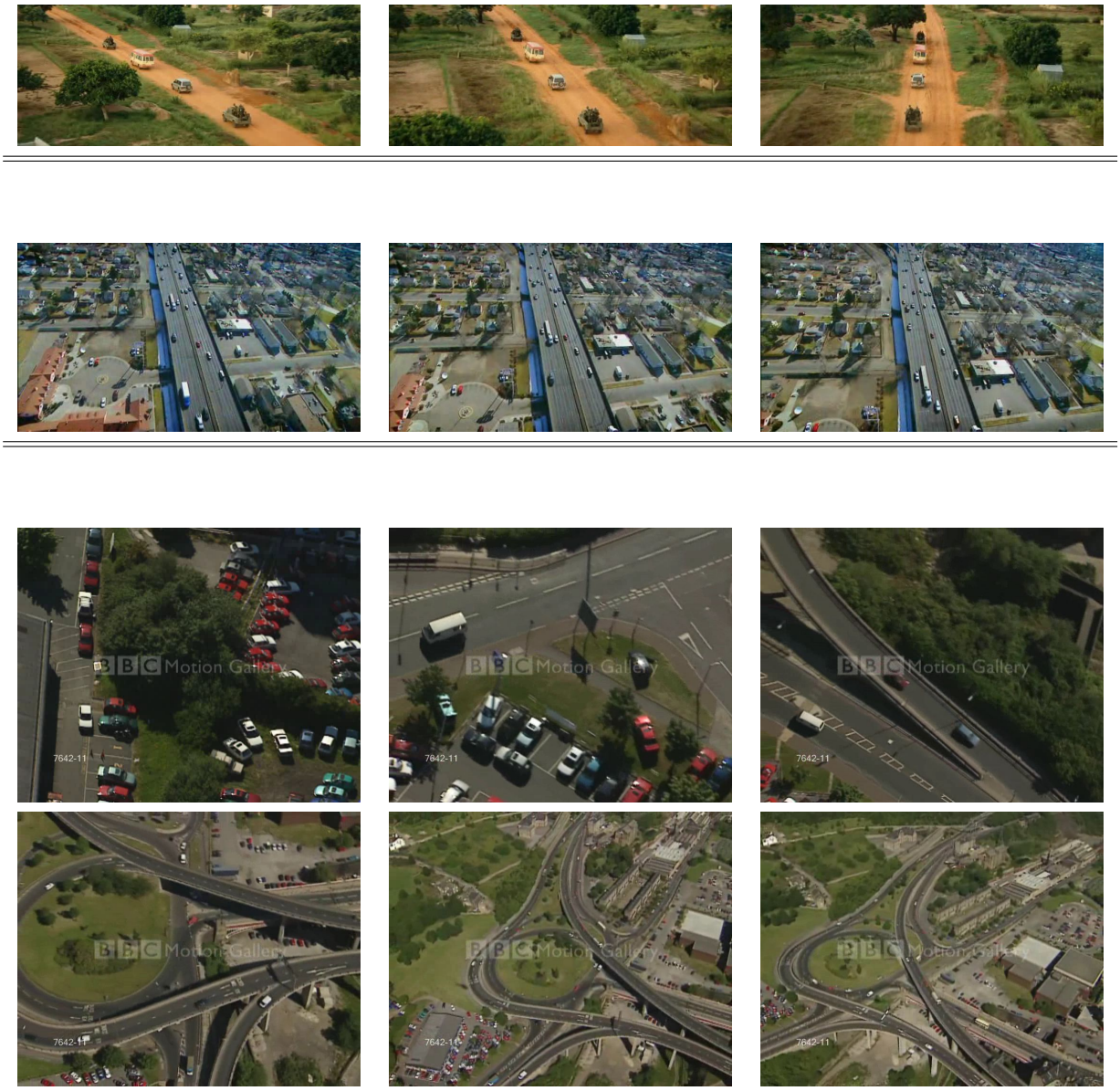
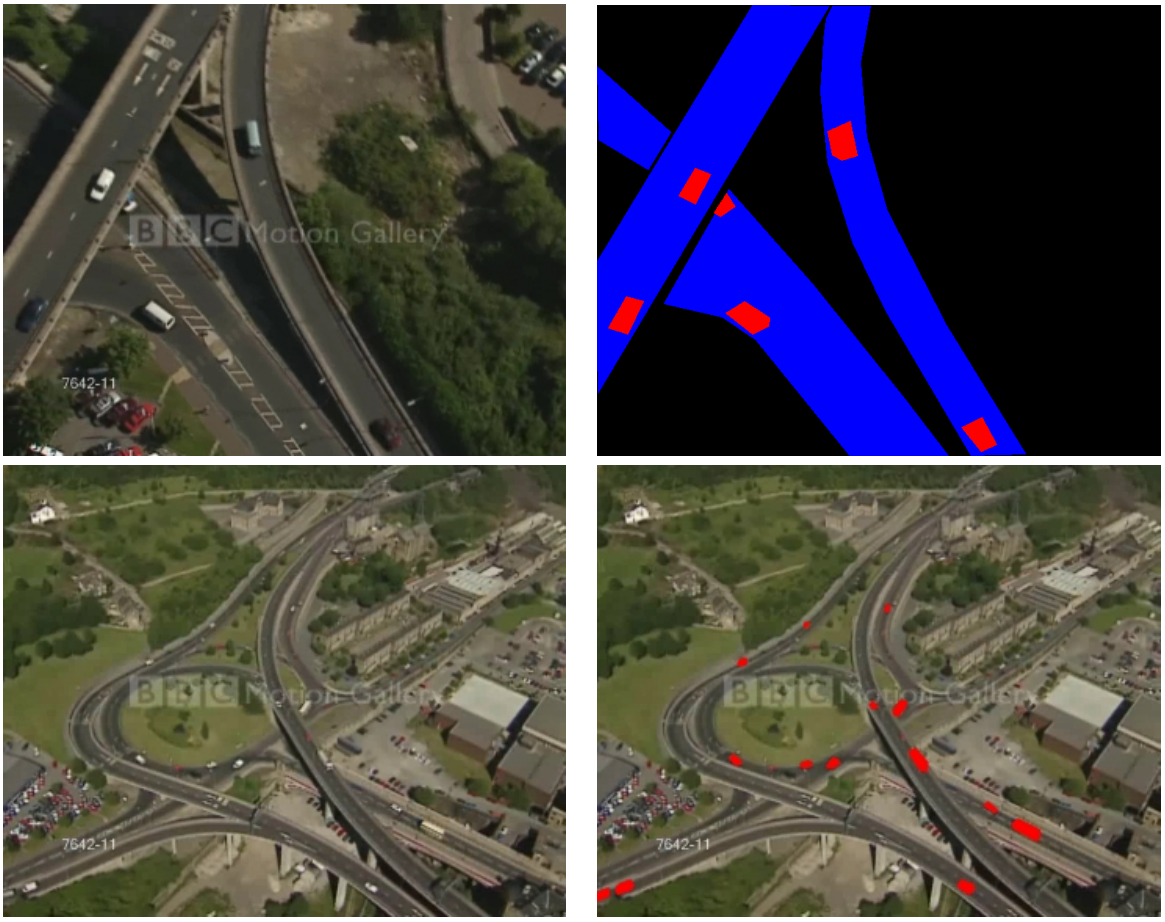


FIG. 4.2 – *Échantillonnage des différentes séquences étudiées, de haut en bas et de gauche à droite. Ligne 1 : séquence "Blood Diamond", 73 images, images 1, 37 et 73. Ligne 2 : séquence "are we destroying Planet Earth", 50 images, images 1,25,50. Lignes 3 et 4 : séquence "BBC", 527 images, images 1,105,210,315,420 et 525.*



(a) (b)

FIG. 4.3 – *Annotation lors de la phase d'apprentissage (1ère ligne) et de test (2ème ligne). Seule la classe "véhicules mobiles" (en rouge) est annotée pour les images de test. Pour les images d'apprentissage, les routes sont annotées en bleu. Le "fond" correspond au complémentaire des deux autres classes.*

signe d'ambiguïté entre deux classes particulières. La matrice de confusion indique ainsi visuellement à la fois les performances de classification et les couples de classes problématiques. Nous nous attachons ici plus particulièrement à la détection d'activité, ou de façon équivalent au sein de l'approche, à la qualité de la classification pour la classe "véhicules mobiles". Les annotations manuelles sont plus légères car seuls les véhicules mobiles doivent être annotés. Un étiquetage par pixel étant trop chronophage, les véhicules sont marqués de manière approchée par des polygones.

Afin d'obtenir des performances sous la forme de pourcentages de précision et rappel, ou de taux d'erreur, dans un objectif de détection, il faut expliciter ces deux termes. Une définition classique de la précision est donnée par le ratio du nombre de pixels correctement classés par rapport au nombre de pixels détectés. Le rappel est le ratio du nombre de pixels correctement détectés par rapport au nombre total de pixels de la classe à détecter sur la vérité terrain. Dans le cadre de notre approche, différents seuillages sur les cartes de probabilités conduiront donc à différents couples de précision et rappel.

Dans un contexte applicatif de surveillance ou de contrôle de circulation par exemple, il peut être utile d'adopter une version "objet" de ces termes. En effet, la vérité terrain souffre parfois d'une qualité image dégradée, avec des effets de bloc, de compression ou de flou notables. Lorsque les véhicules sont de petite taille notamment, l'approximation ou la part libre à l'interprétation inhérentes à l'établissement de la vérité terrain peut perturber l'évaluation par pixels.

Enfin, le calcul des taux de précision et de rappel sur l'ensemble des pixels n'informe

pas sur le nombre de véhicules réellement détectés. En particulier, une non-détection de véhicules de faible taille (dans la portion supérieure de l'image pour une incidence proche de l'horizontale, par exemple) dégradera peu les valeurs. Le calcul des valeurs de précision et de rappel par objets et non plus par pixel apparaît ainsi pertinent. Il faut alors définir les critères permettant de définir un objet à partir des cartes de probabilités ainsi que les critères de détection d'objet :

- les cartes de probabilité pour la classe "véhicule mobile" sont seuillées, ce qui fournit une image binaire. Les composantes connexes de cette image, après une régularisation éventuelle par des opérations morphologiques simples, sont considérées comme étant des objets détectés.
- une composante connexe est considérée comme détection *correcte* s'il existe un objet ou une composante connexe de la vérité terrain "compatible", c'est-à-dire si l'aire de l'intersection entre les deux est supérieure à un pourcentage donné de l'aire de chacune des composantes. Le choix d'un seuil supérieur à 50% semble un bon compromis entre précision et flexibilité.

Pour chaque séquence, les métriques de précision et de rappel sont calculées sur plusieurs dizaines d'images afin d'obtenir des résultats robustes. Les variations des paramètres tels que le seuil de probabilité fournissent différents points de fonctionnement qui permettent d'établir des courbes de précision-rappel. Ces courbes peuvent être traduites sous une forme plus compacte (avec perte d'information) en les limitant à un domaine de fonctionnement précis. Cela revient à fixer un taux de précision ou rappel et conserver la valeur correspondante de l'autre métrique, ou marginaliser les valeurs de précision (respectivement rappel) sur un domaine réduit de rappel (respectivement précision).

En revanche, la longueur des séquences utilisées est limitée et il est donc difficile d'observer les performances de classification sur des données test significativement différentes des données d'apprentissage (changement complet de scène, type de véhicules, aspect des routes...) Toutefois, la séquence "BBC" présente de grandes variations d'incidence et d'échelle ainsi que des rotations de grande amplitude, avec des variations de contenu conséquentes (apparence et hauteur des structures en trois dimensions, tailles et détails des routes, véhicules et objets).

Détection d'activité : approche locale en temps

5

Une première approche pour détecter les différents segments vidéo présentant une activité, c'est-à-dire ici des objets mobiles, consiste à considérer une séquence vidéo comme une suite de couples d'images successives et traiter séparément chacun de ces couples. Les avantages d'une telle démarche sont multiples. D'une part, les traitements algorithmiques peuvent être réalisés sans délai car il n'est pas nécessaire de connaître la suite ou le passé de la séquence. De plus, les données à analyser sont moins volumineuses et demandent donc moins de ressources mémoire que pour un segment temporel plus étendu, pour un choix de primitives équivalent. Ils sont ainsi plus adaptés à une architecture embarquée. D'autre part, il n'est pas nécessaire de définir un repère de référence ni de recaler les autres images d'un segment temporel dans ce repère. En effet, les régions peu texturées et les occultations, notamment aux alentours des gradients de profondeur, ainsi que des déformations radiales ou des variations de qualité, induisent des erreurs de recalage parfois importantes. Une approche locale en temps réduit grandement ces problèmes. La section 4.1 s'efforce de décrire les différentes méthodes de détection d'activité locales en temps.

En revanche, certains résultats parasites, causés par une mauvaise qualité d'image ou la présence de structures en trois dimensions causant des effets de parallaxe, seront difficiles à filtrer en l'absence de cohérence temporelle. Il est également impossible de suivre des objets mobiles et d'interpréter leur trajectoire afin d'en extraire des groupes d'objets de comportement semblable ou de détecter des comportements particuliers sans pistage, régularisation temporelle ou approche globale.

Contributions Les éléments bibliographiques décrivant les approches locales en temps pour la détection d'activité ont été présentés en section 4.1. Nous allons ici détailler l'approche choisie, illustrée par le schéma figure 5.1. Elle consiste à effectuer dans un premier temps une classification locale (section 5.1) entraînée sur une ou plusieurs images de la séquence vidéo considérée. Les cartes de probabilités obtenues complètent les primitives issues directement de la séquence pour une nouvelle classification et sont mises à jour dans un procédé itératif (section 5.2) qui précise peu à peu les résultats. Une dernière étape applique des contraintes sémantiques afin de filtrer les fausses détections restantes (section 5.3) et nécessite de détecter le réseau routier, par exemple à partir des étapes précédentes en ayant choisi les "routes" comme l'une des classes.

L'approche proposée permet de découpler une classification purement fondée sur l'étude de primitives locales, parallélisable et rapide (une fois les classifieurs appris) d'une part, et l'utilisation de règles de contexte sémantiques adaptées aux classes choisies et au cadre d'étude (ici des séquences vidéo aériennes) d'autre part. Le procédé de raffinement itératif des cartes de probabilités permet de les régulariser et d'obtenir des régions cohérentes. Cela facilite l'application de règles de contexte concernant la taille, la forme ou l'aspect de régions préalablement découpées. De plus, le seuillage des cartes finales de

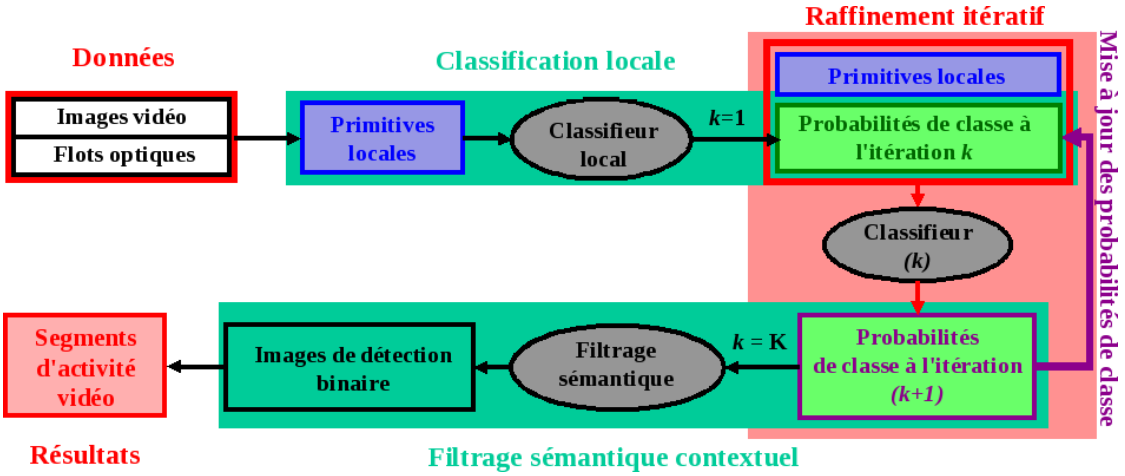


FIG. 5.1 – *Détection d'activité en plusieurs étapes : classification locale, raffinement itératif et application de règles de contexte sémantiques. La classification locale utilise seulement les données (image et mouvement, ou flot optique) pour fournir un premier ensemble de cartes de probabilités (itération 1). Le raffinement itératif intègre ces cartes pour une nouvelle classification sur un espace de primitives augmenté et met à jour les cartes (la régularisation spatiale par régions, étape de traitement accessoire, a été ici omise pour plus de clarté). Après quelques itérations, les cartes obtenues sont filtrées à l'aide de règles de connaissance a priori. Les cartes binaires de détection obtenues après opérations morphomathématiques permettent l'indexation vidéo des segments comportant de l'activité (après une éventuelle régularisation temporelle).*

probabilité, éventuellement suivi d'opérations morphomathématiques, produira des objets binaires moins bruités. Le processus d'apprentissage itératif est peu classique dans un contexte aérien. Il permet d'introduire un compromis entre complexité algorithmique et coût calculatoire d'une part, adéquation aux données traitées d'autre part.

Chaque étape de la méthode peut être adaptée selon l'environnement choisi. Ainsi, pour des classes précises, dans un but d'identification plus que de classification par exemple, il est immédiat de remplacer l'ensemble de primitives locales ici choisies par des primitives d'apparence plus complexes, HoG (histogrammes de gradients), filtres de Gabor ou encore SIFT [153] et son extension GLOH [164]. Les règles de contexte peuvent être définies comme des restrictions d'interactions possibles entre les différentes classes ou instances d'une même classe. L'approche choisie est donc particulièrement flexible et peut être adoptée pour des tâches diverses (classification, identification, détection de comportements atypiques) et dans des environnements variés (vidéo aérienne ou non, caméra fixe ou mobile, visible, infrarouge ou hyperspectral...) Dans le contexte applicatif qui nous intéresse, à savoir des vidéos aériennes, les deux niveaux de prise en compte de l'information de contexte, à partir des données et d'*a priori* de connaissance respectivement, permettent de capturer la structure locale mais aussi globale de la scène. De plus, le cadre d'apprentissage fondé sur des primitives d'apparence, classique dans des applications de classification image, est étendu au domaine de l'interprétation de vidéos aériennes. Les résultats obtenus montrent qu'il s'agit d'une approche efficace qui peut être couplée à une approche purement géométrique afin de réduire le nombre de fausses alarmes.

5.1 Classification locale

Plusieurs éléments doivent être choisis dans une approche de classification : les classes, les primitives et l'algorithme de classification lui-même. Des raffinements tels qu'une implémentation multi-échelles peuvent être inclus dans l'algorithme ou pris en compte directement dans la définition des primitives. Différentes méthodes de régularisation sont également possibles. Un compromis doit être trouvé entre bruit et perte de précision, lorsque certaines instances de classe sont de petite taille et risquent d'être absorbées par une instance d'une classe différente, voisine et plus importante. La régularisation procède alors par

voisinage, fusionnant des pixels ou régions voisins selon la similarité des primitives associées. La régularisation peut être effectuée à plusieurs échelles si un ensemble hiérarchique de partitions, chacune associée à une granularité différente, est disponible. La qualité de la segmentation influe alors directement sur la précision des étiquettes obtenues.

Parmi ces différents raffinements, nous avons inclus une régularisation par régions après une segmentation partielle. L'aspect multi-échelles a été également abordé.

5.1.1 Choix des classes et primitives

Choix des classes L'objectif de l'approche est de détecter les objets mobiles, plus particulièrement les véhicules. Les "véhicules mobiles" forment donc une première classe naturelle. Le "fond", ou tout ce qui ne correspond pas aux classes d' "intérêt", constitue une deuxième classe. Ces deux classes peuvent sembler suffisantes si l'on ne considère que la première étape de détection locale.

Toutefois, afin de profiter des informations de contexte dans la phase de raffinement itératif et lors de l'apport de connaissances *a priori*, il apparaît pertinent de compléter ces deux classes. Les "routes" sont ainsi un élément de contexte adapté à la détection de véhicules mobiles (sous l'hypothèse de circulation sur les routes).

Les effets de parallaxe pouvant être particulièrement importants dans des vidéos aériennes si l'altitude est faible et les bâtiments de grande hauteur (les gratte-ciel en fournissent des exemples typiques), il est envisageable de rajouter, en supplément des trois classes choisies, une classe de "structures en trois dimensions".

Plus généralement, la définition des classes doit être en adéquation avec le but recherché, qu'il s'agisse d'indexation selon des classes imposées (dans un but d'enrichissement d'une base de données par exemple), de détection d'éléments d'intérêt (ici les véhicules mobiles) ou d'évaluation des performances d'algorithmes de classification. La conception de la chaîne de traitement influence également le choix des classes, par l'utilisation d'informations ou d'*a priori* de contexte sémantiques sur les relations spatiales reliant les classes.

Choix des primitives Le choix des primitives est guidé par la nature des données ainsi que les différentes classes. [276] présente ainsi les différentes représentations d'objets dans un but de pistage, divisés en modèles de forme et d'apparence. La classe de "fond" regroupe une grande diversité d'éléments et il est difficile de cerner des primitives particulièrement adaptées.

Si les structures en trois dimensions forment une classe à part entière, et en l'absence de variations importantes de profondeur dans la scène, le mouvement résiduel du fond après recalage sera faible voire nul. L'amplitude du flot résiduel par exemple apparaît comme une primitive efficace pour la caractérisation du fond.

Les véhicules mobiles pouvant présenter des vitesses variables selon les véhicules, cette primitive seule n'est pas suffisante pour cette classe. L'ajout de l'écart-type de la distribution de l'amplitude du flot résiduel sur un patch local apporte une primitive complémentaire afin de distinguer les véhicules. En effet, l'objet étant rigide, le mouvement résiduel après recalage des pixels associés est uniforme et cet écart-type local devrait donc être nul sur des patchs internes au véhicule. Les routes présentent le même comportement que le fond quant au mouvement résiduel, à savoir une profondeur généralement constante et un flot résiduel nul. Ce critère permettant de séparer en théorie routes et fond d'une part, objets en mouvement et structures en trois dimensions d'autre part, est donc retenu.

Si les véhicules et le fond arborent des teintes variées, les routes en revanche sont généralement de couleur plutôt uniforme et distincte. Des primitives de couleur, valeurs des canaux RGB ou saturation, semblent adaptées pour cette classe. Il est possible d'utiliser des primitives d'apparence plus complexes, telles des matrices de cooccurrences ou descripteurs de type SIFT ou histogrammes de gradients. Toutefois, ces primitives sont plus utiles pour la reconnaissance ou détection d'objets particuliers que pour la distinction

TAB. 5.1 – *Primitives utilisées et classes visées*

Primitive	Classe ciblée	Rôle
Intensité (moyenne)	routes	détection de route
Saturation (moyenne)	routes	détection de route
Amplitude du flot résiduel (moyenne)	véhicules	détections de véhicules et 3D
Intensité (écart-type)	toutes classes	distinction intérieur / frontières
Saturation (écart-type)	toutes classes	distinction intérieur / frontières
Amplitude du flot résiduel (écart-type)	véhicules	distinction véhicules / 3D

entre des classes plus générales aux échelles considérées (avec une meilleure résolution, les descripteurs de textures pourraient capturer des détails de textures du fond, des routes ou des véhicules et enrichir l'apprentissage).

Les primitives retenues comprennent donc des primitives temporelles et d'apparence. Les primitives temporelles sont tirées d'une estimation du flot résiduel obtenu après compensation du mouvement dominant. Celui-ci est obtenu par une régression utilisant un M-estimateur [176]. Les valeurs du flot résiduel dépendent de l'échelle et sont normalisées par un facteur estimé (cf. équations (3.5)) à partir des modèles affines globaux obtenus comme décrit précédemment en section 3.2. Ce facteur permet de compenser les variations d'échelle dues au changement de focale et au mouvement de la caméra. La dérive potentielle due au produit cumulatif des facteurs de zoom n'a pas été observée sur les séquences utilisées de quelques dizaines d'images. Elle peut être contrôlée par le choix d'images de référence pour l'estimation du mouvement affine, régulièrement mises à jour afin de prendre en compte la modification de l'environnement.

Pour un pixel donné, les 6 primitives choisies sont calculées comme moyennes et écarts-types sur des patchs 3×3 centrés sur le pixel. Les deux primitives temporelles sont obtenues à partir de l'amplitude du flot résiduel, les quatre primitives d'apparence à partir de l'intensité et de la saturation. La figure 5.2 illustre ces primitives sur une image de la séquence "are we destroying planet Earth".

Il est possible de calculer les moments sur des patchs de taille plus grande, ce qui permet de disposer de caractéristiques plus discriminantes. Toutefois, la simplicité des primitives retenues (moments d'ordre 1 et 2) entraînera alors, en même temps qu'une régularisation ou lissage plus marqué, une perte de précision. Cela est dommageable en présence d'objets de petite taille et la probabilité de mélanger au sein de ces patchs agrandis des pixels de classes différentes est plus élevée. Dans le contexte aérien de l'étude, les objets sont principalement de "petite" taille, avec une dimension image caractéristique de quelques pixels. En revanche, des primitives plus complexes telles que celles mentionnées précédemment pourraient être envisagées sur des patchs de plus grande taille, notamment si la résolution au sol est élevée.

L'ensemble des primitives retenues est particulièrement simple. Les raisons de ce choix sont multiples. Tout d'abord, la complexité algorithmique de la classification dépend directement de la dimension des caractéristiques choisies et une dimension réduite permet de limiter les temps de traitement sans devoir diminuer artificiellement la dimension de l'espace des caractéristiques, par une analyse en composantes principales par exemple. De plus, ces primitives sont génériques et peuvent être utilisées dans un large éventail de séquences vidéo aériennes présentant des contenus variés et des conditions de prise de vue diverses (échelle, incidence, illumination).

Enfin, la prise en compte des zones de séparation entre classes, qu'il s'agisse de bords de route ou de véhicules, peut être gérée indépendamment du choix des primitives image.

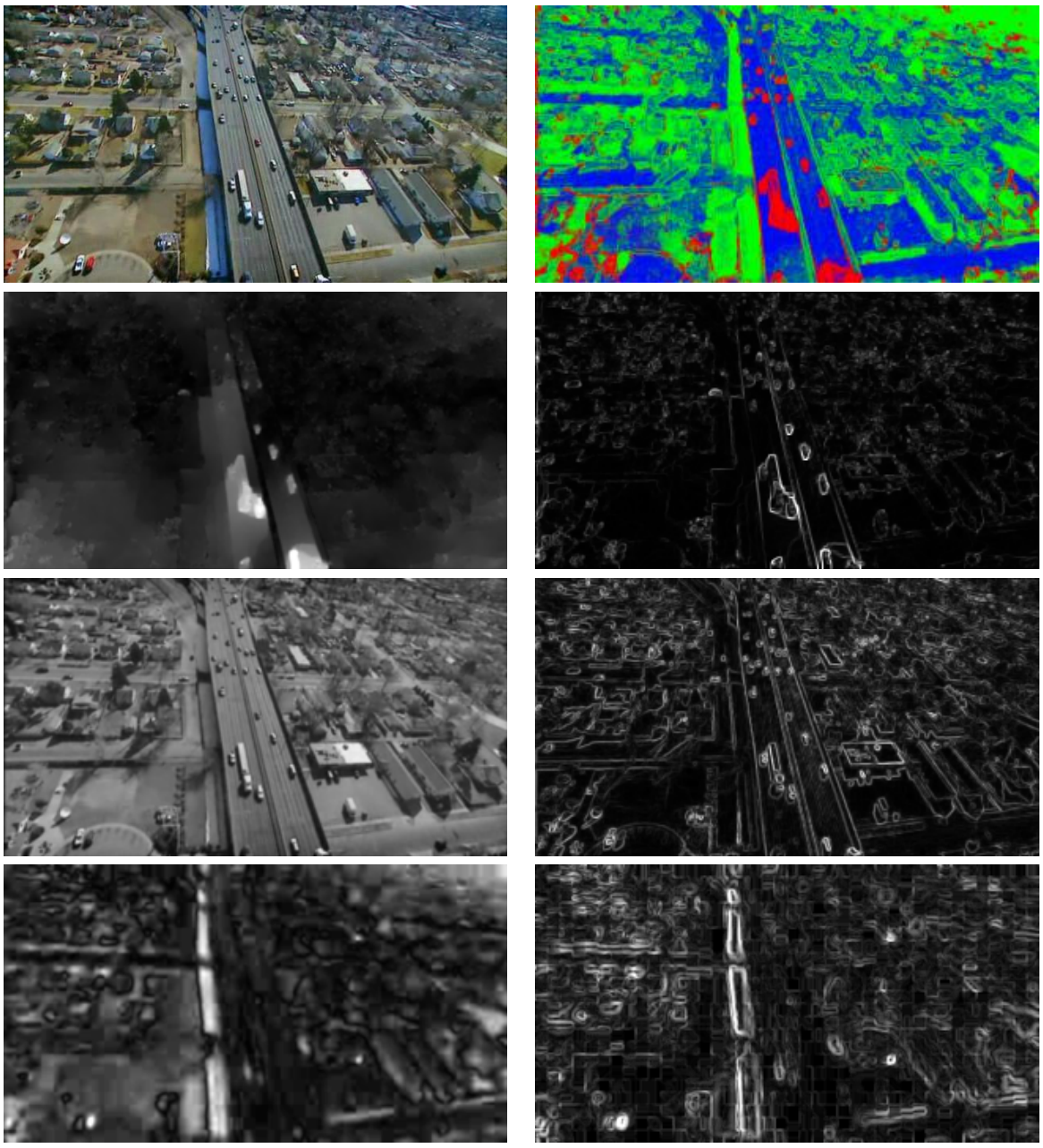
Le modèle retenu propose ici d'intégrer ces relations en considérant les probabilités d'appartenance à chaque classe au sein d'un algorithme d'apprentissage itératif, précisé à la section 5.2. L'apport de règles de connaissance *a priori*, détaillées à la section 5.3, constitue une étape supplémentaire de correction des cartes de probabilité.

5.1.2 Algorithme de classification

Le problème de classification à classes multiples (ici trois classes au minimum, "fond", "routes" et "véhicules mobiles") peut être ramené à un ensemble de classifieurs un-contre-un, les "base learners". Nous avons choisi une approche simple de boosting, rapide et transparente quant aux primitives pertinentes. Chaque "base learner" est obtenu en combinant des classifieurs faibles par Adaboost selon le processus illustré figure 5.3. Nous utilisons un arbre de décision de type CART [42] intégré dans Matlab afin de générer un ensemble de classifieurs faibles avec un taux d'erreur raisonnable à partir des échantillons d'apprentissage, plus précisément six primitives et une étiquette pour chaque échantillon. Chaque nœud interne de l'arbre fournit un classifieur faible sous la forme d'un "decision stump", c'est-à-dire un classifieur binaire associé à un seuil sur une unique primitive. Afin d'accélérer le boosting, et compte tenu du faible nombre de primitives considérées, l'ensemble des "decision stumps" peut être préalablement réduit à ceux dont les taux d'erreur à l'apprentissage sont les plus faibles. L'optimalité de cette présélection n'est pas théoriquement garantie mais la propension du boosting à sélectionner à chaque étape un classifieur faible présentant un faible taux d'erreur paraît naturelle. La comparaison des résultats obtenus avec ou sans présélection ne montre pas de baisse significative des performances de classification. Pour chaque couple de classes, un classifieur "fort" ou "base learner" est donc obtenu lorsqu'un critère d'arrêt est validé (sur l'erreur d'apprentissage ou le nombre d'itérations par exemple) lors du boosting.

Les $\frac{n_c(n_c-1)}{2}$ "base learners", où n_c est le nombre de classes (ici trois dans la version générale du système), sont calculés comme décrit précédemment dans la phase d'apprentissage. Les échantillons sont extraits d'une ou de plusieurs images de la séquence vidéo à analyser voire d'autres séquences et sont associés à des étiquettes manuelles. Le choix des échantillons introduit un compromis entre précision et généralité. Ils sont ici pris pour chaque classe aléatoirement parmi les données d'apprentissage complètes correspondant à la classe. En effet, apprendre les classifieurs sur des images provenant de séquences variées permettra de mieux capturer les primitives communes à chaque classe, au détriment de la précision sur une séquence particulière. De plus, il est possible d'obtenir des résultats de mauvaise qualité si les échantillons présentent des variations trop importantes, avec une variabilité intra-classe égale ou plus importante que les variabilités inter-classe. Lors de la phase de test, les primitives sont calculées à chaque pixel et les "base learners" sont appliqués à ces primitives.

Une transformation logistique symétrique [79] est appliquée aux résultats des "base learners". Cette transformation permet d'obtenir à partir des scores de chacun des $\frac{n_c(n_c-1)}{2}$ "base learners" un ensemble de n_c cartes de probabilité de somme 1 pour chaque pixel. Plusieurs exemples de telles cartes sont présentés figure 5.4. Ces cartes appellent plusieurs remarques. Tout d'abord, la qualité de la détection est variable, avec des meilleurs résultats pour les séquences "VIVID désert" et "BBC" (respectivement troisième et quatrième lignes). Les images extraites de ces séquences présentent peu de parallaxe (la séquence "VIVID" présente tout de même un arbre dans la partie inférieure gauche), au contraire des autres : l'arbre sur la séquence "Blood Diamond", les bâtiments dans la séquence "are we destroying Planet Earth" et "Dubai palace". Cela est particulièrement marqué pour cette dernière séquence. L'effet de parallaxe est particulièrement notable pour la tour centrale. De plus, cette séquence présente une résolution au sol réduite. Les véhicules ne repré-



(a) (b)

FIG. 5.2 – *Illustration des primitives choisies. 1ère ligne : image test et carte des probabilités obtenue après classification locale. 2ème, 3ème et 4ème lignes : moyennes (a) et écarts-types et (b) locaux sur des patchs* 3×3 *de : 2ème ligne, amplitude du flot résiduel ; 3ème ligne, intensité en niveaux de gris ; 4ème ligne, mesure de saturation*

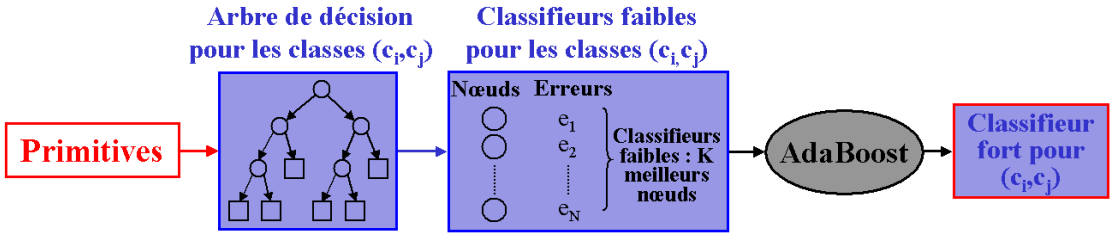


FIG. 5.3 – *Obtention des "base learners" par Adaboost à partir des primitives d'apprentissage*

sentent ainsi sur l'image que quelques pixels de surface. Enfin, l'environnement urbain est particulièrement riche, notamment dans la partie supérieure droite de l'image.

5.1.3 Implémentation multi-échelles et régularisation par régions

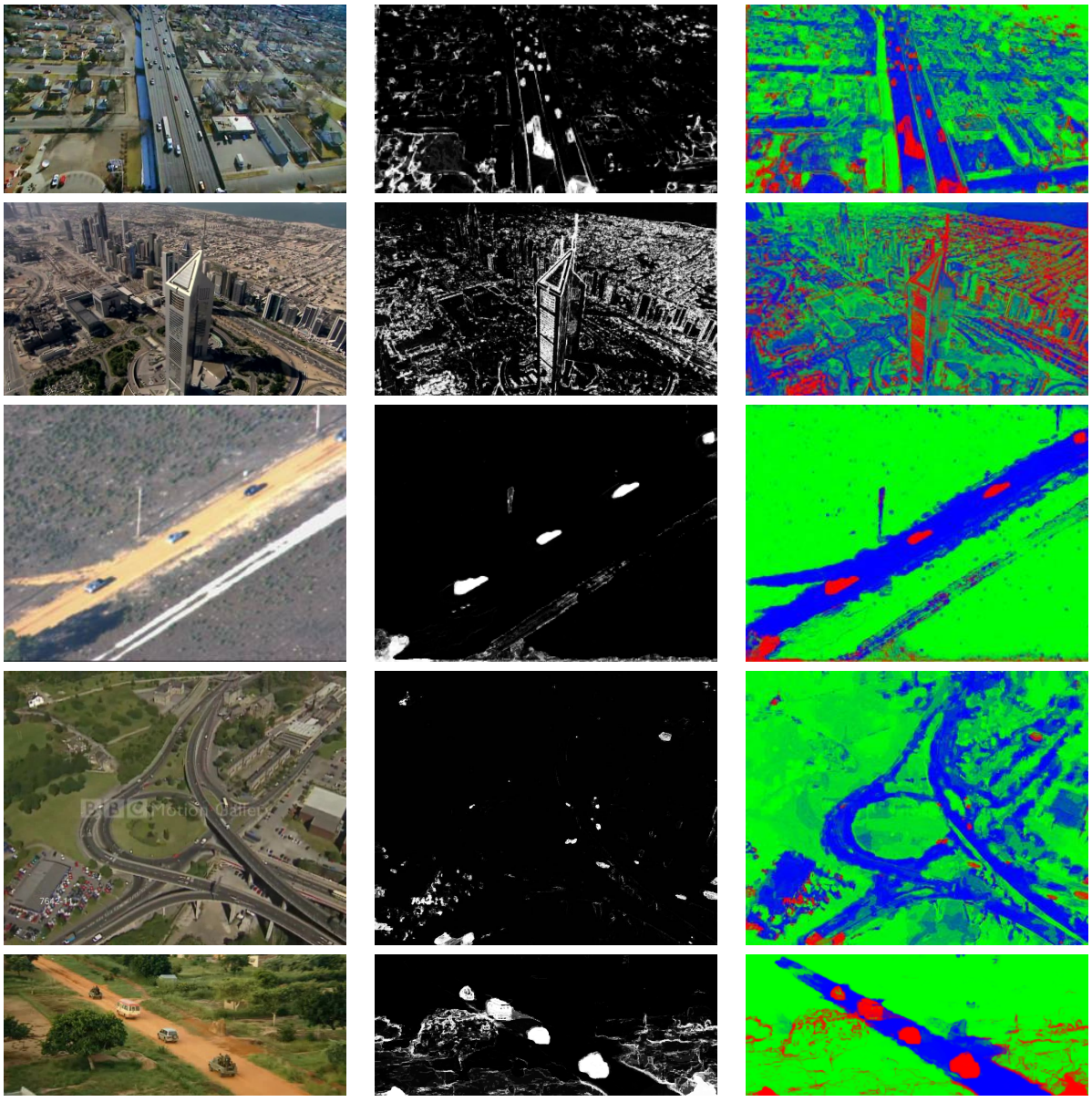
La classification est réalisée pour chaque pixel sur le patch 3×3 centré sur le pixel, indépendamment des voisins. Les cartes de probabilités obtenues sont donc particulièrement bruitées. Outre les multiples détections ponctuelles ne correspondant à aucun objet, les régions correspondant aux objets ne sont pas toutes homogènes. Plusieurs pistes sont envisageables afin d'obtenir des cartes plus régulières. Enfin, des régions détectées sont en fait dues à la parallaxe et doivent être filtrées. Dans un premier temps, il s'agit de régulariser les régions en supprimant les détections ponctuelles et d'obtenir des cartes de probabilités plus homogènes par région.

L'utilisation de patchs de tailles plus importantes est à double tranchant. Les cartes obtenues sont en effet plus lisses et moins bruitées, mais les détails les plus fins disparaissent également. L'idée simple d'effectuer une moyenne des cartes obtenues avec plusieurs tailles de patchs est peu pertinente. Un choix majoritaire ou médian est plus robuste mais pas forcément adapté. Ainsi, un véhicule de petite taille pourrait être associé à une classe dominante "véhicule mobile" à l'échelle adaptée, le patch 3×3, mais à la classe "route" aux échelles plus grossières et un choix tel que proposé ci-dessus privilégierait la classe inexacte. Une pondération des différentes échelles, ou une sélection de l'échelle suivant la structure locale de l'image est donc nécessaire. Une estimation de cette échelle peut être réalisée de plusieurs manières. L'une consiste à analyser l'image originale dans un espace multi-échelles tel qu'une base d'ondelettes. Les coefficients de détail indiquent l'échelle des textures, en partant de l'échelle la plus fine. Il faut toutefois choisir dans cet espace d'ondelettes la taille de la fenêtre sur laquelle observer les coefficients.

Une autre approche consiste à considérer non plus l'échelle mais la structure locale même de l'image. L'idée sous-jacente est de favoriser l'homogénéité des cartes de probabilité au sein d'une région homogène de l'image. La difficulté réside alors dans l'obtention d'une segmentation automatique de l'image séparant correctement des structures ou objets différents, telle une segmentation manuelle, en présence de dégradés de faible amplitude et d'objets fortement texturés. Ces derniers peuvent être éclatés en plusieurs régions homogènes plus simples à classer. En revanche, les éléments difficiles tels que des véhicules peu texturés et de couleur proche de celle de la route adjacente, avec des gradients de séparation quasi nuls ou de l'ordre du bruit, seront fusionnés. Ce problème rejoint celui de l'estimation d'un flot optique précis, délicat en présence de zones faiblement texturées et de gradients peu marqués.

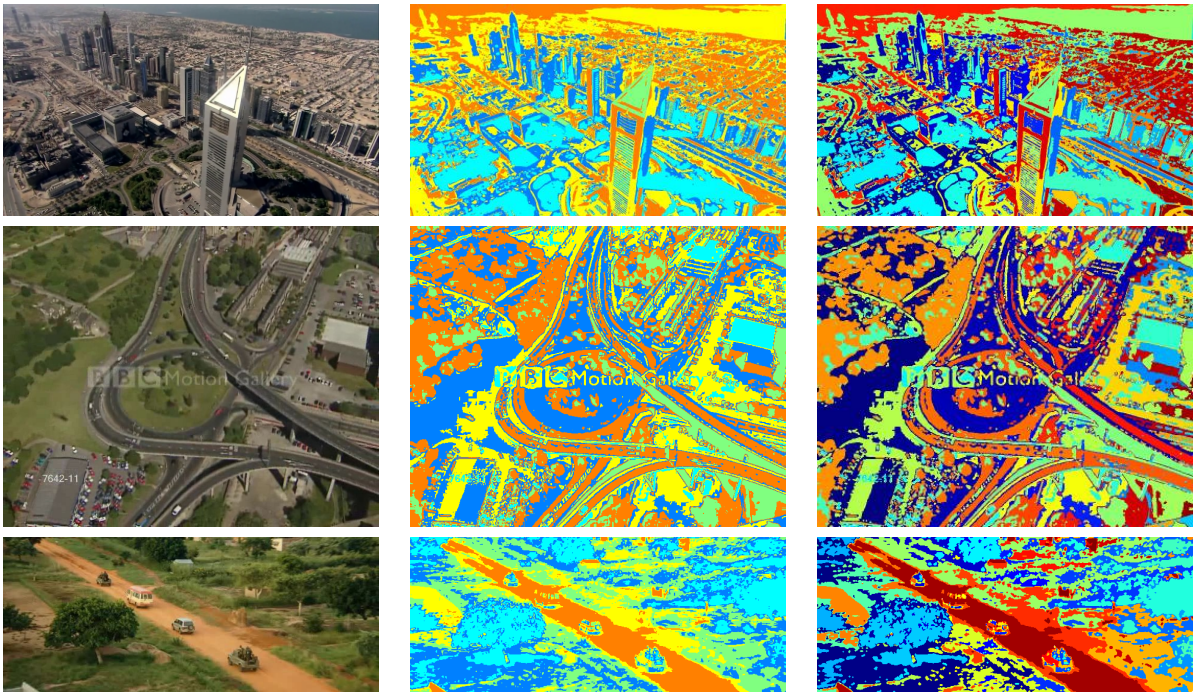
Régularisation par régions Il existe de nombreux algorithmes afin d'obtenir une telle segmentation image en régions homogènes. L'algorithme des K-moyennes est particulièrement simple mais dépend à la fois du nombre de classes choisi et de l'initialisation. En ce sens, la qualité de la segmentation est hautement variable et dépend de la structure de l'image. Un tel algorithme est utile lorsque l'image présente des régions de couleurs contrastées mais fusionnera sinon des régions d'apparence proche. La figure 5.5 donne des exemples de telles segmentations. Sur la première ligne notamment, pour la séquence "Dubai Palace", la segmentation regroupe au sein de mêmes régions des pans de tour et les ombres portées qui correspondent à des structures différentes. Sur la dernière ligne, pour la séquence "Blood Diamond", des parties de véhicules dont les apparences sont proches de celle de la route ou du fond avoisinant, sont intégrées aux régions correspondantes.

Nous disposons ici d'une information dynamique supplémentaire, le flot résiduel. Le but final de la classification étant la détection des régions d'activité ou ici les véhicules mobiles, il semble intéressant d'intégrer cette information dans la segmentation. Ainsi,



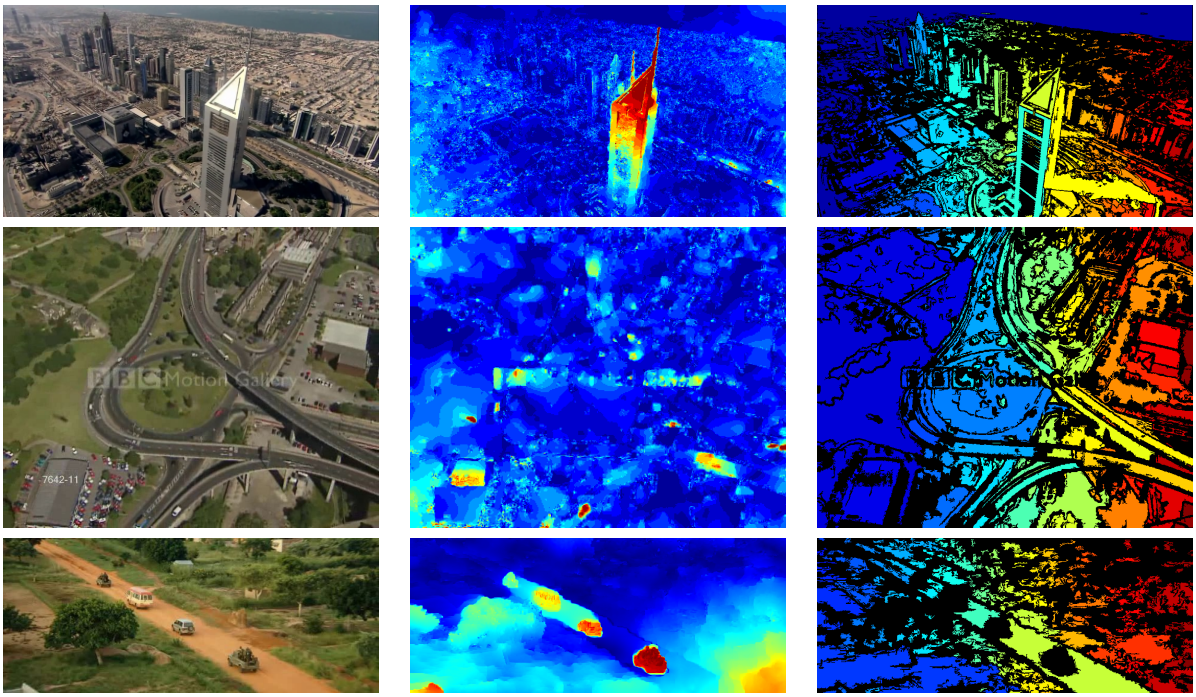
(a) (b) (c)

FIG. 5.4 – *Cartes de probabilités avec* 3 *classes. (a) Image de test (b) Carte de probabilité pour la classe "véhicules mobiles" (c) Composition colorée regroupant les* 3 *cartes. Rouge : véhicule mobile ; vert : fond ; bleu : routes. La somme des trois canaux vaut* 255 *pour chaque pixel, l'intensité n'est donc pas constante.* 1*ère ligne : séquence "are we destroying Planet Earth".* 2*ème ligne : séquence "Dubai Palace".* 3*ème ligne : séquence VIVID "désert".* 4*ème ligne : séquence "BBC7642-11".* 5*ème ligne : séquence "Blood Diamond".*



(a) (b)

FIG. 5.5 – *Exemples de segmentations par K-moyennes. (a) Image à segmenter. (b) Classification avec* 5 *classes. (c) Composantes connexes après opérations morphomathématiques.* 1*ère ligne : séquence "Dubai Palace".* 2*ème ligne : séquence "BBC".* 3*ème ligne : séquence "Blood Diamond".*



(a) (b)

FIG. 5.6 – *Exemples de segmentations partielles par détourage de gradients d'intensité. (a) Image à segmenter. (b) Amplitude du flot résiduel. (c) Composantes connexes après opérations morphomathématiques (les régions noires ne sont pas segmentées).* 1*ère ligne : séquence "Dubai Palace".* 2*ème ligne : séquence "BBC".* 3*ème ligne : séquence "Blood Diamond".*

l'algorithme des K-moyennes peut être lancé sur des primitives de dimension supérieure regroupant des informations de couleur (voire texture) et de mouvement. Cependant, le choix du nombre de classes et des initialisations reste toujours aussi important dans le résultat de la segmentation. De plus, le problème de contraste évoqué pour l'information couleur se retrouve dans le flot qui n'est pas nécessairement exact sur les régions homogènes ou sur les contours de faible contraste. Si cet algorithme fournit de bonnes segmentations sur certaines images et flots associés, les résultats sont de moins bonne qualité lorsque la scène est complexe et présente un nombre élevé de couleurs et de changements progressifs de teinte ou illumination. Le nombre de classes optimal reste de plus à déterminer manuellement.

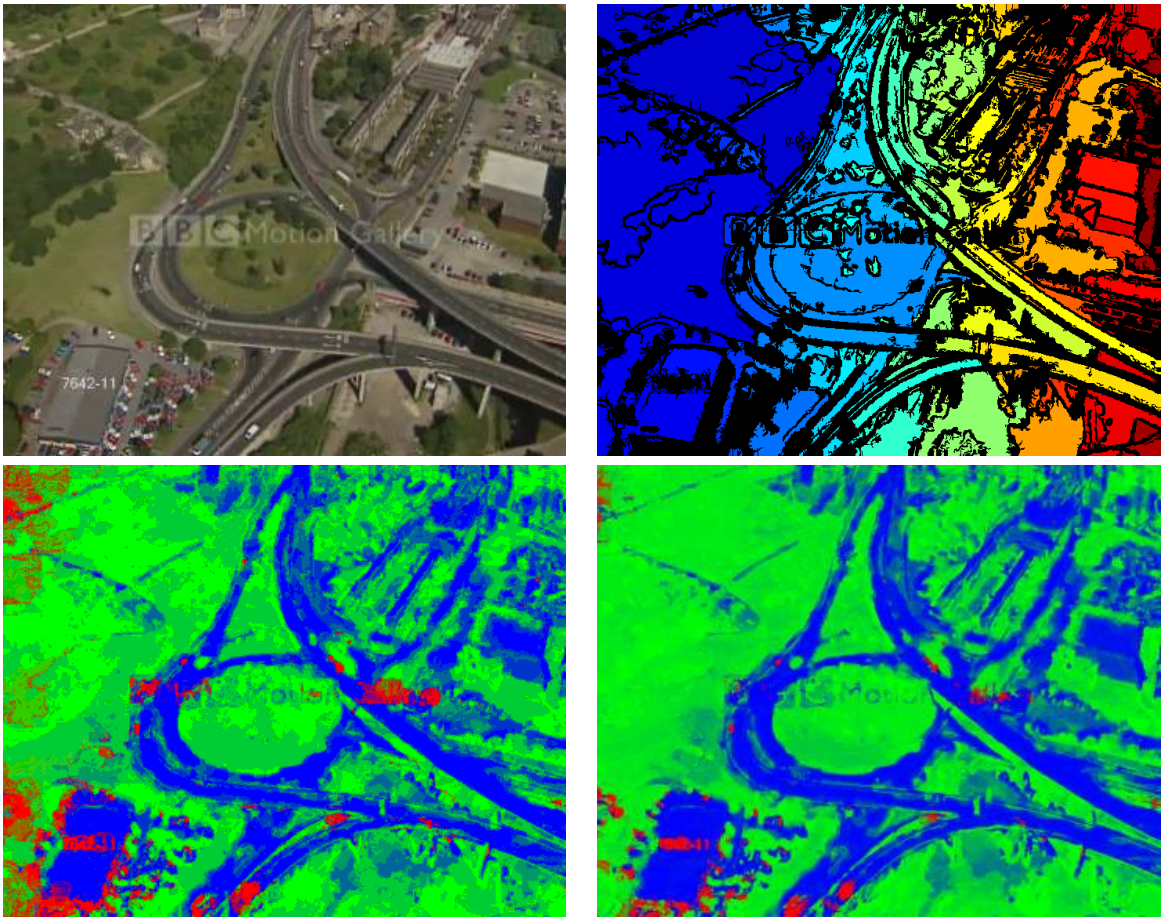
Nous avons donc choisi une autre solution simple qui consiste à utiliser les forts gradients de couleur et de mouvement résiduel comme éléments de séparation de régions. En effet, les gradients de grande amplitude sont des informations fiables permettant d'isoler des régions homogènes tout en autorisant des variations progressives au sein de ces régions. Des applications de filtres morphologiques sur la taille des régions extraites comblent les "trous" ou suppriment au contraire des régions de petite taille (par exemple d'aire inférieure à 5 pixels). Cette approche présente en revanche un inconvénient, la segmentation obtenue étant en effet incomplète. Les zones de discontinuités ainsi que les régions de petite taille ne seront alors pas sujets à la régularisation par régions. La figure 5.6 illustre ce type de segmentation à partir de gradients à la fois du flot résiduel et des canaux de couleur. Outre les régions occupant une faible surface et entourées par des gradients élevés du flot résiduel (véhicules mobiles notamment), les régions fortement texturées ne sont également pas segmentées (régions noires). Cela est particulièrement visible sur la séquence "Dubai Palace", en première ligne de la figure 5.6. La gravité de cette insuffisance est relative dans le sens où une régularisation serait susceptible de mélanger différentes classes coexistant au sein d'une unique région et de biaiser ainsi les probabilités vers des probabilités moyennes peu discriminantes. La segmentation doit ici être vue en effet plus comme une aide permettant de lisser des cartes de probabilités bruitées selon des régions homogènes, que comme un but final de segmentation totale de l'image.

Pour chaque région, un score traduit son homogénéité et son pouvoir de discrimination. Il est souhaitable en effet de ne lisser que les régions homogènes et qui séparent convenablement les différentes classes. Le score α_{reg} est calculé comme produit de deux composantes normalisées entre 0 et 1 :

- le complémentaire de la variance intra-région α_{var} des couleurs et des probabilités (les variances calculées sur chacun des canaux sont ensuite agrégées en un unique score) $\alpha_{var} = \alpha_{couleur} * \alpha_{probas} = \frac{1}{3}\frac{2}{255}\left(var_{reg}(R) + var_{reg}(G) + var_{reg}(B)\right) * \frac{2}{3}\left(var_{reg}(p_1) + var_{reg}(p_2) + var_{reg}(p_3)\right)$. Elle mesure l'homogénéité des couleurs et des probabilités obtenues sur la région. La moyenne géométrique permet ici d'obtenir un critère assez élevé si l'une des deux composantes (couleurs ou probabilités) montre une bonne homogénéité de la région. La variance des couleurs pourrait être étendue à des primitives plus complexes (de texture ou de flot résiduel par exemple) simplement.

- le complémentaire de l'entropie de Shannon empirique α_{ent} des classes normalisée $\alpha_{ent} = -\frac{1}{\log(3)}\sum_{c=1}^{3} p_c^{reg}\log(p_c^{reg})$ où p_c^{reg} représente la probabilité moyenne de la classe c dans la région. Elle exprime la puissance discriminante moyenne de la région.

Le score final $\alpha_{reg} = (1-\alpha_{var}) * (1-\alpha_{ent})$ sera ainsi élevé seulement si les deux composantes elles-mêmes le sont, soit si la région est homogène et discrimine correctement les classes (avec par exemple une probabilité moyenne sur la région de 0.8 pour l'une des trois



(a) (b)

FIG. 5.7 – *Illustration de l'incidence de la régularisation par régions sur les cartes de probabilités.* 1*ère ligne : image et segmentation partielle associée.* 2*ème ligne : cartes de probabilités sans (a) et avec (b) régularisation par régions. Canal rouge : probabilité "véhicule mobile" ; vert : "fond" ; bleu : "route"*

classes). La régularisation par région consiste ensuite à pondérer, pour chaque région de la segmentation (complète ou partielle) et pour chaque classe c, les probabilités moyennes agrégées sur l'ensemble de la région $\{p_c^{reg}\}_{c=1..3}$ avec les probabilités originales du pixel $\{p_c(x_i)\}_{c=1..3}$:

$$\forall c \in \{1,2,3\}, \quad p_c^r(x_i) \;=\; \alpha_{reg} p_c^{reg}(x_i) \;+\; (1-\alpha_{reg})\; p_c(x_i). \tag{5.1}$$

La faiblesse de cette pondération réside dans le mode d'agrégation de chaque composante, entropie de classe ou variance de couleur intégrée uniformément sur l'ensemble de la région. Ainsi l'influence d'un petit nombre de pixels troublant l'homogénéité d'une région de taille importante (par exemple 5 pixels dans une région de 500 pixels) n'influera-t-elle que peu dans le calcul de α_{reg}.

La figure 5.7 illustre la régularisation par régions ainsi effectuée. La seconde ligne montre l'évolution des cartes non lissées (a) en des cartes plus régulières (b). Certaines parties des cartes de probabilité restent inchangées. Il s'agit des régions non segmentées de l'image, qui sont fortement texturées. En revanche, plusieurs régions de végétation sont régularisées, avec disparition de fausses détections (en rouge), notamment au-dessus du "al" de la watermark centrale "Motion Gallery", ou encore dans la partie gauche de l'image (notamment le coin supérieur ainsi que le coin inférieur). Il est également intéressant de noter que le réseau routier (à dominante bleue) et le fond (à dominante verte) sont mieux séparés, avec une importante diminution des régions "vert-bleu".

5.1.4 Résultats et performances

Les performances de la classification locale selon l'approche exposée ci-dessus dépendent d'un certain nombre d'éléments :

- l'ensemble d'apprentissage ainsi que l'ensemble de test,
- un prétraitement des données,
- les primitives choisies,
- une méthode éventuelle de régularisation,
- les métriques de caractérisation des performances.

Les ensembles d'apprentissage et de test conditionnent en effet les résultats obtenus par plusieurs biais. D'une part, un écart des données test par rapport aux données d'apprentissage dégradera d'autant plus les performances qu'il est important. "Écart" doit ici être compris relativement aux primitives choisies. Il importe donc de minimiser cet écart, ce qui peut passer par :

- augmenter la similitude entre la base d'apprentissage et la base de test. Cela suppose de disposer au préalable d'informations sur la base de test. Si les séquences test sont des segments vidéo aériens en contexte urbain, il est alors préférable de sélectionner des données test d'un environnement similaire.
- un traitement préable des données : un lissage anisotropique (un filtre bilatéral par exemple) conservant les contours mais atténuant de faibles variations d'intensité ou couleur fournira de la sorte des données moins susceptibles de conduire à des cartes de probabilité bruitées.
- un choix adapté de primitives robustes à des changements d'environnement ou d'objets, de structures ou de textures des différentes classes. L'utilisation de primitives purement fondées sur la couleur, avec une variabilité intra-classe sur ces primitives supérieure à la variablité inter-classes, est ainsi peu pertinente.

Afin de bien représenter l'ensemble des instances possibles, il importe lors de la phase d'apprentissage de fournir des données suffisamment variées pour chacune des classes

choisies. D'un autre côté, il faut éviter un surapprentissage contre-productif dès que les données test s'écarteront trop des données utilisées pour l'apprentissage. Enfin, le volume des données choisies pour l'apprentissage influe directement sur le temps de calcul des classifieurs. L'application de prétraitements permet de réduire ou supprimer des hautes fréquences de l'image ou de parvenir à une estimation plus précise du flot optique (à l'origine d'une partie des primitives choisies ici). Ainsi, les objets considérés dans les séquences vidéo aériennes étudiées ont un mouvement rigide et, localement en temps, approximativement affine. Un lissage du flot résiduel permet donc de rectifier ou filtrer des valeurs aberrantes. Les occultations posent en revanche problème pour ce dernier prétraitement.

Nous avons choisi de faire varier plusieurs paramètres afin d'observer leur incidence sur les performances de classification :

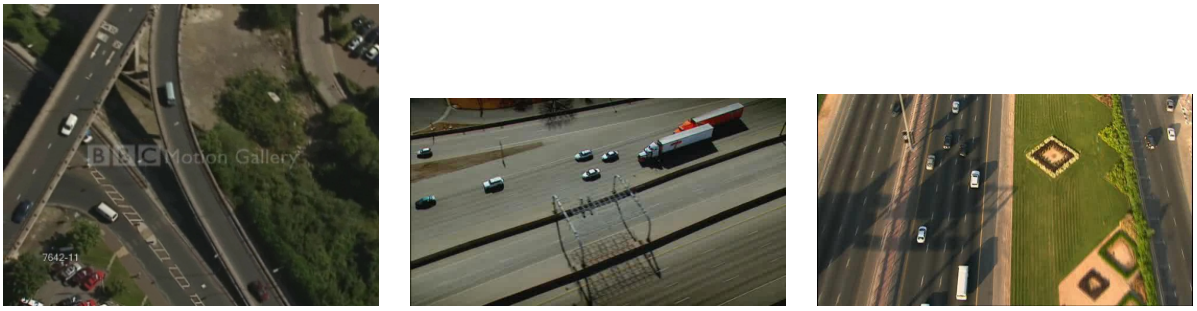
- le nombre et la provenance des images desquelles sont extraites les échantillons de test,
- le nombre d'échantillons pour chaque classe (pour chaque image),
- un lissage préalable des images,
- un lissage temporel des flots résiduels,
- une régularisation multi-échelles : différentes tailles de patchs sont utilisées pour le calcul des primitives. Les données peuvent être agrégées à la fin, les cartes de probabilités finales étant une moyenne pondérée des cartes obtenues à chaque échelle. Les primitives issues de chaque taille de patch peuvent être aussi agrégées en un seul ensemble de primitives pour une unique classification.
- une régularisation par régions : il s'agit là plus d'obtenir une cohérence spatiale suivant les régions découpées par segmentation plutôt qu'un changement d'échelle,
- une régularisation temporelle des cartes de probabilités obtenues lors de la phase de test : aux occultations près, après recalage, la cohérence temporelle des classes rajoute des contraintes permettant de corriger d'éventuelles fausses classifications. Le cas particulier de véhicules mobiles de vitesse variable, considérés comme immobiles en-deçà d'une vitesse donnée, est plus délicat. La classification se traduisant par des cartes de probabilités, la distinction entre deux classes (véhicules d'une part, fond ou routes d'autre part) reste continue.

Données test : nombre de séquences d'origine et nombre d'échantillons Nous avons utilisé 3 images issues de séquences différentes pour l'apprentissage et observé les cartes de probabilités obtenues dans différentes conditions :

- images de test tirées de la même séquence que l'une des images d'apprentissage (figure 5.10) ou d'une nouvelle séquence (figure 5.9) ;
- classification à partir des classifieurs appris sur l'une des trois images (troisième colonne des figures 5.10 et 5.9) ;
- cartes de probabilité obtenues en utilisant les trois images d'apprentissage (quatrième colonne des figures 5.10 et 5.9).

Il apparaît sans surprise que les résultats de classification sur une image tirée de la même séquence que l'image d'apprentissage (figure 5.10, troisième colonne, première ligne) sont bien plus fidèles que ceux obtenus à partir d'images tirées d'autres séquences (figure 5.10, troisième colonne, deuxième et troisième ligne).

Nous avons considéré trois méthodes permettant de prendre en compte l'ensemble des trois images d'apprentissage. Les deux premières (quatrième colonne, première et deuxième ligne des figures 5.10 et 5.9) effectuent une moyenne des cartes de probabilité obtenues à partir d'une seule image d'apprentissage (troisième colonne).



(a) (b) (c)

FIG. 5.8 – 3 *images d'apprentissage. (a) "BBC" (b) "are we changing Planet Earth" (c) "Dubai Dream séquence 15"*

La première consiste en une simple moyenne arithmétique et n'intègre donc aucun critère évaluant la pertinence de chaque carte (carte couleurs représentant les trois probabilités sur les canaux R,G et B en chaque pixel, obtenue à partir d'une image d'apprentissage).

La deuxième pondère, en chaque pixel, chacune des trois cartes selon l'entropie de classe. Cette entropie (ici normalisée entre 0 et 1) mesure simplement en chaque pixel x_i la certitude des classifieurs : $E(x_i) = -\frac{1}{\log(3)} \sum_{c=1}^{3} p_c(x_i) \log(p_c(x_i))$ où $p_c(x_i)$ représente la probabilité d'appartenance du pixel x_i à la classe c selon l'ensemble de classifieurs considérés. Une entropie nulle correspond à une certitude absolue du classifieur (probabilité de 1 pour l'une des classes et 0 pour chacune des autres), alors qu'une entropie maximale, égale à 1, correspond à une incertitude maximale (les probabilités de chaque classe sont égales à $\frac{1}{3}$). Nous avons ici considéré cette entropie comme un critère de qualité de la classification, et une pondération plus importante est donc accordée à la carte de probabilités pour laquelle l'entropie est plus faible, et ce en chaque pixel. Il s'agit donc d'une moyenne pondérée et normalisée : $p_c^{ent}(x_i) = \frac{1}{\sum_{j=1}^{3}(1-E^j(x_i))} \sum_{j=1}^{3}(1 - E^j(x_i))p_c^j(x_i)$, où $p_c^j(x_i)$ représente la probabilité de la classe c par l'ensemble de classifieurs obtenu avec l'image d'apprentissage j, et $E^j(x_i)$ l'entropie de classe normalisée pour le pixel x_i et pour ce même ensemble de classifieurs.

La troisième consiste à apprendre un ensemble de classifieurs directement sur l'ensemble des images d'apprentissage. La fusion des données d'apprentissage provenant de différents contextes afin d'obtenir un unique ensemble de classifieurs capturera ainsi les différences communes séparant les classes dans un ensemble varié de contextes. En revanche, la variabilité intra-classes sera d'autant plus importante que le nombre de contextes d'origine pour les images d'apprentissage sera élevé.

Les figures 5.9 et 5.10 illustrent ces différents points :

- les résultats obtenus lorsque les contextes des données test et d'apprentissage sont différents (figure 5.9) correspondent peu à la réalité ou vérité terrain. Ils sont biaisés en faveur d'une classe (route pour l'image d'apprentissage de la figure 5.8 (c) notamment, route et objet mobile pour celle de la figure 5.8 (b)). Les résultats obtenus avec l'image d'apprentissage 5.8 (a) sont plus proches de la réalité, les contextes correspondants sont plus proches (même séquence pour la figure 5.10, fond varié et aspect proche des routes pour la figure 5.9).
- les moyennes simples améliorent peu les résultats, mais la moyenne entropique fournit des cartes plus conformes à la réalité. Les cartes obtenues par l'ensemble de classifieurs appris sur la fusion des différentes images d'apprentissage sont plus proches des résultats attendus.
- Les résultats restent toutefois inférieurs pour une séquence donnée à ceux obtenus avec pour seule image d'apprentissage une image extraite de la séquence de test.

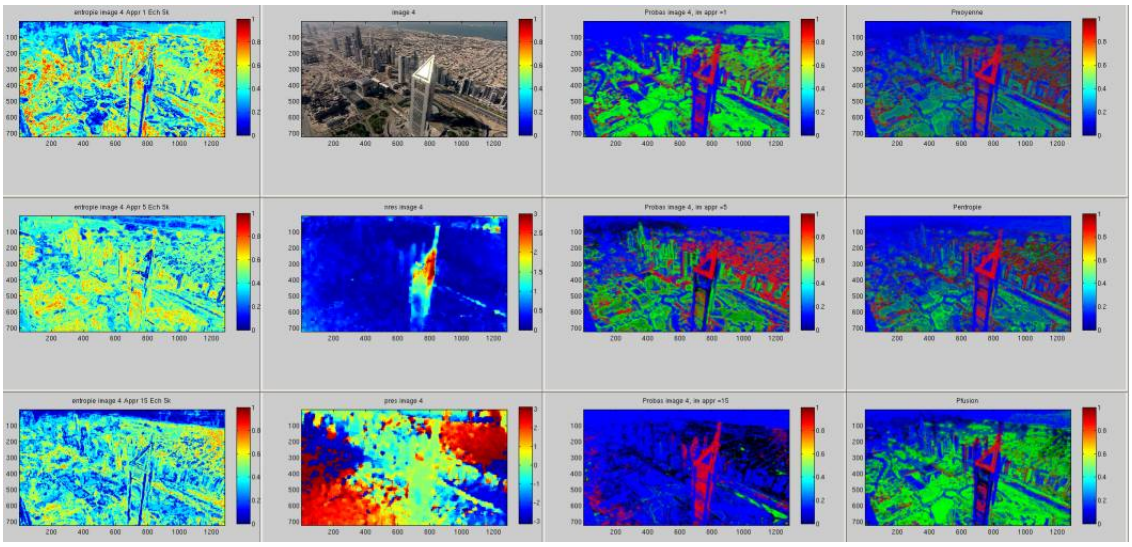


FIG. 5.9 – *Influence des données d'apprentissage. Résultats sur une image de la séquence autres que les séquences d'où ont été tirées les images d'apprentissage.* 1*ère colonne : entropies (pouvoir séparateur empirique) de classe selon l'image d'apprentissage, respectivement figure 5.8 (a), (b) et (c).* 2*ème colonne : image de test, amplitude et direction du flot résiduel associé.* 3*ème colonne : cartes de probabilités obtenues avec les classifieurs respectivement appris sur (a), (b) et (c).* 4*ème colonne : moyenne uniforme des* 3 *cartes ; moyenne entropique ; cartes obtenues avec un ensemble de classifieurs appris sur la concaténation des primitives de 5.8 (a), (b) et (c).*

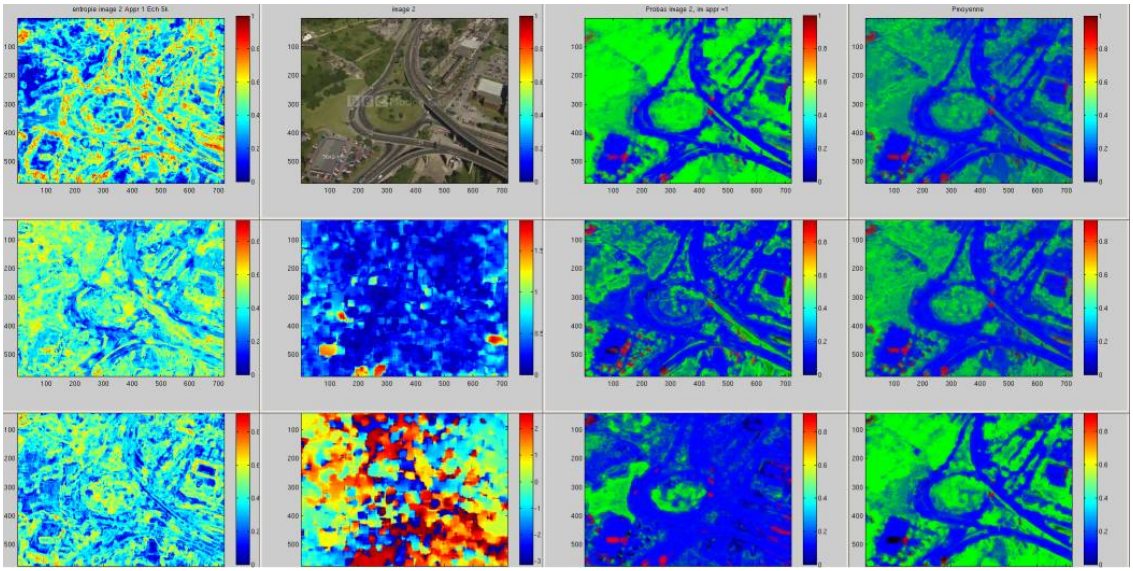
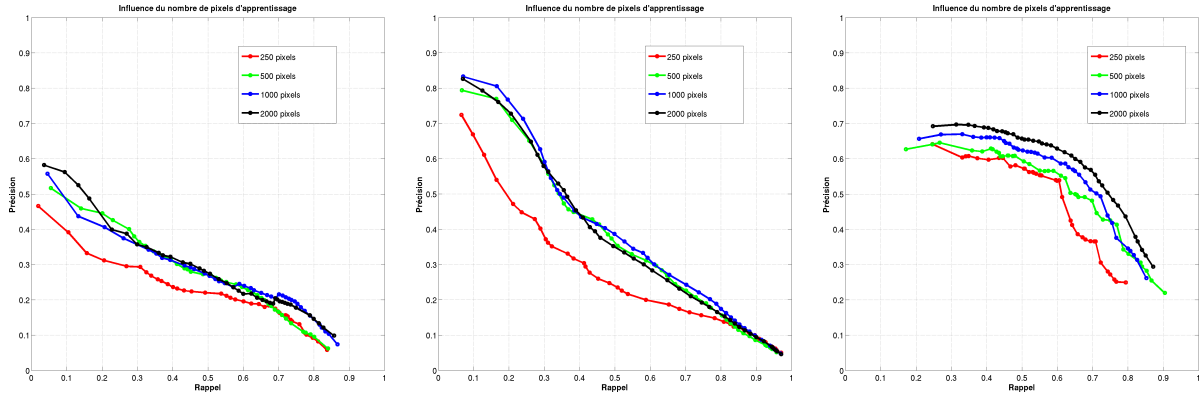


FIG. 5.10 – *Influence des données d'apprentissage. Résultats sur une image de la séquence de laquelle a été tirée 5.8 (a).* 1*ère colonne : entropies (pouvoir séparateur empirique) de classe selon l'image d'apprentissage, respectivement figure 5.8 (a), (b) et (c).* 2*ème colonne : image de test, amplitude et direction du flot résiduel associé.* 3*ème colonne : cartes de probabilités obtenues avec les classifieurs respectivement appris sur les images de la figure 5.8 (a), (b) et (c).* 4*ème colonne : moyenne uniforme des* 3 *cartes ; moyenne entropique ; cartes obtenues avec un ensemble de classifieurs appris sur la concaténation des primitives des images de la figure 5.8 (a), (b) et (c).*



(a) (b) (c)

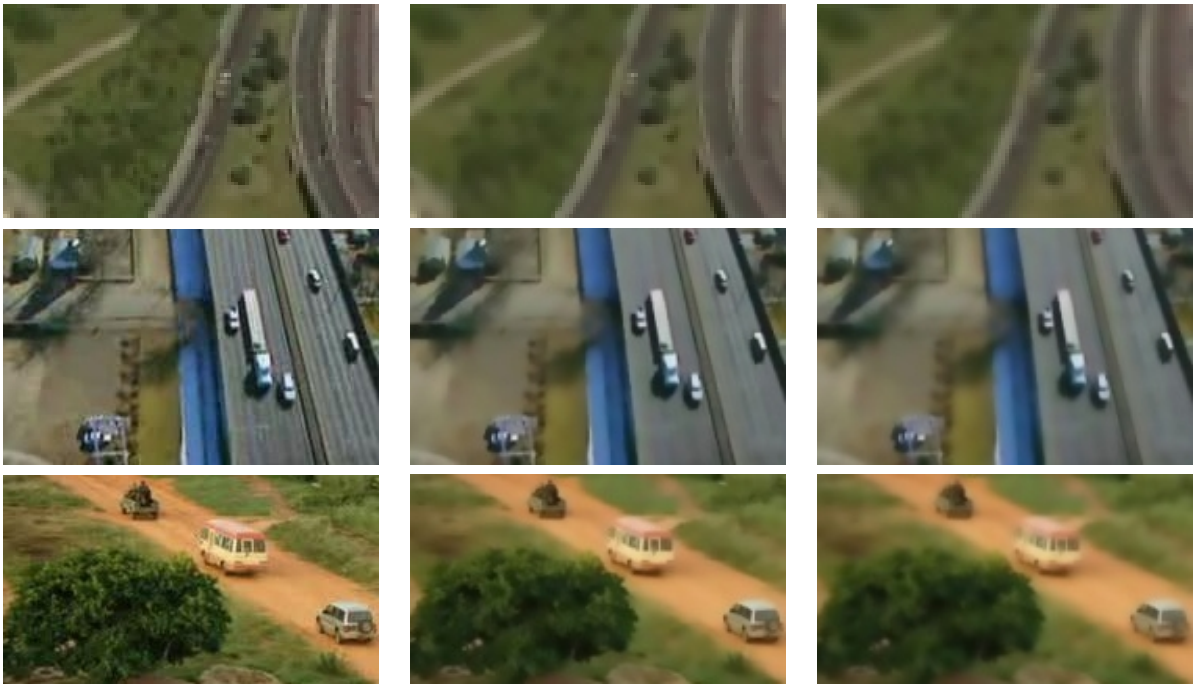
FIG. 5.11 – *Influence du nombre de pixels d'apprentissage sur les métriques de précision et rappel sur trois séquences, (a) "BBC" (b) "are we changing Planet Earth" (c) "Blood Diamond". Respectivement* 250 *(rouge),* 500 *(vert),* 1000 *(bleu) et* 2000 *(noir) pixels sont pris pour chaque classe sur l'image d'apprentissage (tirée de la séquence).*

Nombre d'échantillons La figure 5.11 illustre l'influence du nombre d'échantillons d'apprentissage sur les performances de classification. Les pixels utilisés pour l'apprentissage sont respectivement au nombre de 250, 500, 1000 et 2000 pour chaque classe. Nous constatons une amélioration des résultats pour chacune des séquences, lorsque le nombre d'échantillons augmente. Toutefois, notamment pour la séquence "are we changing Planet Earth", l'amélioration est moins notable au-delà de 500 pixels. Cela peut s'expliquer par un aspect relativement uniforme des classes, qui peuvent alors être représentées par un nombre inférieur de pixels. Prendre trop d'échantillons peut causer un problème de sur-apprentissage par rapport à l'image d'apprentissage.

Nous avons conservé pour la suite 2000 pixels par classe. Le choix de ce paramètre n'influe pas les temps de calcul lors de la phase de test car la classification s'effectue sur l'intégralité des images de test quel que soit le nombre d'échantillons choisis lors de la phase d'apprentissage. Le calcul des classifieurs est en revanche légèrement plus long, entre 3 et 5 secondes supplémentaires sur une moyenne de 20 secondes. L'arbre de décision est calculé par l'algorithme CART de Matlab et l'étape de boosting a été codée en mex (code $C++$ pour intégration dans Matlab).

Prétraitements : lissage spatial des images et temporel des flots résiduels Les images originales de la séquence peuvent présenter un bruit (notamment de compression) plus ou moins marqué. Il peut sembler logique d'effectuer un lissage spatial de ces images afin d'obtenir notamment des régions de route plus lisses et de réduire le nombre des artefacts. Nous avons utilisé un filtre bilatéral [236] permettant de conserver les contours tout en réduisant le bruit. Ce lissage effectue une moyenne pondérée des valeurs (intensité ou couleurs) des pixels voisins, dont les poids dépendent à la fois de la distance euclidienne des pixels et de leurs différences radiométriques. La figure 5.12 illustre ce filtrage pour des détails d'images de chaque séquence traitée avant et après application avec un effet plus ou moins marqué, pour une fenêtre de rayon 5, d'écart-type spatial 3 et d'écart-type en intensité 0.1 puis 0.2 (les valeurs d'intensité étant normalisées entre 0 et 1). Nous avons comparé les performances de classification avec ou sans application préalable d'un filtre bilatéral sur les images de la séquence à traiter. Les paramètres du filtre ont été choisis de façon à obtenir un compromis entre lissage et effacement des détails avec une fenêtre de rayon 5, d'écart-type spatial 0.3 et d'écart-type en intensité 0.1.

Les performances de classification présentées sur la figure 5.13 pour un lissage modéré ne montrent pas de réelle amélioration. Dans le cas de la séquence "Blood Diamond" même, la précision est, à rappel égal, inférieure. De plus, ce pré-traitement introduit tout de même un surcoût temporel non négligeable d'une dizaine de secondes par image pour

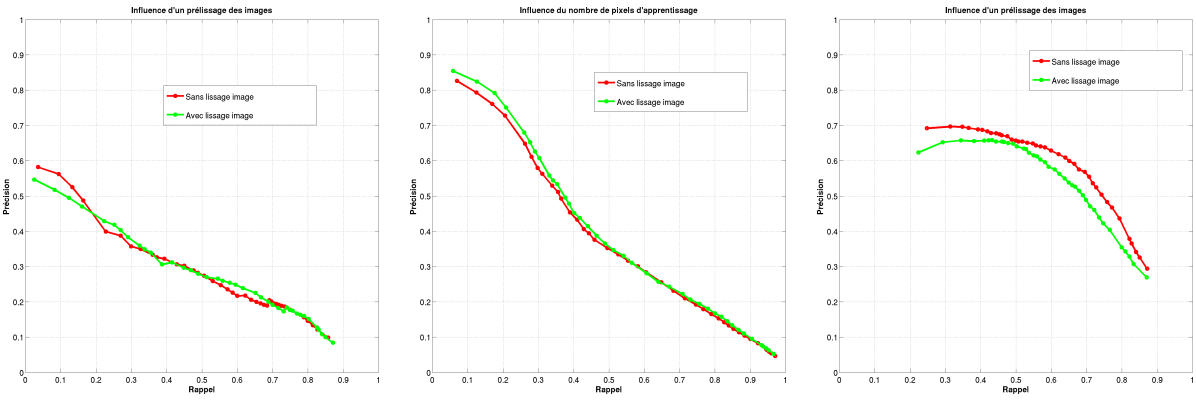


(a) (b) (c)

FIG. 5.12 – *Détails d'images avant (a) et après lissage bilatéral d'écart-type en intensité (b) modéré (0.1) et (c) plus élevé (0.2) sur trois séquences : 1ère ligne, "BBC" ; 2ème ligne, "are we changing Planet Earth" ; 3ème ligne, "Blood Diamond".*

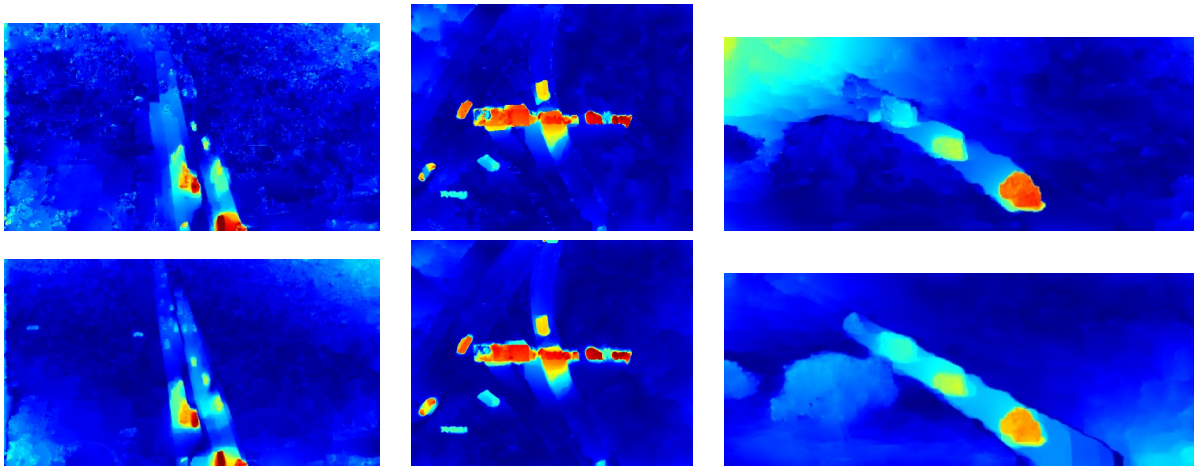
une image 256×624 et proportionnelle à la définition (nombre de pixels) de l'image. Nous n'avons donc pas conservé ce traitement.

Un autre pré-traitement possible concerne plutôt les données de mouvement représentées ici par le flot optique résiduel. Les algorithmes d'estimation du flot optique ne fournissent pas toujours des résultats corrects, en présence de larges régions homogènes ou pour des raisons de bruit ou d'occultation. Ainsi peut-on observer par exemple des différences de flot peu importantes aux frontières d'objets en mouvement ou des flots inhomogènes sur des régions correspondant à un unique objet (sans effet de parallaxe). Il est ainsi envisageable de lisser ces flots, en effectuant une moyenne temporelle du flot à l'instant t avec les flots en $t-1$ et en $t+1$. Toutefois, il faut pour cela recaler ces deux derniers flots dans le repère correspondant à l'instant t. En raison des erreurs possibles du recalage et des occultations, ce choix n'a pas été retenu.



(a) (b) (c)

FIG. 5.13 – *Influence d'un pré-lissage des images de test sur les métriques de précision et rappel sur trois séquences, (a) "BBC" (b) "are we changing Planet Earth" (c) "Blood Diamond". Respectivement sans (rouge) et avec pré-lissage (vert).*



(a) (b) (c)

FIG. 5.14 – *Amplitudes des flots résiduels avant (1ère ligne) et après (2ème ligne) lissage temporel (a) "are we changing Planet Earth" ; (b) "BBC" (c) "Blood Diamond".*

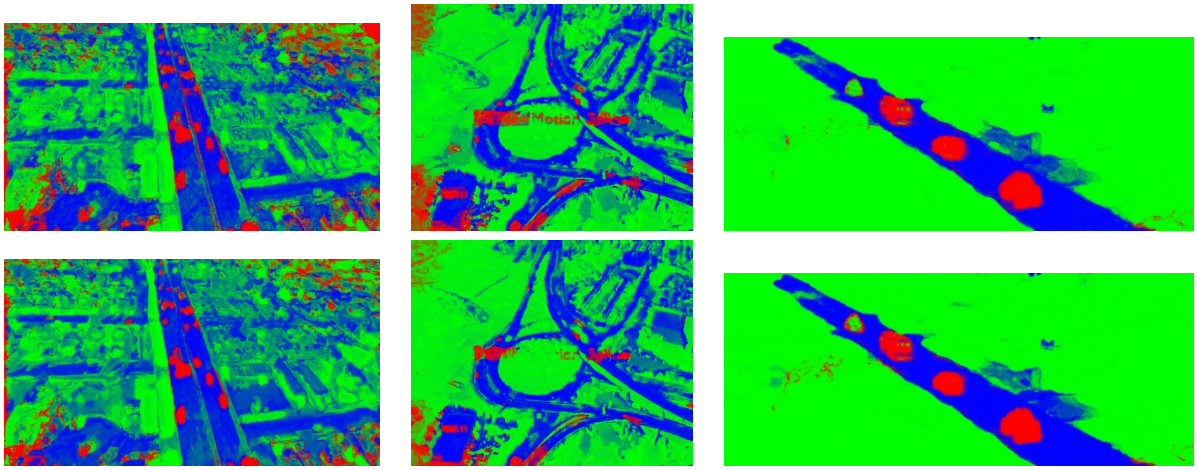
En considérant le mouvement résiduel localement affine en temps, une autre approche consiste à effectuer une moyenne des flots résiduels ayant pour origine l'image en un temps t et les images voisines aux temps $t-K, t-K+1, ..., t-1, t+1, ..., t+K-1, t+K$, en annulant le mouvement global (ici approché par une affinité) et en compensant le facteur d'échelle (le mouvement résiduel entre t et $t+2$ est, selon l'hypothèse de flot résiduel affine en temps, deux fois plus important qu'entre t et $t+1$) :

$$\mathbf{u}_{res}^{lisse}(t) = \frac{1}{2K}\left[\left(\sum_{t'=t-K}^{t-1}\frac{1}{t-t'}\mathbf{u}_{res}(t')\right) + \left(\sum_{t'=t+1}^{t+K}\frac{1}{t'-t}\mathbf{u}_{res}(t')\right)\right]$$

où $\mathbf{u}_{res}(t)$ désigne le flot résiduel en t et $\mathbf{u}_{res}^{lisse}(t)$ désigne le flot après lissage. Nous avons choisi K égal à 2. En effet, cela permet de disposer de 4 flots différents pour la moyenne tout en conservant un intervalle de temps suffisamment court (à une fréquence d'échantillonnage de 25 images par seconde) pour approcher raisonnablement l'hypothèse d'affinité locale en temps. La figure 5.14 donne quelques exemples de tels lissages temporels, en affichant les amplitudes des flots résiduels avant et après lissage. Les performances sont améliorées pour chacune des séquences, ce qui apparaît lors du passage des courbes rouges (sans traitement) aux courbes vertes (avec lissage préalable des flots résiduels) sur la figure 5.16.

Post-traitements : régularisation spatiale et temporelle des probabilités Les prétraitements ont pour but d'obtenir des données de meilleure qualité à fournir à l'algorithme de classification, par un lissage spatial des données image ou temporel des données de mouvement. Il est également possible d'appliquer ces traitements aux sorties de l'algorithme, soit les cartes de probabilités obtenues. Une segmentation en régions des images permet ainsi de régulariser les cartes de probabilités au sein de chaque région. Cette régularisation présentée au paragraphe 5.1.3 peut toutefois se révéler contre-productive si la segmentation ne sépare pas correctement les différentes classes.

La régularisation temporelle des cartes de probabilités suppose une identité des classes entre images consécutives, en suivant le flot optique. Une moyenne temporelle des différentes cartes de probabilités après recalage est ainsi susceptible de fournir des cartes moins équivoques. Deux problèmes apparaissent toutefois. D'une part, le recalage est sujet à des erreurs d'estimation de flot optique et aux occultations. D'autre part, une simple moyenne des cartes recalées, sans prendre en compte de critère de qualité, peut aussi réduire la qualité des résultats. C'est pourquoi la pondération des cartes fait intervenir un critère de qualité du flot, le critère "aller-retour". Ce critère mesure entre deux images $I1$ et $I2$ le



(a) (b) (c)

FIG. 5.15 – *Cartes de probabilités avant (1ère ligne) et après (2ème ligne) les post-traitements de régularisation spatiale et temporelle (a) "are we changing Planet Earth" ; (b) "BBC" (c) "Blood Diamond".*

TAB. 5.2 – *Différents pré- et post-traitements : temps de calcul*

Traitement	Étape	Intérêt	Temps de calcul (phase test)
Lissage spatial des images	données	faible	1 min./image (1M pix)
Lissage temporel des flots résiduels	données	élevé	1 min./image (1M pix)
Régularisation par régions	probabilités	variable	17 sec./image (1M pix)
Régularisation temporelle	probabilités	élevé	32 sec./image (1M pix)

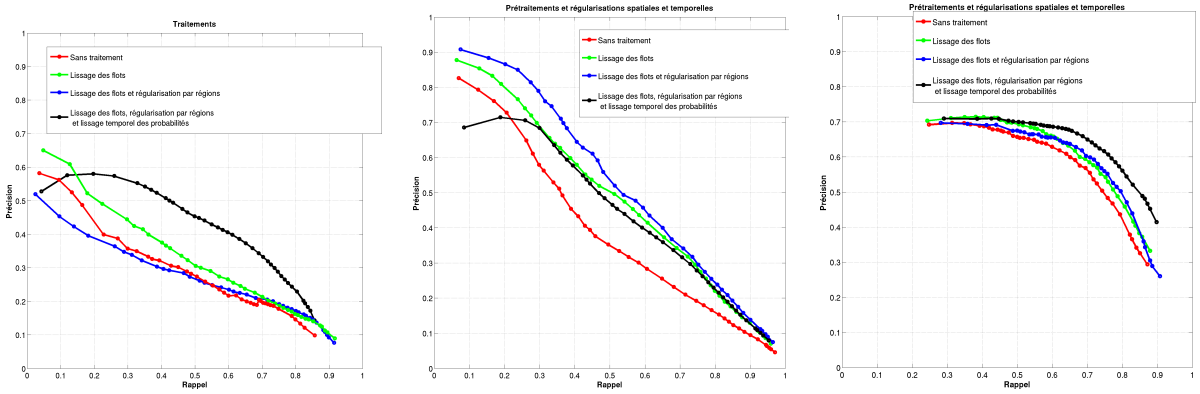
mouvement résiduel obtenu en additionnant le flot optique direct "aller" et le flot optique "retour". En notant $\mathbf{u_1}$ et $\mathbf{u_2}$ les flots optiques respectivement de $I1$ vers $I2$ et de $I2$ vers $I1$, le critère c_{AR} est défini pour un point $\mathbf{x} = (x, y)$ de $I1$ par :

$$c_{AR}(\mathbf{x}) = \mathbf{u_1}(\mathbf{x}) + \mathbf{u_2}\left(\mathbf{x} + \mathbf{u_1}(\mathbf{x})\right).$$

La moyenne n'est effectuée que si le critère est suffisamment bas, signe d'un recalage correct. Ainsi, le coefficient de pondération avant normalisation est ainsi nul en un pixel $\mathbf{x}$ de l'image centrale si $c_{AR}(\mathbf{x})$ (de l'image centrale vers l'image suivante ou précédente) est supérieur à 1 pixel en amplitude, égal à 1 si $c_{AR}(\mathbf{x})$ est d'amplitude nulle et linéairement décroissant entre ces deux valeurs de l'amplitude du critère. La figure 5.15 montre l'effet (deuxième ligne) de ces deux post-traitements sur des cartes de probabilités obtenues avec lissage temporel des flots (première ligne). Sur les séquences "are we destroying planet earth" et "BBC" (première et deuxième colonne), on observe une réduction des fausses détections, notamment dans les coins. En revanche, ce lissage temporel ajoute quelques pixels de fausse détection pour la séquence "Blood Diamond", dans la partie gauche de l'image : il s'agit d'une perturbation par les cartes des instants voisins (précédent et suivant), inexactes à cet endroit. Un lissage temporel sur un segment plus étendu dans le temps serait plus robuste, en supposant les fausses détections non cohérentes (dans le cas contraire, il s'agit probablement d'effets de parallaxe qui ne seront alors pas filtrés). Mais cela suppose un recalage de bonne qualité, ce qui devient plus difficile pour des écarts temporels plus croissants.

5.1.5 Discussion

Choix des primitives Les primitives ici utilisées restent particulièrement simples, mais permettent déjà d'obtenir une première classification approximative. L'apport de primitives plus complexes se heurte à différents problèmes. D'une part, la capture de structures



(a) (b) (c)

FIG. 5.16 – *Effets des différents pré- et post-traitements sur les performances de détection sur les séquences (a) "BBC" ; (b) "are we changing Planet Earth" (c) "Blood Diamond" : sans traitement (rouge), avec lissage préalable des flots résiduels (vert), avec lissage des flots et régularisation des cartes de probabilités par régions (bleu) et avec lissage des flots, régularisation par régions et lissage temporel (noir)*

à plus grande échelle (des éléménts géométriques tels que lignes ou des textures spatialement étendues) nécessite un nombre important de primitives (par exemple des familles de filtres orientés à différentes échelles). D'autre part, les primitives de région sont inapplicables dans le cadre d'une approche jusqu'ici pixellique. Enfin, les variabilités internes aux classes sont importantes, ce qui limite l'ensemble des primitives susceptibles de séparer efficacement, toujours dans une approche locale et pixellique, les différentes classes. Des essais incorporant des primitives issues d'une décomposition en ondelettes de l'image n'ont par exemple montré aucune amélioration.

Sélection des données test Dans notre cadre applicatif, l'analyse est effectuée "hors ligne", l'intégralité des séquences vidéo à traiter est donc disponible. Il est alors possible de sélectionner pour la phase d'apprentissage une ou plusieurs images accompagnées d'une "vérité terrain" manuellement annotée, à partir d'une ou plusieurs séquences. Plusieurs cas de figure ont été considérés, différents compromis entre précision et généricité.

En effet, dans une recherche de généricité, il semble souhaitable d'apprendre des classifieurs à partir de plusieurs sources (ici des images de séquences vidéo). Ces classifieurs seront toutefois peu performants, et ce d'autant plus que les environnements "appris" sont variés. La variabilité intra-classes sur les différentes séquences d'apprentissage devient en effet importante par rapport à la variabilité inter-classes. Par exemple, en ne considérant que les trois classes "véhicules mobiles", "routes" et "fond", la présence d'effets de parallaxe ainsi que de grandes variations d'apparence du fond au sein des séquences d'apprentissage fournissent des classifieurs bien moins discriminants pour une séquence nouvelle sans parallaxe et avec un fond encore différent, en classifiant par exemple des véhicules mobiles d'apparence et mouvement résiduel proches des structures en trois dimensions apprises comme fond.

Il est par conséquent préférable de n'utiliser lors de l'apprentissage que des images issues de la séquence de test, en sachant que les classifieurs alors appris présenteront des performances réduites sur de nouvelles séquences vidéo de contenus différents.

Estimation de la dérive du contenu La nature aérienne de cette séquence, entraîne des évolutions parfois importantes d'apparence (changement de scène, d'échelle) ainsi que des caractéristiques de mouvement. Il est alors nécessaire d'apprendre de nouveaux classifieurs sur des images plus proches des segments vidéos concernés. Une méthode simple consiste à utiliser des images de référence (utilisées pour apprendre des ensembles de classifieurs) régulièrement réparties le long de la séquence.

En revanche, un tel échantillonnage ne tient pas compte de l'évolution réelle du contenu

et traitera indifféremment des segments peu variés ou au contraire particulièrement riches. Il faut pour cela disposer d'un critère estimant l'importance de la dérive du contenu. La caractérisation du mouvement global développée au chapitre 3 indique notamment les variations d'échelle ainsi que le déplacement d'ensemble mais n'informe pas sur la variation d'apparence du contenu. Une mesure de l'évolution du contenu dans un espace de textures à déterminer compléterait ces informations de localisation et d'échelle.

Il reste toutefois difficile d'établir un lien évident entre performances de classification et une estimation de la variation de contenu qui agrège l'ensemble des classes tout en sur-représentant le fond, classe qui présente une grande variabilité interne. Comme pour chaque approche d'apprentissage, cet impact des différences entre données d'apprentissage et données de test sur les performances peut être en partie quantifié par la multiplication des évaluations en faisant varier les données. Afin de pouvoir approcher une garantie de performances (succès de détection avec un pourcentage supérieur à 90% par exemple), il faudrait ainsi rechercher un seuil critique de variation de contenu (indépendant du contenu ou automatiquement adapté en fonction) susceptible de fournir cette garantie.

Régularisation spatio-temporelle des données et des cartes de probabilités Les cartes de probabilités obtenues par classification locale sont bruitées. Cela provient de la forme même de la classification, qui ne considère qu'un patch de taille réduite autour de chaque pixel à classer, indépendamment du reste de l'image. Il importe donc de lisser ces cartes, en jouant sur plusieurs leviers.

D'une part, les primitives choisies doivent être rendues robustes aux changements d'environnement. Dans notre approche, la normalisation du flot optique résiduel par compensation des changements d'échelle fournit des valeurs adaptées aux classifieurs appris sur une échelle donnée. En revanche, cela rajoute un traitement supplémentaire qui peut dériver ou être sujet à erreur.

D'autre part, les données image et de flot optique résiduel peuvent être lissées spatialement ou temporellement. Le tableau 5.2 résume les caractéristiques des quatre traitements étudiés. Le lissage spatial apporte peu, voire dégrade les performances de classification lorsqu'elle gomme des véhicules contrastant peu avec les routes environnantes. La régularisation spatiale par régions des cartes de probabilités peut améliorer les résultats, sur les séquences "are we changing Planet Earth" et "Blood Diamond" par exemple, mais aussi les dégrader, sur la séquence "BBC". Une segmentation mélangeant au sein d'une même région des éléments de classes différentes, tels le fond avec des véhicules faiblement contrastés, tend à noyer ces derniers au sein du fond, d'autant plus que les véhicules représentent une faible portion de la région mixte.

En revanche, le lissage temporel, appliqué au préalable sur les flots optiques puis en post-traitement sur les cartes de probabilités, corrige respectivement une partie des artefacts du flot ainsi que des pixels ou régions dont les classes ne présentent aucune cohérence temporelle.

L'ensemble de ces traitements introduit un surcoût temporel non négligeable. Seuls les traitements temporels semblent donc pertinents, mais la charge calculatoire supplémentaire doit également être prise en compte, et ce d'autant plus que les données à traiter sont de grande taille, le surcoût dépendant linéairement de la définition (nombre total de pixels) des images.

Insuffisance d'une classification purement locale Une approche complémentaire afin d'améliorer la qualité des résultats (cartes de probabilités voire composantes connexes binaires) consiste à appliquer un ensemble de règles de connaissances dans une approche "top-down". Ces règles filtrent alors des pixels ou régions dont les classes présentent des incohérences sémantiques : ainsi, un "véhicule" isolé au milieu de régions de "fond", ou

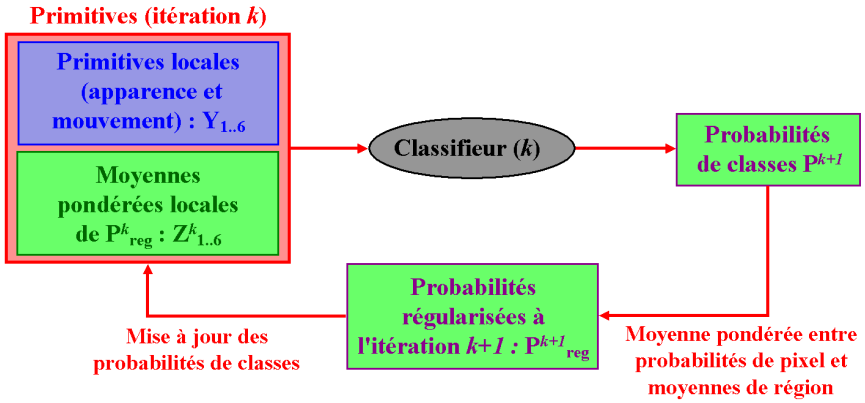


FIG. 5.17 – *Raffinement itératif des cartes de probabilités. Les cartes résultant du classifieur de l'itération précédente fournissent des primitives locales supplémentaires complétant les primitives locales d'origine fondées sur l'apparence et le mouvement (indépendantes de l'itération).*

un segment classé en "route" de taille minime peu vraisemblable. Outre ce contexte sémantique imposé par des règles de connaissance, et dans le sillage de la régularisation par régions, il apparaît pertinent d'apprendre les configurations de pixels ou structures localement. Ces deux axes font l'objet des sections suivantes, raffinement itératif et application de règles de contexte sémantiques.

5.2 Procédé itératif de raffinement

Après les traitements de régularisation appliqués aux cartes de probabilité issues de la classification locale, ces dernières restent partiellement bruitées. Ces probabilités sont alors incorporées au sein d'une approche itérative comme informations de contexte qui enrichissent l'ensemble des primitives. La première itération ne considère que les primitives locales d'apparence et de mouvement et permet d'initialiser les cartes de probabilités. Lors de la phase d'apprentissage, de nouveaux classifieurs sont appris sur le nouvel ensemble de primitives. Cet ensemble comprend la partie constante au cours des itérations des primitives d'apparence et de mouvement d'une part, les primitives de contexte qui sont mises à jour à chaque itération d'autre part. Ce procédé itératif, inspiré d'Auto-context [241], généralise ce dernier dans le cadre spatio-temporel de séquences vidéo. La définition des primitives de contexte et la régularisation par régions s'en distinguent également.

La figure 5.17 illustre le procédé. A chaque itération, un classifieur est appliqué sur un espace étendu de primitives composé d'une part de primitives d'apparence et de mouvement résiduel indépendantes du procédé itératif, d'autre part de primitives issues des cartes de probabilité obtenues à l'itération précédente. Ces itérations correspondent à l'apprentissage (respectivement l'application des classifieurs correspondants lors du processus de test) de la structure locale. L'approche supervisée permet d'adapter la classification à un environnement proche des données à traiter et le procédé itératif complète une première classification locale fondée sur la seule apparence image et de mouvement. Chaque étape d'implémentation est aisée et peut être adaptée à d'autres problèmes en modifiant les primitives en conséquence.

5.2.1 Primitives

Les primitives sont donc composées de deux sous-ensembles distincts :

- les 6 primitives $\{Y_l\}_{l=1..6}$ d'apparence et de mouvement décrites au paragraphe 5.1.1, avec $Y_l = \{Y_l(x_i)\}_{x_i}$ où x_i parcourt l'ensemble des pixels de l'image ;

– les primitives à l'itération k tirées des cartes de probabilités obtenues comme résultat de l'itération précédente $k-1$, $\{Z_m^k\}_{m=1..6}$ avec $Z_m^k = \{Z_m^k(x_i)\}_{x_i}$.

Les cartes de probabilités sont initialisées pour la première itération ou classification locale ($k = 1$) par une distribution uniforme sur l'ensemble des classes. Ainsi, en chaque pixel x_i, la probabilité associée pour chaque classe c vaut $p_c^1(x_i) = \frac{1}{3}$. Cette initialisation permet d'intégrer la classification locale décrite à la section 5.1 comme première itération du procédé. En effet, les cartes uniformes n'étant pas discriminantes, elles seront ignorées lors de l'élaboration des classifieurs lors de la phase d'apprentissage.

Les primitives correspondantes visent à capturer la distribution locale des probabilités. Cela revient à privilégier ou au contraire à pénaliser certaines configurations. Par exemple, les configurations présentant des pixels de forte probabilité "véhicule mobile" ou "route" entourées par des pixels de forte probabilité "fond" sont fortement improbables et sont signe d'artefacts à corriger en augmentant dans ce cas précis fortement la probabilité "fond" du pixel central. Les primitives proposées comprennent :

– des moyennes spatialement uniformes des probabilités du patch 3×3 pour chaque classe. Il y a donc pour chaque pixel x_i, 3 telles primitives $\{Z_m^k(x_i)\}_{m=1..3}$. Pour chaque classe $m = 1..3$, la primitive correspondante est donc donnée par : $Z_m^k(x_i) = \frac{1}{9}\sum_{x_j \in N(x_i)} p_c(x_j)$ avec $N(x_i)$ le patch 3×3 centré en x_i ;

– des moyennes de ces mêmes probabilités, pondérées par la similarité d'apparence et de mouvement entre le pixel central et les voisins du patch. L'hypothèse sous-jacente est que des pixels d'apparence et de mouvement semblables devraient appartenir à la même classe. Les primitives correspondantes $\{Z_m^k(x_i)\}_{m=4..6}$ sont décrites ci-après.

Les moyennes uniformes constituent un sous-échantillonnage spatial des cartes de probabilité. Une sélection de ces primitives revient donc à lisser spatialement les cartes de probabilité et pourrait être remplacée par un simple filtre moyenneur appliqué à l'ensemble des cartes. Ces primitives correspondent donc à un compromis entre régularisation et précision.

Les moyennes pondérées au contraire prennent en compte la structure locale de l'image et du mouvement résiduel associé. Elles privilégient un lissage spatial à partir des seuls pixels voisins d'apparence spatio-temporelle proche. Ces primitives permettent d'introduire un élément d'*a priori* semblable aux potentiels de clique dans une formulation markovienne. L'inclusion d'un terme de similarité du flot peut être en revanche contre-productive si ce dernier est estimé de façon imprécise et fait disparaître des discontinuités de mouvement. Pour un pixel donné x_i et le patch de taille 3×3 centré sur x_i, la primitive de moyenne pondérée $Z_{c+3}^k(x_i)$ pour la classe c est donnée par :

$$Z_{c+3}^k(x_i) = \sum_{x_j \in N(x_i)} W_c(x_i, x_j) p_c^{k-1}(x_j) \tag{5.2}$$

où $N(x_i)$ représente les pixels voisins du patch, $p_c^{k-1}(x_j)$ la probabilité pour la classe c du pixel x_j obtenue à l'itération précédente $k-1$. Les poids $W_c(x_j)$ sont calculés comme suit :

$$W_c(x_i, x_j) = \left(1 - \frac{|I(x_j) - I(x_i)|}{\max(I)}\right) + \left(1 - \frac{AE(\mathbf{u}(x_i), \mathbf{u}(x_j))}{180}\right) \tag{5.3}$$

où I est l'image d'intensité et $\mathbf{u}$ le flot résiduel de dimension 2. $AE(\mathbf{u}(x_i), \mathbf{u}(x_j))$ représente l'erreur angulaire entre les deux flots vectoriels $\mathbf{u}(x_i)$ et $\mathbf{u}(x_j)$ et prend des valeurs entre 0° et 180°. La fusion arithmétique plutôt que géométrique évite de trop "aplatir" les coefficients de pondération. Ces 3 primitives sont ensuite normalisées. Pour chaque pixel, l'ensemble des primitives contient donc au total 12 primitives ($6 + 2n_c$ avec n_c le nombre de classes ici

TAB. 5.3 – *Points de fonctionnement (précision égale au rappel), métriques pixelliques (cf. section 4.3.2)*

	Blood Diamond	are we	BBC
classification locale (itération 1)	0.64	0.41	0.39
itération 2	0.66	0.48	0.4
itération 5	0.6	0.43	0.23

égal à 3 mais plus élevé si on rajoute d'autres classes telles que par exemple "*structures en trois dimensions*").

5.2.2 procédé itératif

Lors de l'apprentissage, de nouveaux classifieurs "base" $\{Cl^k(c_i, c_j)\}_{i \neq j}$ pour les couples de classes (c_i, c_j) sont appris sur les données d'apprentissage à chaque itération k et fournissent par transformation logistique symétrique un classifieur multi-classes pour cette même itération k. Le principe reste identique à celui employé pour la classification locale décrite en section 5.1, mais sur l'ensemble étendu de primitives $(\{Y_l\}_{l=1..6}, \{Z^k_m\}_{m=1..6})$ présenté au paragraphe 5.2.1. Lors de la phase de test, pour chaque itération k, les classifieurs appris correspondants $(Cl^k(c_i, c_j))$ seront appliqués aux primitives étendues avant la transformation logistique.

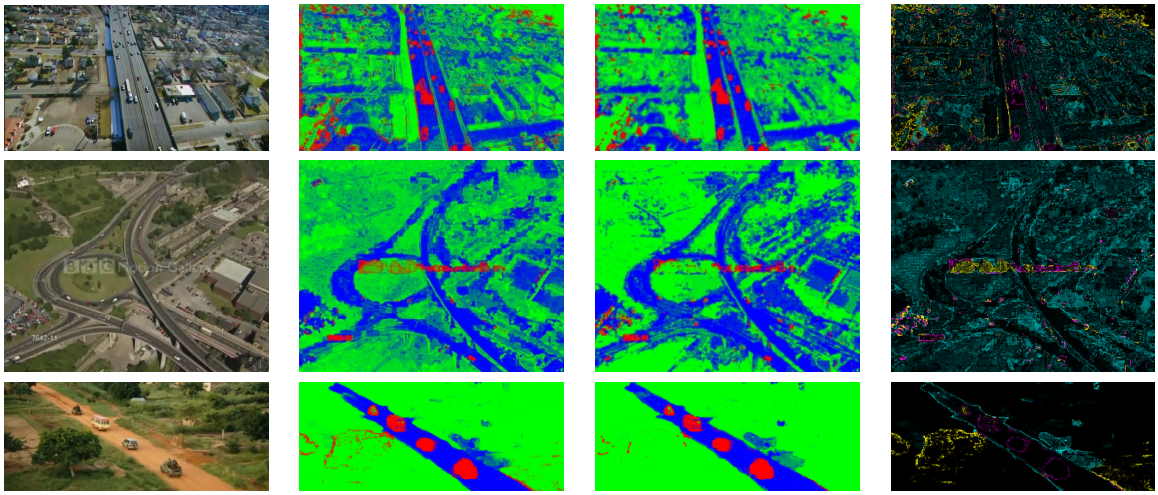
Pour la première itération $k = 1$, les cartes de probabilités sont initialisées par une distribution uniforme sur l'ensemble des classes. Les primitives de contexte correspondantes, moyennes de ces cartes, suivent donc également une distribution uniforme sur les classes et sont ainsi non discriminantes (cf. paragraphe 5.2.1). L'application des classifieurs $\{Cl^1(c_i, c_j)\}_{i \neq j}$ fournit après transformation logistique des cartes de probabilités $\{p^2_c\}_{c=1..3}$ dont découle un ensemble de primitives de contexte $\{Z^2_m\}_{m=1..6}$ participant à la classification dès la deuxième itération. Les classifieurs appris sur les données d'entraînement seront ensuite appliqués dans le même ordre sur les primitives locales d'apparence et mouvement ainsi que les résultats de classification, ou informations de contexte.

En comptant la première itération d'initialisation, le nombre total d'itérations, est supérieur à 2. Il est toutefois souhaitable de limiter le nombre maximal d'itérations, pour des raisons de temps de calcul aussi bien que d'efficacité. En effet, un nombre important d'itérations conduit aux problèmes mentionnés précédemment de redondance, de surapprentissage et de lissage trop marqué des cartes de probabilités.

Régularisation par régions Après chaque itération, les cartes de probabilités obtenues sont régularisées selon le principe détaillé au paragraphe 5.1.3. Ces cartes sont alors utilisées comme données pour fournir les primitives de contexte $\{Z^k_m\}_{m=1..6}$, comme illustré sur la figure 5.17. L'étendue spatiale des patchs 3×3 est réduite et des cartes de probabilités bruitées conduisent à des primitives peu discriminantes. La régularisation limite cet effet et les cartes obtenues après plusieurs itérations sont beaucoup plus régulières. En cas de segmentation imprécise en revanche, des détails isolés (par exemple des véhicules situés dans la partie supérieure d'une image, avec une incidence rasante de la caméra et donc de faibles dimensions) peuvent disparaître plus rapidement.

5.2.3 Résultats

La figure 5.18 ainsi que le tableau 5.3 illustrent et précisent l'évolution des résultats en fonction des itérations, selon les métriques pixelliques de précision et rappel présentées à la section 4.3.2.



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

FIG. 5.18 – *procédé itératif avec régularisation, cartes de probabilités. 1ère ligne : séquence "are we destroying Planet Earth" ; 2ème ligne : séquence "BBC" ; 3ème ligne : séquence "Blood Diamond". (a) Image test (b) Cartes après classification locale (itération 1) (c) Cartes après l'itération 5 (d) Différence image absolue entre les deux itérations*

Les différentes séquences utilisées pour évaluer l'intérêt de cette approche itérative montrent globalement une amélioration des performances, tant visuellement que par le biais des métriques de précision et rappel. La disparité des résultats selon la séquence considérée est intéressante. En effet, la scène décrite et sa complexité diffèrent notablement d'une séquence à l'autre. Ainsi, la séquence "Blood Diamond" présente peu de variation de contenu, avec un fond et une route d'apparence plutôt homogènes au cours du segment temporel étudié. En revanche, les deux autres séquences présentent non seulement une incidence importante, avec des variations de taille d'objet conséquentes, mais la taille des objets à détecter est également beaucoup plus petite. De plus, l'apparence du fond est par endroits (sur les bâtiments) très proche de celle de la route, ce qui complique fortement la classification (ici initialisée à la première itération par le biais de primitives d'apparence et de mouvement résiduel). Enfin, la séquence "BBC" comporte des variations importantes d'échelle, la taille des véhicules passant ainsi de plusieurs dizaines de pixels dans l'image d'apprentissage, à quelques pixels pour une partie du segment temporel d'évaluation : si l'échelle est partiellement prise en compte par une normalisation des primitives de flot résiduel, la taille des patchs ne varie pas et les primitives d'apparence sont ainsi moins adaptées pour la classification de scènes à une autre échelle.

Une grande partie des fausses détections disparaît au fur et à mesure des itérations et les pixels incertains, pour lesquels les probabilités de différentes (deux ou trois) classes sont proches, font apparaître une classe plus probable. C'est par exemple le cas pour la séquence "BBC", sur la deuxième ligne de la figure 5.18, pour laquelle les pixels "verts-bleus" (probabilité "route" proche de la probabilité "fond") après la classification locale sont après 4 itérations beaucoup plus précis. Une partie de la watermark centrale (le "BC" de la watermark "BBC Motion Gallery") est également régularisé en des pixels "fond" ou "route". Enfin, certains objets peu visibles ou présentant des cavités apparaissent plus clairement ou sans cavité, tels que les voitures et camionnettes blanches au centre et dans le coin inférieur gauche de l'image "BBC" ou encore, sur la séquence "are we destroying planet Earth" de la première ligne, le groupe de deux voitures et un camion blancs au centre ainsi que la voiture bleue située en-dessous de ce groupe. Les objets de taille moyenne ou supérieure (supérieure en aire à une douzaine de pixels) sont quasiment tous détectés.

Le surcoût calculatoire est linéaire en fonction du nombre d'itérations, mais l'application des classifieurs lors de la phase de test est rapide car il ne s'agit que de pondérations

locales par patchs des primitives. L'indépendance entre patchs permet de paralléliser les calculs dans un but d'optimisation supplémentaire. Les classifieurs doivent cependant être appris pour chaque itération lors de la phase d'apprentissage, l'obtention des classifieurs faibles par CART étant le processus le plus coûteux en temps.

Il apparaît dans les trois séquences, sur les métriques quantitatives, qu'un nombre trop élevé d'itérations nuit à la qualité des résultats. En l'occurrence, les régularisations par régions successives permettent d'obtenir des cartes plus lisses, au détriment des objets de faible taille qui sont peu à peu noyés parmi le fond ou la route. Un équilibre doit donc être trouvé entre cohérence des régions et conservation de ces petits objets. Toutefois, ce problème n'apparaît que dans les cas où l'incidence et l'échelle sont telles que les objets présents dans la scène sont de "petite" taille sur l'image (d'aire approximativement inférieure à une douzaine de pixels).

5.2.4 Discussion

Une telle approche présente toutefois plusieurs limites. Les relations de voisinage capturées restent locales, même si la poursuite des itérations propage de proche en proche l'information à des régions de l'image de plus en plus lointaines. Il n'est donc pas possible d'apprendre ainsi facilement des relations de longue portée. De plus, si les primitives choisies pour représenter ces relations de voisinage sont simples, quel que soit le nombre d'itérations, les interactions associées à des textures ou structures géométriques complexes ne pourront pas être apprises. En revanche, un faible nombre d'itérations (moins d'une dizaine) sera suffisant pour intégrer ces relations de premier ordre. Ainsi, le choix d'interactions limitées à des couples de pixels, utilisé dans notre méthode, permet de limiter le temps de calcul mais est insuffisant pour intégrer la différence entre des variations de formes car il agrège sur chaque patch l'ensemble des couples de pixels voisins. Des relations plus complexes pourraient ainsi être apprises en agrandissant l'espace des paramètres à l'ensemble des interactions pour des cliques d'ordre deux voire d'ordre supérieur, sans les agréger et en augmentant le nombre d'itérations. Le risque associé à cette extension de paramètres est un sur-apprentissage des relations de voisinage sur les données d'entraînement. Ce risque est limité si les données d'entraînement sont suffisamment variées et proches des données de test. Les tentatives d'intégrer de telles relations de cliques ont fourni des résultats mitigés. Si cela conduit à un nombre restreint de fausses détections, le nombre de bonnes détections est également réduit.

L'approche itérative se révèle dans certains cas contre-productive. En effet, les cartes de probabilités obtenues après classification locale, bruitées, sont régularisées sur le modèle des interactions apprises sur les données d'entraînement. Si des objets de faible taille, suite à des variations d'incidence ou d'échelle, apparaissent au cours de la séquence et sont peu contrastés, les probabilités associées tendront peu à peu vers les probabilités de la classe dominante environnante, à savoir routes ou fond. Après seuillage, les pixels ne seront donc pas étiquetés en tant que véhicules mobiles. Une piste consiste à équilibrer les probabilités du pixel lui-même avec les primitives fournies par les interactions de cliques de façon à conserver les pixels déjà étiquetés avec certitude (pour lesquels la probabilité de l'une des classes est proche de 1). Seuls les pixels "incertains", pour lesquels aucune classe ne se dégage, bénéficieraient-ils alors des itérations subséquentes.

En l'absence de régularisation, le procédé itératif converge (au moins sur les données d'apprentissage), l'erreur de classification décroissant strictement à chaque itération. En revanche, la régularisation vient perturber cet aspect du procédé. La pondération entre probabilités pixelliques et probabilités moyennes introduit en effet au sein de chaque région une dépendance vis-à-vis de l'ensemble des pixels de la région. Il existe donc un compromis entre convergence et régularisation. En pratique, les données de test sont dif-

férentes des données d'apprentissage, parfois de manière significative, et la régularisation améliore les résultats. Elle comporte toutefois le risque de supprimer des détails de petite taille et peu contrastés.

5.3 Filtrage contextuel par des règles de connaissance a priori

Les informations de contexte sémantique pertinentes dépendent du scénario considéré, soit, dans un objectif de classification, des données ainsi que des classes choisies. Ce scénario consiste ici en la détection de véhicules mobiles sur le réseau routier. Les résultats intermédiaires obtenus proviennent d'une classification itérative présentée ci-dessus sur les trois classes "véhicule mobile", "fond" et "route". Deux critères de vérification simples, introduisant des *a priori* de contexte sémantique, sont considérés :

- la distance entre un véhicule au réseau routier doit être faible voire nulle (selon la définition, objet ou par pixel, de la distance) ;
- le mouvement d'un véhicule doit être approximativement aligné sur la direction locale principale de la route environnante (les dépassements d'autres véhicules présentent une déviation par rapport à cet alignement mais la déviation reste limitée).

Pour cela, il est tout d'abord nécessaire d'extraire le réseau routier des images de la séquence, puis estimer l'axe local du réseau. Cette étape fait l'objet du paragraphe 5.3.1.

5.3.1 Réseau routier : obtention et estimation d'axe

La qualité du réseau routier estimé est primordiale lors de cette étape. En effet, des segments routiers non détectés amèneraient à filtrer des objets proches alors considérés comme "hors route". Une estimation robuste de la direction locale du réseau est nécessaire pour disposer d'un critère d'alignement fiable (mesure découlant directement de l'écart entre l'orientation locale du réseau et celle du flot résiduel après compensation du mouvement dominant). Hormis les cas particuliers de croisement, pour lesquels aucune direction principale n'est définie, l'erreur d'estimation de l'angle est directement liée au pouvoir discriminant du critère d'alignement. Dans le pire des cas, une erreur de 90° signifie que le critère est inutile. Les obstacles sont multiples, pour l'obtention du réseau et la détermination de la direction principale : "trous" ou discontinuités dus par exemple à un véhicule ou ombre de bâtiment, ou encore structures ambiguës telles que des toits de bâtiments d'apparence semblable à celle des routes. Les irrégularités dans les contours du réseau obtenu faussent l'estimation de la direction principale et donc le critère d'alignement. Les défauts de détection ou au contraire les bâtiments détectés comme routes amènent respectivement à filtrer des objets véritables ou conserver des artefacts. Il est possible d'améliorer la qualité d'estimation de la direction principale en ajoutant des *a priori* de contours rectilignes et parallèles mais cela suppose de disposer déjà d'un réseau peu bruité.

Nous considérons ici ne pas pouvoir établir de réseau routier à partir de cartes correspondant à la zone observée, par méconnaissance des paramètres de vol, difficultés de recalage ou encore imprécision ou absence de cartes. De nombreuses approches s'attachent à l'extraction automatique du réseau routier à partir de séquences vidéo. Nous proposons ici deux méthodes simples.

Obtention du réseau routier La première consiste à seuiller la carte de probabilité pour la classe "routes" obtenue par le procédé itératif. Cela est rapide mais fournit un réseau bruité avec des contours irréguliers. Afin d'éviter l'apparition de "trous" au sein du réseau routier dus à la présence de véhicules, les régions classées comme "véhicules mobiles" (avec une probabilité de classe correspondante supérieure à 0.5 par exemple) et à proximité des

régions classées comme "routes" sont intégrées au réseau retenu. Des exemples de réseaux routiers ainsi obtenus sont présentés figure 5.19 (b).

Une partie des artefacts, groupes de points isolés ou "trous", peut être rectifiée par des opérations morphomathématiques (figure 5.19 (c)) mais il est difficile d'adapter automatiquement les paramètres correspondant au contexte. En effet, un filtrage sur la largeur des zones détectées peut sembler suffisant pour supprimer les structures en trois dimensions classées comme "routes", mais il existe également des routes larges à plusieurs voies.

Conserver l'ensemble du réseau routier estimé conduit à une carte de distance dont les valeurs sont sous-évaluées, ce qui correspond à un relâchement de la contrainte de distance. Cela permet de ne supprimer aucun (ou peu de) véhicule mobile. En revanche, des fausses détections sont conservées. Les directions principales estimées localement sont peu fiables et le critère associé est susceptible de dégrader les performances de détection.

Une autre approche consiste à extraire directement du contenu de l'image le réseau routier. En suivant l'hypothèse que l'apparence des routes change peu au cours d'une même séquence vidéo, nous avons choisi de modéliser ces dernières au travers de 7 primitives : l'intensité I, les trois canaux de couleur R, G et B ainsi que les rapports $\frac{R}{G}$, $\frac{R}{B}$ et $\frac{G}{B}$. Ces trois dernières primitives permettent d'apprendre la couleur de la route sans avoir recours à une distribution pluridimensionnelle lourde, en combinant simplement un apprentissage sur chaque primitive.

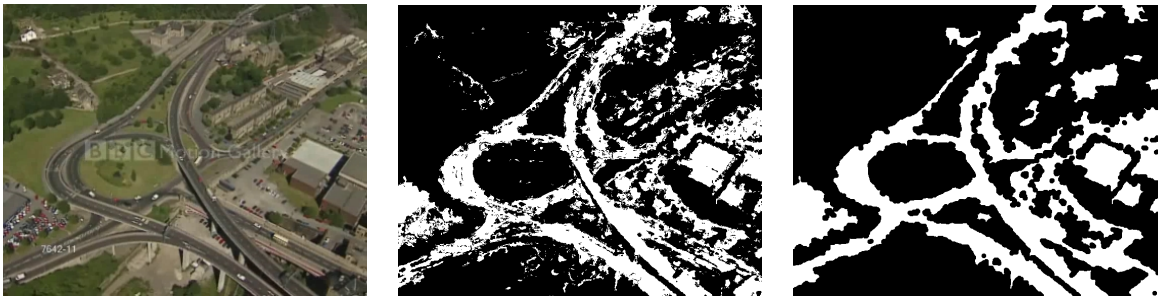
Pour chacune de ces primitives, un histogramme est établi sur une image d'apprentissage (plus précisément sur les pixels de cette image étiquetés manuellement comme appartenant à la classe "routes"). L'apparence des routes étant relativement uniforme, les histogrammes présentent chacun des "pics".

Pour chaque primitive, les queues de la distribution empirique obtenue sont écartées, ce qui revient à poser un seuil inférieur et un seuil supérieur encadrant les valeurs les plus probables. Dans l'éventualité où les distributions seraient multimodales, il faudrait alors extraire un seuil inférieur et un seuil supérieur par pic sur le même principe.

Pour chaque image de la séquence, un double seuillage (dans le cas d'une distribution empirique présentant un unique mode) fournit une carte binaire par primitive. Le produit de ces 7 cartes donne un réseau binaire grossier.

Plusieurs opérations morphomathématiques sont alors effectuées sur ce réseau binaire. Une première fermeture permet de lisser les contours et de combler d'éventuels "trous" de faible taille. Les "petites" régions, fines et d'aire faible, sont traitées comme du bruit et filtrées. Une illustration des réseaux obtenus par cette méthode est présentée figure 5.20, à comparer avec les réseaux obtenus à partir de la classification, figure 5.19.

En l'occurrence, il peut sembler étrange que la qualité du réseau routier obtenu par histogrammes soit supérieure à celle du réseau obtenu par classification. L'approche par histogrammes est en effet plus simple et n'introduit aucun élément de contexte. Toutefois plusieurs éléments peuvent justifier cette différence. Le réseau routier est extrêmement bien représenté par une couleur donnée, du gris foncé, ce qui explique les bons résultats obtenus par histogrammes. L'approche par classification est également multi-classes et prend ainsi en compte la séparation entre routes et fond d'une part, routes et objets mobiles d'autre part. La classification est fondée sur des primitives simples extraites de patchs (moments d'ordre 1 et 2 sur ces patchs) et les données d'apprentissage consistent simplement en quelques centaines d'échantillons pour chaque classe. Enfin, les primitives de mouvement sont particulièrement importantes pour la classe de véhicules mobiles : si une erreur d'estimation de flot conduit à un flot résiduel important sur une portion de route, la probabilité de la classe "véhicules mobiles" correspondante sera élevée. En revanche, l'analyse purement colorimétrique considérera des toits de bâtiment de même colorimétrie que la route comme étant de la route, au contraire de l'approche par classification si l'effet de parallaxe (et donc le flot résiduel) est suffisamment marqué.



(a) (b) (c)

FIG. 5.19 – *Estimation de réseau routier par classification. (a) Image de séquence vidéo (b) Réseau grossier (c) Réseau final*

(a) (b) (c)

FIG. 5.20 – *Estimation de réseau routier par histogrammes. (a) Image de séquence vidéo (b) Réseau grossier (c) Réseau final*

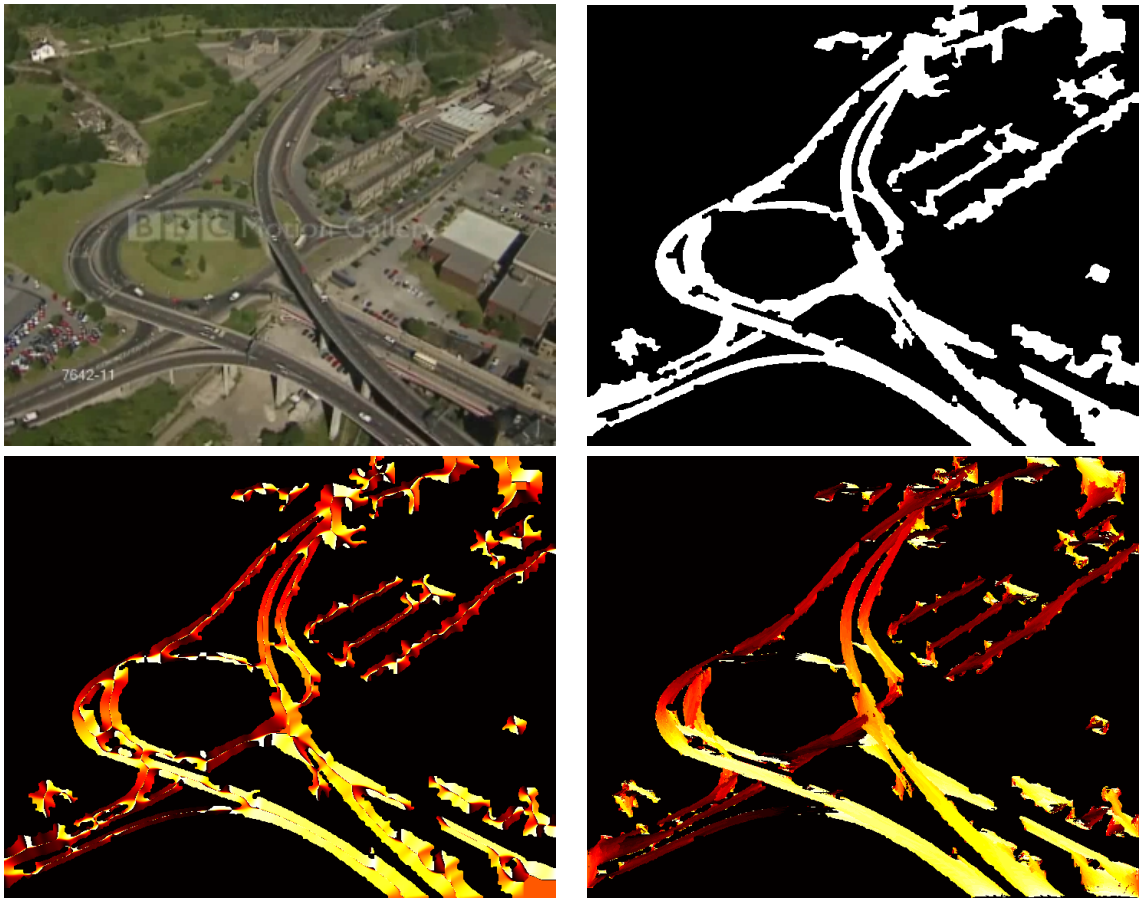
Estimation de la direction principale locale du réseau routier L'hypothèse implicite est que le réseau routier présente localement une direction principale, dans le sens de la longueur qui se trouve être le sens de circulation. Les croisements ou carrefours sont des cas particuliers pour lesquels il faut relâcher le critère d'alignement. En effet, outre les deux (ou plus) directions principales associées aux voies qui se rejoignent au croisement, les véhicules y circulant sont susceptibles de changer continûment de direction et le critère d'alignement n'y est donc pas pertinent. Il faut donc que la méthode estimant la direction de la route soit en mesure de détecter les croisements, ou plus généralement, de fournir une valeur de confiance afin de relâcher en conséquence le critère d'alignement.

Une première méthode consiste à estimer la direction des gradients sur les bords de route. En effet, sous l'hypothèse évoquée ci-dessus d'une route localement rectiligne, l'orientation locale des gradients des contours ou bords est orthogonale à la direction principale du segment de route considéré. Le réseau routier obtenu étant bruité, les gradients locaux présentent de multiples orientations. Il importe donc de lisser au préalable le réseau, par exemple à l'aide d'un noyau gaussien.

Si ce prétraitement est efficace pour des routes isolées, il brouille en revanche les contours aux abords des intersections ou lorsque plusieurs routes sont proches, ce qui conduit à des gradients inexploitables. De plus, la taille et surtout la variance du noyau gaussien nécessaires dépendent du niveau de dégradation du réseau et doivent être ajustés manuellement pour des résultats optimaux.

Cette méthode présente ainsi plusieurs défauts et la carte des orientations obtenue n'est associée à aucun indicateur de qualité. En revanche, elle est très rapide, la convolution par un noyau gaussien et le calcul de l'orientation des gradients étant des opérations peu coûteuses (un calcul en GPU est envisageable pour un noyau de grande taille avec un rayon de plusieurs dizaines de pixels).

Une méthode plus fastidieuse apporte des résultats plus précis ainsi qu'un critère de fiabilité. Elle consiste à filtrer la carte binaire du réseau routier par un ensemble de masques



(a) (b)

FIG. 5.21 – *Estimation de la direction principale locale du réseau routier. 1ère ligne : (a) image de séquence vidéo ; (b) réseau routier estimé par histogrammes. 2ème ligne : angle trigonométrique de la direction estimée (a) Par lissage par noyau gaussien et direction des gradients ; (b) par filtres orientés. Les valeurs sont prises dans* $[0, 180[$ *en degrés. Les couleurs "chaudes" (ici jaune - blanc) correspondent à des valeurs proches de* 180 *et les valeurs "froides" (ici rouge foncé) à des valeurs proches de* 0*.*

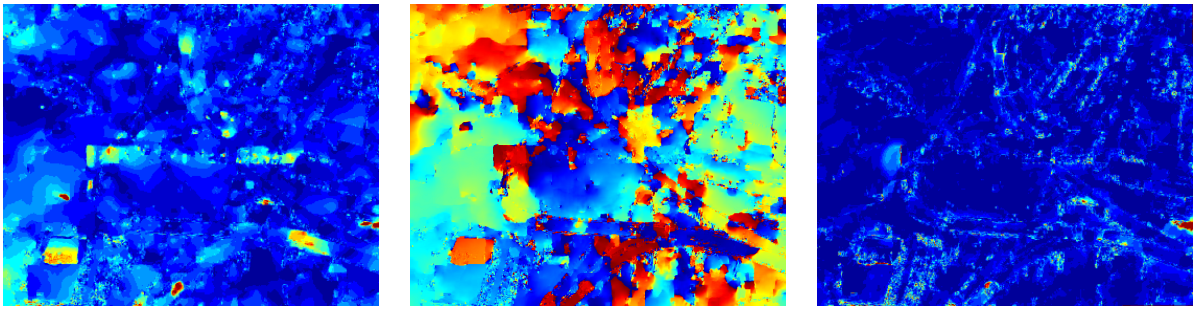
d'orientation et de longueur variables. Le maximum de la réponse correspond à l'orientation locale, pour une longueur de masque suffisante, c'est-à-dire qui permet de départager les réponses (pour un pixel situé au centre de la route dans sa largeur, une longueur trop faible, de l'ordre d'une demi-largeur de route, conduira à un ensemble de réponses identiques pour l'ensemble des orientations, la structure locale du réseau binaire étant à cette échelle isotrope).

Cette méthode est plus lente et donne des orientations également bruitées aux abords des intersections. Le bruit est toutefois moins marqué que pour la première méthode et une mesure du pouvoir discriminant (pic plus ou moins marqué de la réponse en fonction des orientations des masques) est disponible. Une illustration de cette approche est donnée figure 5.21, deuxième ligne, (b) et à la figure 5.23 (sur laquelle (c) représente le score de fiabilité).

5.3.2 Filtrage contextuel

Les critères de connaissance *a priori* sont appliqués aux cartes de probabilités issues du procédé itératif de classification. Les différentes étapes permettant d'obtenir l'image finale de détection peuvent être ainsi résumées :

- obtention du réseau routier,
- calcul de la carte de distance au réseau et de la carte d'orientation locale de ce même réseau,



(a) (b) (c)

FIG. 5.22 – *Flot résiduel après compensation affine du mouvement dominant sur une image de la séquence "BBC" : (a) amplitude (b) direction (c) critère d'erreur aller-retour. Les valeurs affichées sont prises dans* $[0,2]$ *pour l'amplitude en pixels,* $[-180,180[$ *en degrés pour la direction,* $[0,1]$ *pour le critère d'erreur en pixels. Les couleurs "chaudes" (ici rouge foncé) correspondent aux valeurs maximales, les couleurs "froides" (ici bleu foncé) aux valeurs minimales.*

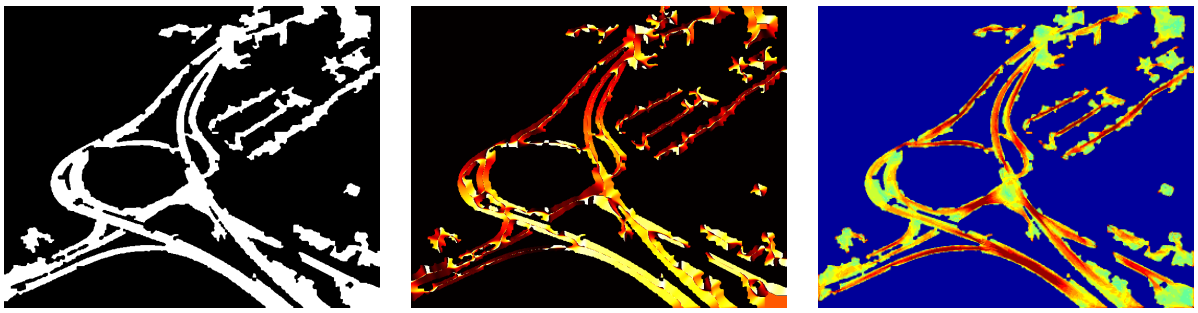
- calcul des cartes de scores de distance et d'alignement correspondants,
- calcul du score final pour la classe "véhicule mobile",
- obtention d'une image de détection binaire par seuillage du score,
- application de traitements morphomathématiques classiques sur l'image binaire (suppression régions / remplissage de trous de dimensions réduites).

Plusieurs de ces points méritent d'être explicités. Ainsi, la distance au réseau et l'orientation locale peuvent être calculées de diverses manières. Nous avons choisi d'utiliser une simple distance euclidienne et l'estimation par filtres orientés décrite précédemment pour l'orientation du réseau. Les cartes de distance et d'orientation sont transformées en des scores de valeurs comprises entre 0 et 1 par des fonctions de seuillage doux à deux paramètres contrôlant respectivement le seuil et la pente. Le score de distance est une fonction directe de la distance au réseau, le score d'alignement est fonction de l'écart angulaire entre la direction du flot résiduel et l'orientation locale du réseau. Un score nul indique que le critère correspondant invalide totalement l'appartenance à la classe "véhicules mobiles". Au contraire, un score de 1 indique que le critère confirme cette appartenance. Mathématiquement, ces scores sont intégrés par un modèle produit pour le calcul du score final : nous avons défini ce dernier comme le produit de la probabilité "véhicules mobiles" après classification itérative, par chacun des scores de contexte.

L'intérêt affiché des critères de connaissance est de réduire les fausses détections, tout en conservant les véritables détections. En pratique, suivant la qualité du réseau et de l'estimation des directions associées, ainsi que la précision du flot résiduel, les critères peuvent conduire à une dégradation des performances. Il est donc important de prendre en compte des informations complémentaires de fiabilité, du flot résiduel et du réseau. Ainsi, seules les régions de l'image pour lesquelles le flot et le réseau routier sont jugés fiables seront prises en compte pour l'application des critères de contexte.

5.4 Résultats

Les figures suivantes donnent des exemples des différentes données et résultats obtenus lors de cette étape de filtrage contextuel. La figure 5.22 montre un exemple de flot résiduel (amplitude et direction) ainsi que le critère aller-retour associé, défini à la section 5.1.4 au paragraphe "Post-traitements". La direction est utilisée pour le calcul du score d'alignement, mais il faut également disposer d'un critère de qualité ou "confiance". Ce critère dépend à la fois de la précision du flot (critère aller-retour) et de celle de la carte d'orientation du réseau routier.



(a) (b) (c)

FIG. 5.23 – *(a) Réseau routier estimé (b) Orientations correspondantes (c) Score de fiabilité de l'orientation (estimation de l'amplitude du "pic" de réponse de convolution par des filtres orientés).*

La figure 5.23 illustre, à partir d'une estimation du réseau routier en (a), les directions calculées en (b) et le score de fiabilité de ces orientations, soit le pic de la réponse au filtre d'orientation, en (c). Si les éléments fins du réseau routier sont associés à un score élevé (en rouge foncé), les embranchements ainsi que les "blocs" de route (provenant d'une estimation imprécise du réseau) présentent des scores réduits (en vert voire jaune).

La figure 5.24 représente les scores correspondants de distance et d'alignement. Ces cartes de scores appellent plusieurs commentaires. D'une part, le score de distance est imprécis : afin de compenser les erreurs d'estimation du réseau, une marge "de sécurité" est nécessaire afin de ne pas filtrer accidentellement des véhicules situés sur une route non détectée. Cela est notamment le cas pour la portion de route masquée par le car blanc dans la partie inférieure droite de l'image ou celle masquée par la watermark dans la portion inférieure gauche de l'image. Les images (a) et (b) sur la première ligne de la figure 5.21 montrent les "trous" correspondants du réseau routier. L'image de score de distance (b) de la figure 5.24 illustre sur ces exemples l'intérêt d'un relâchement du score de distance. D'autre part, le score d'alignement (forcé à une valeur de 1 en dehors du réseau routier) apparaît bruité. L'orientation locale estimée peut en effet être elle-même bruitée et peu fiable, notamment aux intersections et aux abords des artefacts du réseau : l'image du réseau sur la figure 5.24 (a) est en effet dentelée, notamment à cause des occultations créées par la présence des véhicules. De plus, la direction du flot résiduel présente elle-même des variations rapides.

La figure 5.25 souligne l'intérêt du filtrage contextuel par les règles de connaissance *a priori*. Les cartes de probabilité objet (ici en niveaux de gris) obtenues après filtrage ne sont pas parfaites mais sur chacune des trois images (extraites des différentes séquences étudiées), on constate une diminution importante des fausses alarmes : véhicules fixes et bâtiments sur les coin inférieur gauche et inférieur droit ainsi qu'une partie importante de la watermark sur la séquence "BBC" (première ligne) ; éléments de végétation et l'arbre sur la séquence "Blood Diamond" (deuxième ligne) ; bâtiments aux coins de la séquence "are we changing Planet Earth" (troisième ligne). Sur cette dernière séquence notamment, le réseau routier occupe une majeure partie de l'image. La marge de sécurité supplémentaire évoquée ci-dessus pour le score de distance rend ainsi celui-ci peu efficace sur cette séquence. Le score d'alignement apporte également peu car le réseau est bruité, occulté par la végétation et les ombres, ce qui conduit à des orientations locales peu fiables. Afin d'éviter une suppression erronée de vraies détections, le score d'alignement correspondant à ces régions est élevé.

Les tableaux 5.4 et 5.5 détaillent de manière plus quantitative l'apport des différentes étapes. Les métriques détaillées à la section 4.3.2 sont utilisées afin de fournir une appréciation quantitative fondée sur les pixels ou même par régions / objets.

Nous comparons les résultats à ceux d'une autre méthode ou baseline fondée sur la

TAB. 5.4 – *Points de fonctionnement (précision égale au rappel), métriques pixelliques (cf. section 4.3.2)*

	Blood Diamond	are we	BBC
classification locale (itération 1)	0.64	0.41	0.39
classification locale avec règles de connaissance	0.82	0.52	0.47
itération 2	0.66	0.48	0.4
itération 2 avec règles de connaissance	**0.83**	0.56	**0.50**
itération 5	0.6	0.43	0.23
itération 5 avec règles de connaissance	0.81	**0.58**	0.3
baseline fondée sur un MRF	0.5	0.41	0.25

TAB. 5.5 – *Mesures de précision et rappel objet (cf. section 4.3.2). Les nombres de véhicules indiqués sur la vérité terrain correspondant à la somme des nombres de véhicules sur l'ensemble des images*

	Précision	Rappel
Séquence "Blood Diamond" (288 véhicules sur la vérité terrain)		
classification locale (itération 1)	0.51	0.57
itération 2 avec contexte (règles de connaissance)	0.50	**0.81**
baseline fondée sur un MRF	0.48	0.35
Séquence "are we" (854 véhicules sur la vérité terrain)		
classification locale (itération 1)	0.5	0.36
itération 2 avec contexte (règles de connaissance)	0.49	**0.55**
baseline fondée sur un MRF	0.54	0.16
Séquence "BBC" (420 véhicules sur la vérité terrain)		
classification locale (itération 1)	0.512	**0.39**
itération 2 avec contexte (règles de connaissance)	0.531	0.345
baseline fondée sur un MRF	0.43	0.302

minimisation d'un champ de Markov aléatoire (MRF) à deux classes (véhicules et le complémentaire) utilisant des cliques d'ordre 2. Les potentiels unaires sont définis comme étant les amplitudes du flot résiduel normalisées par l'amplitude maximale des objets mobiles sur l'image d'apprentissage (après annulation du facteur d'échelle introduit par le mouvement global). Les potentiels de clique binaires, pour l'ensemble des couples reliant le pixel étudié avec ses 8-voisins, sont définis par des moyennes pondérées des différences de couleur et des erreurs angulaires définies au paragraphe 5.2.1 entre les flots résiduels du couple de pixels voisins considéré.

La minimisation de l'énergie globale, sur l'ensemble de l'image, est obtenue par une approche de coupure minimale / flot maximal [38], ce qui fournit une carte de détection binaire. Plusieurs cartes, donnant différents couples de précision - rappel, sont obtenus en faisant varier le poids des potentiels unaires dans l'énergie globale : plus ce poids est élevé, plus le taux de non-détection est bas, mais cela augmente également le taux de fausse alarmes (et donc fait diminuer la précision).

Le critère de distance à la route améliore de façon significative les résultats. Pour la séquence "are we changing Planet Earth", l'amélioration est plus marquée à l'itération 5, ce qui corrige en partie le sur-apprentissage (les performances sans apport de contexte sont sinon inférieures à l'itération 5). Les fausses alarmes après la classification locale sont dues à la fois à une estimation erronée du flot optique résiduel à cause d'artefacts de compression (séquence "BBC") ou une compensation affine inexacte (dans les coins inférieur gauche et supérieur droit pour la séquence "are we changing Planet Earth", cf. figure 5.14 (a)). Les fausses détections dues à la parallaxe sont rectifiées en partie (les bâtiments ou les poteaux dans les documentaires "BBC" et "are we changing Planet Earth" ainsi que l'arbre dans le film "Blood Diamond"). Cependant, les structures en trois dimensions proches de segments de route détectés ne sont pas filtrés.

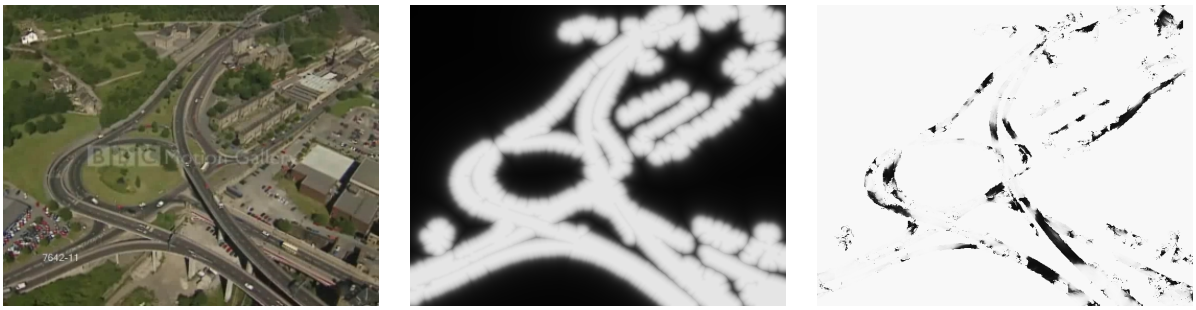
Dans la majorité des cas, l'application de l'ensemble des traitements fournit les meilleurs résultats. Cela n'est toutefois pas vérifié dans le cas de métriques objet pour la séquence "BBC". Dans cette dernière, les véhicules de taille réduite dans la partie supérieure de la séquence ne sont pas détectés à la suite du filtrage morphologique des petits objets, ce qui dégrade les performances de rappel objet. La métrique de rappel pixellique est moins impactée car ces véhicules représentent une part minime dans le nombre de pixels total de l'ensemble des objets.

Le score d'alignement se révèle peu utile. La qualité du réseau routier estimé est souvent insuffisante pour garantir une orientation fiable du réseau, sans compter les erreurs possibles du flot résiduel.

Enfin, le sur-apprentissage apparaît rapidement, au bout de 5 itérations voire moins suivant la séquence considérée. Il semble donc préférable de limiter le nombre d'itérations à 1 ou 2 itérations supplémentaires après la classification locale, et de filtrer les fausses alarmes par les règles de contexte sémantique (ici de distance et alignement par rapport à la route) et en imposant une cohérence temporelle. La cohérence temporelle peut intervenir à plusieurs niveaux : entre les cartes de probabilité pour chacune des classes, mais aussi entre les apparences et les champs de déplacement des régions détectées comme objets (dont les mouvements doivent être localement affines en temps).

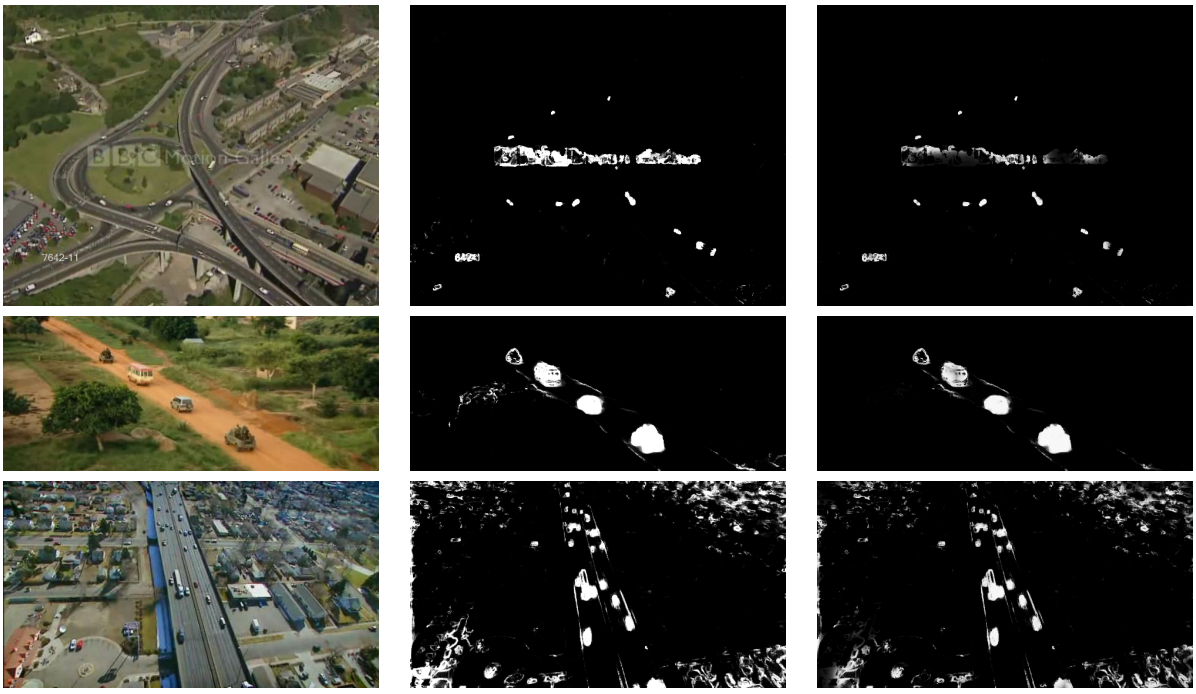
5.5 Discussion

Plusieurs remarques découlent de l'étude des résulats présentés ci-dessus. Tout d'abord, la qualité du réseau est primordiale. Un réseau de mauvaise qualité conduit à des scores de contexte peu fiables qui peuvent masquer des détections véritables ou au contraire conserver des fausses détections. Cela est particulièrement important pour le score d'alignement. En effet, ce score dépend à la fois de la précision du flot résiduel et de l'orientation estimée du réseau routier. Si le flot résiduel est dans l'ensemble de bonne qualité, le calcul



(a) (b) (c)

FIG. 5.24 – *(a) Image test (b) Score de distance à la route (c) Score d'alignement entre direction du flot résiduel et orientation du réseau routier. Les valeurs sont dans l'intervalle* $[0, 1]$ *et sont d'autant plus élevées que le pixel associé est lumineux.*



(a) (b) (c)

FIG. 5.25 – *(a) Image test ; probabilité objet (b) avant filtrage (c) après filtrage.* 1*ère ligne : séquence "BBC". 2ème ligne : séquence "Blood Diamond". 3ème ligne : séquence "are we changing Planet Earth"*

de l'orientation locale du réseau est peu robuste à des défauts du réseau. L'introduction d'un critère de fiabilité pour l'orientation du réseau permet ainsi de pondérer l'influence du score d'alignement. Le score de distance est plus robuste à des défauts et la présence de trous de faible taille a moins d'importance que pour le score d'alignement.

Un algorithme spécifique d'estimation de réseau routier permettrait certainement de fournir des réseaux de meilleure qualité. Il est également envisageable de disposer de cartes ou systèmes d'information géographique permettant après recalage de tracer directement dans le repère de l'image un réseau routier précis.

Un autre point concerne la robustesse des deux critères énoncés. Ces derniers sont invariants par rotation, et le second est également invariant par échelle. Il est possible de normaliser la distance à la route si l'échelle locale est connue. Cette échelle dépend à la fois de l'échelle globale de l'image et des variations d'échelle au sein d'une même image dues à une incidence rasante ou des différences de profondeur. L'échelle globale peut être estimée dans l'absolu ou plus simplement par rapport à une référence. En revanche, une estimation précise des variations d'échelle à image fixée est plus délicate car elle nécessite des cartes de profondeur ainsi qu'une estimation de l'angle d'incidence.

Ces critères de contexte sémantique limitent la détection de véhicules en mouvement aux véhicules circulant sur un réseau routier repérable. Ainsi, des véhicules sur des chemins, ou des objets en mouvement autres, tels que des piétons, cyclistes ou animaux traversant des champs ou des espaces piétons par exemple, ne satisfont pas ces critères. Il est alors possible de définir de nouvelles classes telles que "piéton" ou "cycliste" et des critères associés. Ainsi, les classes "finales" peuvent être plus riches (introduction de sous-catégories) que les classes utilisées lors du procédé itératif.

D'autres règles de connaissance sont à explorer, notamment pour les cas d'application cités ci-dessus et qui ne sont pas pris en compte par les deux critères développés. Il peut s'agir de règles par pixel telles que celles présentées, ou de règles par régions. L'introduction de telles règles est susceptible de grandement restreindre la quantité des fausses alarmes ne satisfaisant pas ces dernières. Ainsi, de telles règles pourraient inclure :

- des modèles d'apparence adaptés à la classe à détecter, ici les véhicules ;
- des règles d'alignement entre une région et la vitesse de celle-ci : un véhicule se meut dans une direction proche de son axe ;
- des règles reliant la vitesse et la taille d'une région : un véhicule n'est pas censé dépasser une certaine vitesse et ses dimensions sont limitées inférieurement et supérieurement ;
- des règles imposant une cohérence du flot résiduel au sein de la région : un objet rigide doit présenter une vitesse uniforme. Cependant, une incidence importante (proche de l'horizontale) associée à une faible altitude du porteur aérien de la caméra, peuvent causer dans certains cas des variations de flot résiduel au sein d'un même objet (il s'agit alors d'un objet en trois dimensions) ;
- dans le cadre de bâtiments présentant des variations de profondeur et bien segmentés, le flot résiduel correspondant présente une structure affine qui pourrait être détectée ;
- à plus haut niveau, une mesure de la cohérence de mouvement entre objets (même alignement des régions et directions des flots résiduels associés proches) serait indicative d'ensembles de véhicules situés sur une même voie par exemple. De tels critères ne prendraient toutefois pas en compte l'existence de routes courbes ou un trafic trop faible pour faire appraitre de tels ensembles d'objets. Ce principe de "contexte de mouvement" (et d'apparence) [6] peut être notamment utilisé pour

l'analyse de circulation routière à partir de séquences vidéo aéroportées.

De manière plus générale, l'introduction de règles de connaissance *a priori* ciblant une classe d'intérêt particulière permet de préciser les résultats en éliminant un grand nombre de fausses détections. Ces éléments présentent toutefois un risque : une estimation imprécise des scores correspondants peut dégrader les performances de détection en filtrant des objets réels et en conservant des fausses alarmes.

5.6 Conclusion et perspectives

Deux ensembles d'informations apparaissent complémentaires, d'une part l'information locale d'apparence spatio-temporelle (couleurs, textures, mouvement), d'autre part l'information de contexte. Le "contexte" revêt ici plusieurs sens complémentaires : contexte spatial ou sémantique, local ou régional, primitives directement issues de la séquence vidéo ou résultats de classification.

Les seules informations d'apparence locale s'avèrent insuffisantes devant la variabilité intra-classe élevée en regard de la variabilité inter-classes. S'il est possible de construire des primitives ou des combinaisons de primitives permettant de séparer les différentes entités de la scène (ici véhicules mobiles, routes, bâtiments, fond par exemple) pour des conditions de prise de vue et des exemples d'entités donnés, le risque de surapprentissage est élevé. L'annulation de l'évolution des conditions de prise de vue (CPDV) par recalage introduit de nouvelles difficultés : disparités de profondeur, déformations d'apparence, recalage délicat pour des CPDV trop différentes.

L'apport du contexte, à plusieurs niveaux, permet de relâcher les exigences de précision d'une approche purement fondée sur des informations d'apparence. Le contexte permet de filtrer une partie des fausses détections voire de recouvrer des détections manquées. L'apport de règles de connaissance *a priori* adaptées au type d'objets à détecter, les véhicules en mouvement pour l'application présentée dans ce chapitre, dépend toutefois des performances de classification ou détection d'objets ou de structures sémantiques de contexte. Dans l'application développée, la qualité du réseau routier estimé est ainsi primordiale pour la pertinence des critères de distance et d'alignement comme filtres de fausses alarmes.

Cela illustre l'importance des algorithmes "bas niveau", fournissant les premiers ensembles de primitives pour les tâches de classification, détection, de reconnaissance ou encore d'identification (DRI). La qualité de l'estimation du flot optique, la définition de primitives pertinentes (manuelle ou automatique par apprentissage sur une base de données) selon les classes choisies, conditionnent directement les performances des approches de DRI. Des étapes préliminaires de lissage du flot optique ou des post-traitements sur les résultats de classification (respect de la cohérence temporelle des étiquettes) améliorent ainsi sensiblement la précision des détections au prix d'un temps de calcul accru.

Par rapport à l'approche développée dans ce chapitre, plusieurs perspectives méritent d'être mentionnées.

La simplicité des primitives et des données d'apprentissage utilisées est certes intéressante d'un point de vue calculatoire. En revanche, une approche plus complète reposant sur une base de données beaucoup plus importante et un ensemble de primitives plus riche pourrait résoudre en partie le problème de surapprentissage. Un grand nombre de primitives pourrait alors nécessiter une sélection automatique de primitives à partir de la base d'apprentissage. Qui plus est, la constitution d'une telle base de données, avec un but d'exhaustivité dans les conditions de prise de vue et apparence spatio-temporelle des classes, paraît toutefois délicate pour des raisons de disponibilité des données mais surtout en raison de l'explosion de la variabilité intra-classe rapportée à la variabilité inter-classe. La création de bases de données restreintes à un environnement (urbain, montagneux, ru-

ral, désertique...) et des conditions de prise de vue données afin de contourner le problème de variabilité nécessiterait de pouvoir associer à une nouvelle séquence vidéo la base de données pertinente, grâce à une interaction avec l'utilisateur humain voire de manière automatique.

L'inclusion de considérations géométriques telles que des contraintes épipolaires ou des critères de taille et forme sur des régions représente une autre forme de contexte à explorer. Ces derniers critères dépendent toutefois de la qualité des régions obtenues par des algorithmes de DRI préalables.

Sur un plan théorique, la définition d'un modèle incorporant des termes d'apparence et de contexte présenterait l'avantage d'une optimisation globale, sans schéma séquentiel dépendant particulièrement des performances des premières briques. Certains travaux ont proposé d'utiliser un formalisme à base de champs aléatoires conditionnels (CRFs). Ils définissent des potentiels unaires d'apparence et des potentiels de clique afin de prendre en compte le contexte, dans un cadre hiérarchique fondé sur une segmentation multi-échelles [130]. L'introduction de contraintes géométriques et d'interactions longue distance entre régions non voisines n'est toutefois pas traitée.

Détection d'activité : approche globale en temps

6

Problématique Les activités sont des événements cohérents dans le temps. Cela est d'autant plus marqué dans l'acception ici choisie du terme "activité", à savoir des véhicules en mouvement, dont les déplacements peuvent être considérés comme localement affines en temps et dont la cohérence temporelle est donc particulièrement claire (par opposition à des mouvements plus complexes de fluides par exemple). La détection d'activité est par conséquent une tâche pouvant bénéficier d'une approche globale en temps. Il semble donc naturel de suivre une telle approche globale pour obtenir des détections moins bruitées, voire des pistes correspondant aux trajectoires de chaque véhicule détecté. Outre une diminution du nombre de fausses alarmes, la disponibilité de telles pistes permettrait alors une étude du comportement de chaque entité, avec à la clef un filtrage ou une classification éventuels : suppression de détections improbables (par exemple dues à des effets de parallaxe), classification de la nature ou du comportement des entités (piétons, véhicules à deux ou quatre roues, allure modérée ou rapide...) respectivement.

Toutefois, les conditions de prise de vue inhérentes à la nature des données traitées, à savoir d'une part un mouvement complexe cumulant mouvement du capteur, effets de parallaxe et d'occultation, et mouvement des entités mobiles, d'autre part une variation de qualité image, rendent l'exploitation de la dimension temporelle ardue. Les opérations de recalage nécessaires pour associer les détections au cours du temps et produire des pistes cohérentes doivent être capables de maîtriser ou compenser au mieux ces difficultés. En supposant cette condition remplie, l'association d'entités voire de pistes élémentaires au cours du temps peut rester ambiguë, notamment en présence d'objets de taille réduite et d'apparences et de trajectoires semblables.

Une approche globale en temps nécessite dans un premier temps une compensation du mouvement apparent dominant. Les données image et de mouvement exprimées dans un référentiel commun sont en effet plus aisées à associer. Sans ce recalage, il reste possible d'associer par exemple des points d'intérêt par leur apparence, mais cela sera plus délicat en cas d'objets multiples d'apparences proches. Il sera également difficile de construire un modèle de comportement : un changement de focale par exemple modifie l'amplitude du mouvement image perçu, même si la vitesse réelle de l'objet n'a pas varié.

Contributions Dans ce chapitre, nous nous proposons d'étudier plusieurs approches parmi les méthodes globales exposées dans la section 4.2. Dans le contexte d'application évoqué, des séquences vidéo aériennes de complexités et d'environnements observés variés, une telle étude a pour but d'éclairer les points positifs et les sources de difficultés de chaque approche, voire de définir les domaines d'application souhaitables pour chacune (qu'il s'agisse de contraintes sur le mouvement dominant, la qualité du recalage, la résolution, l'importance des effets de parallaxe...)

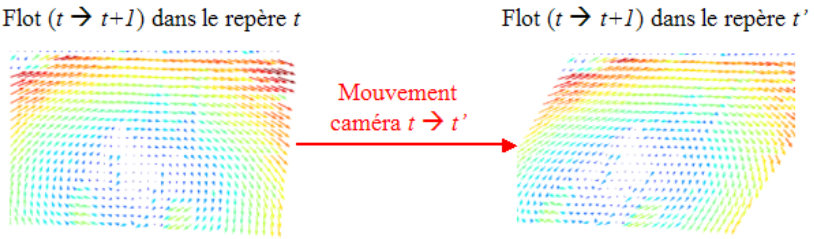


FIG. 6.1 – *Le flot optique correspondant au mouvement d'une image t peut être recalé de manière globale, par exemple par le biais d'un modèle paramétrique. En revanche les valeurs (amplitude et direction) du flot sont déformées.*

Une première distinction peut être effectuée entre des approches de "régularisation" d'une part, les approches globales "directes" d'autre part. La "régularisation" consiste à filtrer les détections obtenues indépendamment sur chaque image d'un segment temporel en vérifiant leur cohérence temporelle. Des détections isolées sont ainsi généralement des fausses détections dues à des erreurs de détection. Il s'agit d'une approche globale par "extension" car elle opère tout d'abord en deux dimensions avant d'intégrer la dimension temporelle.

Par contraste, les approches "directes" prennent en compte directement le volume spatio-temporel (ou des primitives extraites de ce volume). Les données considérées sont donc plus riches. Après recalage, il s'agit de détecter des tubes spatio-temporels d'activité (ce qui fournit par là-même des étiquettes afin de séparer les différentes entités en mouvement) ou des structures dans un espace de primitives adapté (par exemple, des variétés d'ordre 2 dans l'espace (x, y, v_x, v_y) en agrégeant les points d'un segment temporel et en supprimant la dimension du temps [279]).

6.1 Recalage

L'étape de recalage est nécessaire afin de compenser le mouvement global, voire le mouvement précis de chaque pixel et obtenir des données comparables dans un unique repère. Dans le cadre d'un recalage global, seul le mouvement d'ensemble est compensé. Subsiste le mouvement résiduel dû aux objets en mouvement ainsi qu'aux effets de parallaxe et aux erreurs d'estimation, estimation du mouvement complet par pixel ou du modèle global. Les données qui nous intéressent ici englobent le contenu image ou d'apparence et le contenu dynamique estimé à partir de la séquence d'images.

Sous l'hypothèse d'illumination constante, les données d'apparence ne nécessitent pas de traitement supplémentaire après recalage. Ce recalage s'applique également au flot tel qu'illustré sur la figure 6.1. En revanche, il importe d'annuler les potentiels changements d'échelle et d'orientation pour les données dynamiques (le flot optique, résiduel ou non, par exemple) afin d'obtenir des valeurs comparables.

La solution la plus rapide et qui a été choisie, consiste à ne prendre alors en compte que les changements apportés par le mouvement global, en appliquant l'inverse du modèle estimé (par exemple une affinité) au champ vectoriel de déplacement (flot) résiduel.

Flot optique Le flot optique, dont le principe et un état de l'art des méthodes d'estimation sont fournis en annexe A, fournit un champ de déplacement dense entre deux images. Ce champ permet après interpolation des intensités image d'obtenir une représentation image approchée de l'une des images dans le repère de la seconde, exceptées les régions occultées sans correspondances. Un critère d'erreur tel que la mesure du flot résiduel "en suivant le point" dans un trajet aller-retour, précisé à la section 5.4), donne une indication de fiabilité pour le champ de déplacement estimé.

Recalage global ou recalage complet par pixel ? Dans un premier temps, il peut sembler préférable d'effectuer un recalage complet, pour chaque pixel, sur une image de référence. Cela revient à compenser totalement le mouvement entre deux images. La position d'un pixel dans cette image serait alors associée au fil temporel d'un élément de l'espace réel suivi dans le temps. La suite des images obtenues après recalage permet ensuite de créer un modèle ou profil d'apparence en chaque pixel dans le repère de référence. Une fois le volume spatio-temporel obtenu, il est possible de paralléliser le traitement des profils de pixel (ou de patchs).

Cependant, si un recalage complet est envisageable pour de brefs instants, il est plus délicat de faire de même pour des intervalles temporels prolongés. Il faut en effet cumuler les recalages et interpolations car le recouvrement entre les images disparaît avec le mouvement de la caméra et du porteur. Un recalage global (ou moyen) sur l'ensemble de l'image, paramétrique par exemple, est plus robuste et tolère de nombreux outliers dans le calcul du flot complet. Il est cumulable avec une dérive raisonnable. Enfin, un recalage global permet de faire ressortir les objets mobiles ainsi que les structures en trois dimensions (effets de parallaxe) par différence image ou par étude du flot résiduel.

6.2 Régularisation temporelle

La régularisation temporelle peut intervenir à plusieurs étapes du processus de détection.

Il peut s'agir d'un post-traitement appliqué à un ensemble d'images de détection, binaires ou réelles, dans le but d'obtenir une image de détection finale de meilleure qualité ; ou encore des pistes d'entités, ce qui permet également de trier les résultats en ne conservant que les pistes cohérentes.

La régularisation peut être au contraire établie sur les données ou primitives même, avant classification ou détection. Dans le cas où la caméra est fixe, il peut ainsi être judicieux de réaliser un modèle de fond à partir d'une séquence vidéo. Une différence par rapport au fond permet ensuite d'extraire les entités mobiles. Il n'y a de plus dans ce cas particulier ni éléments de parallaxe, ni erreur de recalage. Dans le cas d'une caméra mobile, un recalage global préliminaire peut ramener au cas de la caméra fixe, tout en introduisant des défauts de recalage et des effets de parallaxe.

Nous présentons dans un premier temps une approche tirant parti d'une régularisation des données. Les cartes de détection obtenues ne seront pas fournies avec des pistes associées, mais les cartes peuvent être présentées dans un repère commun, ce qui facilite l'étude de cohérence temporelle. Dans un deuxième temps, nous étudions plutôt une régularisation *a posteriori* des résultats de détection par une approche de Graph Cut à étiquettes multiples en trois dimensions (deux dimensions spatiales auxquelles s'ajoute la dimension temporelle). La représentation des résultats et des extensions possibles seront évoquées dans un troisième temps.

6.2.1 Critères obtenus par régularisation temporelle et constitution de modèles

Le recalage dans un repère de référence d'un segment temporel (par exemple 50 ou 100 images consécutives) permet de disposer d'un volume spatio-temporel. Ce volume doit théoriquement présenter une redondance sur les régions de fond. En pratique, les erreurs de recalage et le bruit variable des images introduisent des variations sur les images recalées. Les éléments mobiles ainsi que les structures en trois dimensions sont une autre source de variations. La première approche présentée et testée ici consiste à extraire plusieurs critères de présence d'objets mobiles par régularisation temporelle. La figure 6.2 illustre l'extraction et la composition de ces critères afin d'obtenir une carte de score de détection finale.

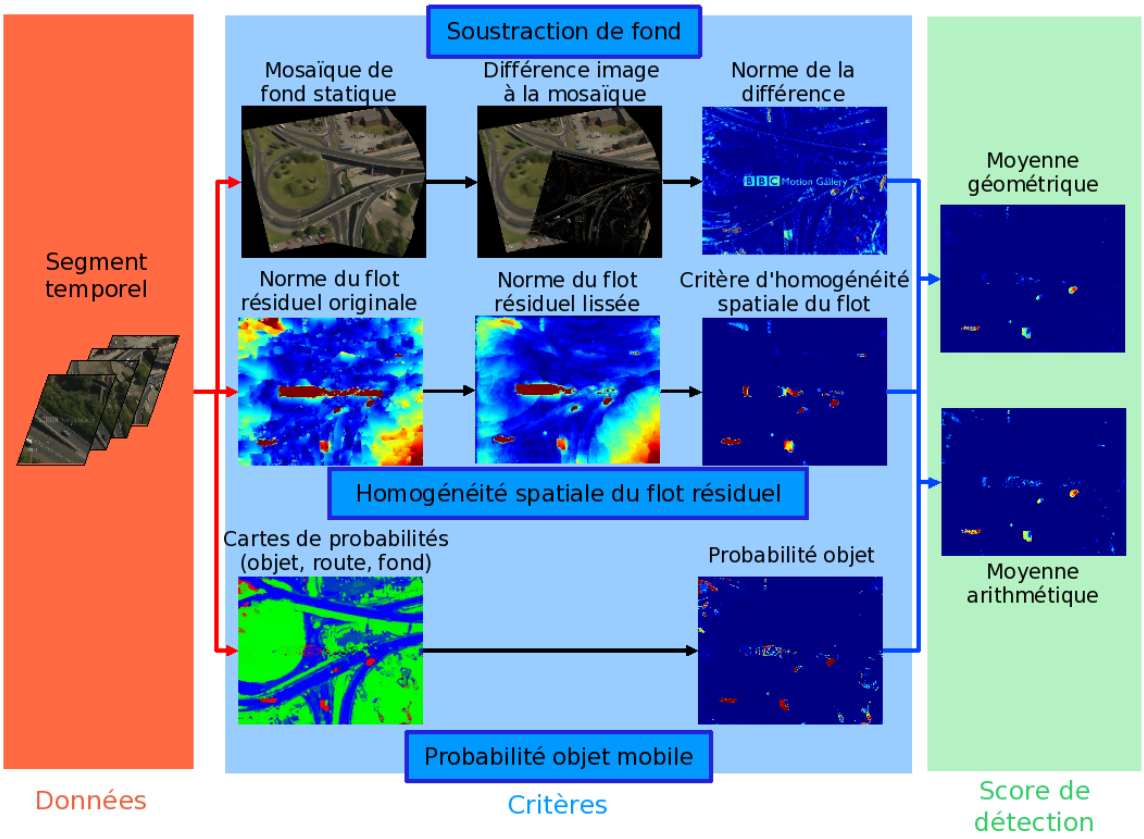


FIG. 6.2 – *Le segment temporel (données) permet de créer une mosaïque de fond (première ligne) après recalage global, d'où l'on tire des images de différence. En chaque image, toujours après recalage, la norme du flot résiduel est d'abord lissée puis fournit un critère d'homogénéité spatiale du flot (deuxième ligne). Des cartes de probabilités par exemple issues d'un algorithme de classification locale apportent un troisième critère de probabilité objet (troisième ligne). Les trois critères ainsi obtenus sont ensuite fusionnés, ici par moyenne géométrique ou arithmétique afin de fournir une carte de scores de détection.*

Le premier critère est une méthode de type "suppression de fond" avec caméra mobile. Les données sont traitées par bloc afin de produire une mosaïque de fond statique M_{fond}. Le critère correspondant $c_{fond}(t)$ pour une image t est donné par :

$$c_{fond}(t) = \alpha \| M^t_{fond} - I^t_{aff} \|^2_2$$

où

- I^t_{aff} est l'image t (couleur) après recalage affine dans le repère de la mosaïque,
- M^t_{fond} représente la partie de la mosaïque M_{fond} (couleur) restreinte à l'emprise de I^t_{aff},
- α est un critère de normalisation afin d'obtenir des valeurs du critère entre 0 et 1 (ici $\frac{1}{\sqrt{3}}$ avec des valeurs des canaux R,G et B de I^t_{aff} et M^t_{fond} prises entre 0 et 1).

L'idée sous-jacente au critère tiré des flots résiduels vient de trois ensembles de comportements possibles, fond, structures en trois dimensions ou objets mobiles. Le flot résiduel correspondant aux régions de fond (hors structures en trois dimensions) devrait être nul et n'indiquer ainsi aucun objet en mouvement. Le flot résiduel associé aux structures en trois dimensions, généralement des bâtiments, présente localement une structure affine due aux variations de profondeur par rapport à la caméra. Une application de filtre médian spatial sur les images de flots résiduels (composantes ou norme) devrait donc créer des flots sensiblement identiques en apparence. En revanche, les objets mobiles évoluent au milieu du fond (*a priori* sans structures en trois dimensions). L'application de ce même filtre sur les régions correspondantes devrait donc renvoyer un flot résiduel nul ou faible (médian spatial du flot environnant correspondant au fond), très différent du flot résiduel de l'objet mobile.

Ce critère est obtenu en plusieurs étapes illustrées figure 6.5 :

- la première étape consiste à régulariser la norme du flot résiduel $n_r(t)$ (image (a)) par une moyenne pondérée des images de normes après recalage complet (image (b)) :

$$n^{reg}_r(t) = \frac{1}{\sum_{t'=t-2}^{t+2} (1 - c_{AR}(t'))} \sum_{t'=t-2}^{t+2} \left(1 - c_{AR}(t')\right) n^{rec}_r(t')$$

  où $n^{rec}_r(t')$ correspond à l'image de norme de flot résiduel après recalage en chaque pixel dans le repère au temps t et $c_{AR}(t')$ le critère aller-retour de fiabilité du flot optique utilisé pour le recalage et calculé ici par l'algorithme Huber-L1 ;

- $n_{res}(t)$ est l'image obtenue après recalage affine (dans le repère de la mosaïque) de la norme du flot résiduel $n^{reg}_r(t)$;

- une version spatialement lissée $n^{median}_{res}(t)$ de la norme $n_{res}(t)$ est obtenue par application d'un filtre médian spatial sur des fenêtres de 61×61 pixels ;

- le critère $c_{flot}(t)$ pour une image t (image (c)) est obtenu selon la formule suivante :

$$c_{flot}(t) = \min\left(1, \beta \max\left(0, n_{res}(t) - n^{median}_{res}(t)\right)\right)$$

  où β est un critère de normalisation dépendant de la norme du flot résiduel après recalage (dans l'exemple de la séquence "BBC", avec une norme significative de flot résiduel fixée à 1.5 pixels, $\beta = \frac{1}{1.5}$).

Un dernier critère peut être extrait de cartes de probabilités obtenues comme sorties d'algorithmes de classification locale en temps. Ce critère $c_{prior}(t)$ pour une image t correspond à la carte de probabilité objet après recalage dans le repère de la mosaïque.

L'application de différents critères utilisant ces différentes mosaïques, cartes ou champs lissés conduit ainsi à des cartes de critères. Différentes opérations de fusion sont ensuite appliquées à ces cartes (somme dans le temps, moyenne, médian...) afin de produire une carte de score finale (en l'occurrence une carte de présence d'objets mobiles). En résumé, les différents critères sont formés par :

- la différence image (couleur) par rapport à la mosaïque de fond ;
- des flots résiduels lissés dans le temps. La différence de ces flots par rapport à des médians spatiaux de ces mêmes flots fournit des critères de présence d'objet mobile ;
- des cartes de probabilités, ou des cartes binaires, issues par exemple de traitements de classification tels que ceux présentés au chapitre 5 après une éventuelle régularisation temporelle.

Établissement d'une mosaïque de fond Plusieurs points doivent être pris en compte pour la construction d'une mosaïque dans le cas d'une caméra mobile. Il faut tout d'abord recaler les images d'un segment temporel dans un repère unique, généralement celui de l'image centrale du segment afin de minimiser les déformations dues au recalage. Les erreurs de recalage ainsi que les effets de parallaxe nuisent à la qualité globale de la mosaïque si celle-ci est construite par simple moyenne.

Il est alors intéressant de construire plutôt le médian des différentes images obtenues après recalage global (ici un recalage affine). La mosaïque obtenue ici doit filtrer les objets mobiles en créant un fond statique. En effet, dans le cas d'objets mobiles, les pixels de l'objet n'apparaissent, à une position image donnée dans un repère commun, que pendant un nombre limité d'images, d'autant plus réduit que la vitesse de l'objet est élevée. Le filtre médian devrait ainsi renvoyer des valeurs correspondant au fond et non à l'objet, sans "fantôme" apparaissant par une simple moyenne. Pour les effets de parallaxe, l'effet d'un tel filtre dépend de la structure et de l'apparence du bâtiment : les bords les plus décalés par rapport à leur position dans l'image centrale, correspondant aux extrémités temporelles du segment, n'apparaissent que pendant un faible nombre d'images par rapport aux pixels du fond, présents dans les images "centrales". En revanche, si le reste des éléments en trois dimensions devrait être moins perturbé que par une moyenne des images de la séquence (toujours après recalage global), l'apparence présentera tout de même des modifications par rapport à l'image centrale.

Le figure 6.3 montre un exemple de telles mosaïques. Les véhicules circulant dans la partie inférieure de l'image centrale sont "gommés" dans la mosaïque. En revanche, la voiture blanche engagée sous le "pont", la voiture rouge vers le centre de la mosaïque sont apparentes. Cela est dû à leur vitesse nulle ou très faible. Les deux voitures blanches en haut de l'image centrale, de vitesse faible, ne sont pas totalement effacées, et apparaissent sous la forme de "fantômes".

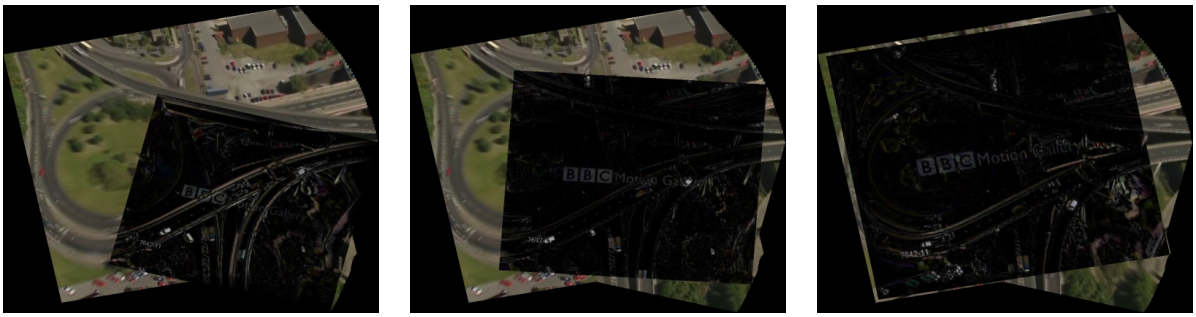
La figure 6.4 montre plus précisément les différences entre la mosaïque et les images d'origine du segment après leur recalage dans le repère de référence. Se détachent particulièrement les véhicules (d'autant plus que leur apparence contraste avec celle du fond), les structures en trois dimensions (les bords des routes à altitude variable) et la watermark du documentaire.

Le deuxième ensemble de primitives est tiré de la norme du flot résiduel. La figure 6.5 décrit les différentes étapes de l'obtention de ce critère, précisées page 121.

Ces deux critères (différence par rapport à une mosaïque de fond et critère de régularité spatiale du flot résiduel) peuvent être combinés en l'état ou avec un résultat de classification provenant par exemple de la méthode décrite au chapitre 5. La figure 6.6 donne ainsi les moyennes arithmétique et géométrique des trois critères ainsi disponibles (images (b)

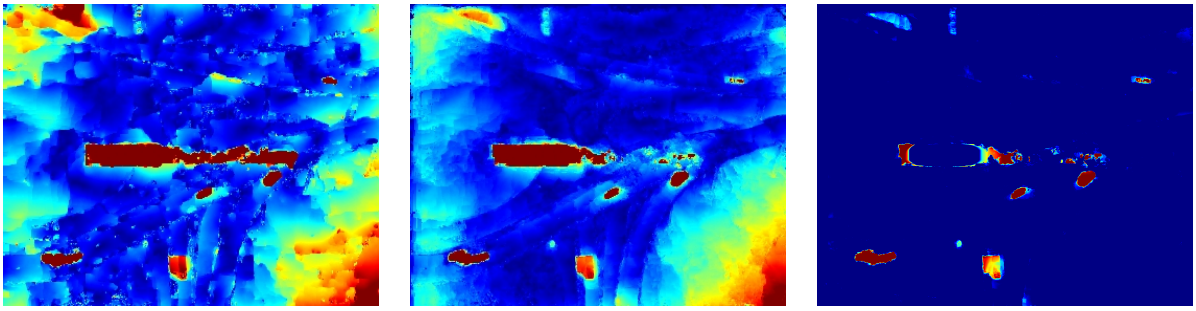
(a) (b)

FIG. 6.3 – *(a) Mosaïque de fond obtenue à partir d'un segment de* 51 *images de la séquence "BBC". (b) Image centrale du segment*



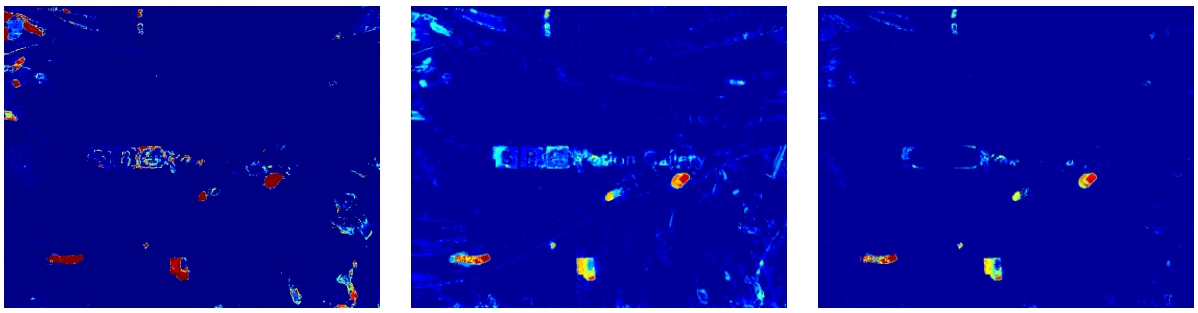
(a) (b) (c)

FIG. 6.4 – *Différences image entre la mosaïque et quelques images du segment après recalage dans le repère de la mosaïque (ou de l'image centrale du segment).*



(a) (b) (c)

FIG. 6.5 – *Critère de présence d'objet mobile à partir de la norme du flot résiduel. (a) Norme originale du flot résiduel ; (b) norme lissée ; (c) critère.*



(a) (b) (c)

FIG. 6.6 – *(a) Probabilité objet (ici obtenue par le processus de classification itératif décrit au chapitre 5) ; (b) moyenne arithmétique des trois critères ; (c) moyenne géométrique des trois critères*

et (c)).

Le seuillage des cartes obtenues respectivement à partir de la carte de probabilité objet et de la moyenne géométrique des trois critères fournit des détections d'objets (qui peuvent être par la suite traitées par régularisation temporelle et opérations morphomathématiques). La figure 6.7 montre une illustration des détections obtenues, superposées à l'image d'origine. Les seuils ont été choisis de manière à obtenir un rappel équivalent.

Il apparaît sur cet exemple que l'introduction de critères supplémentaires obtenus grâce à un lissage temporel (à partir de la mosaïque de fond et de la norme lissée du flot résiduel) diminue de manière importante les fausses alarmes. En revanche, dans les deux cas, les objets de faible contraste ne sont pas détectés. Cela est dû à une estimation faussée du flot optique (et donc du flot résiduel) : sur la figure 6.5, le véhicule sombre situé dans le coin inférieur gauche de l'image (cf. images (a) et (b), première ligne de la figure 6.7) ne correspond à aucune zone de flot résiduel notable. Le critère correspondant et la carte de probabilité obtenue (la classification dépend fortement du flot résiduel) portent donc des valeurs faibles. Le véhicule situé dans le coin supérieur droit apparaît sur le critère de flot ainsi que le critère de mosaïque mais pas sur la carte de probabilité (d'où son absence dans la moyenne géométrique). La figure 6.8 détaille respectivement pour chacun de ces objets l'image originale correspondante, la norme du flot résiduel après lissage et la moyenne arithmétique des trois critères. La moyenne arithmétique permet dans le deuxième cas de "récupérer" la détection au contraire de la moyenne géométrique, deux des critères fournissant une réponse non nulle.

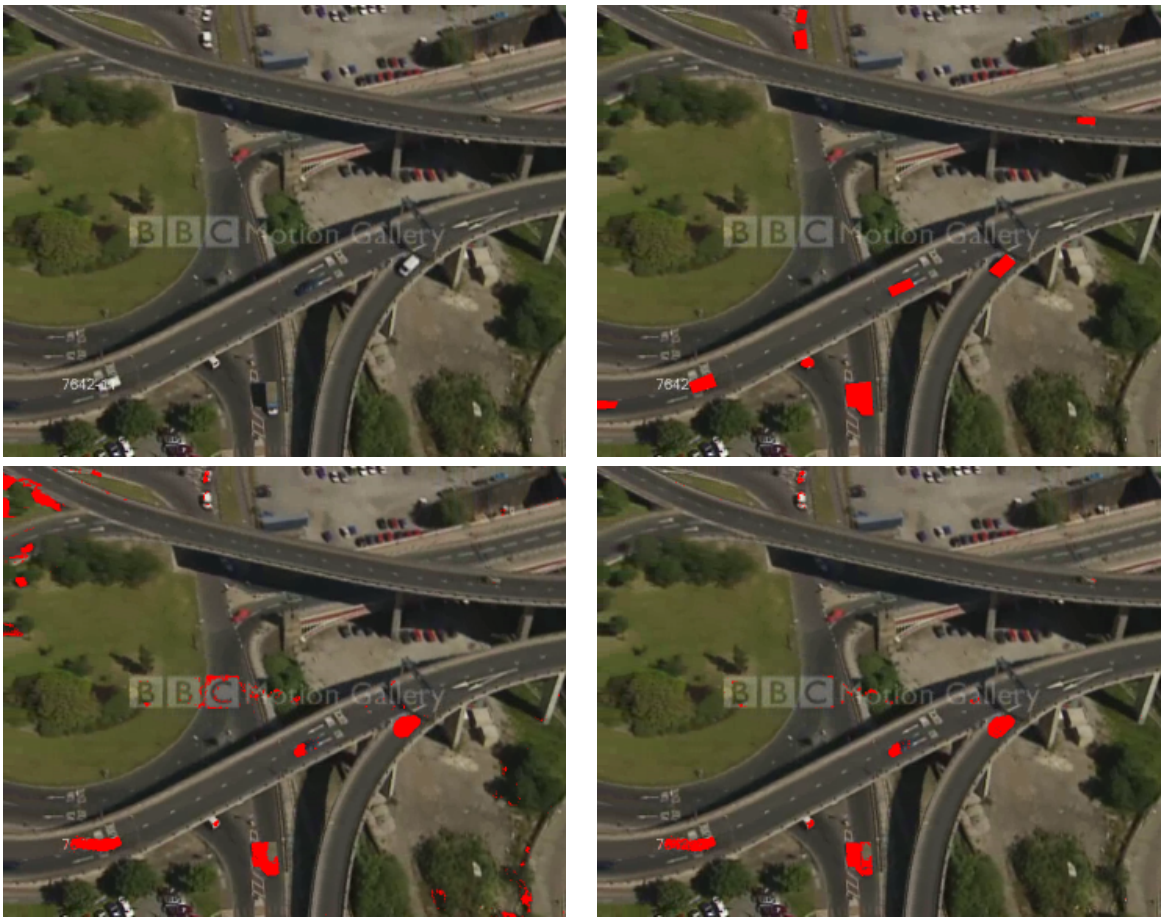
La moyenne arithmétique correspond ainsi à un choix intermédiaire, avec un rappel plus important mais une précision décrue. Il s'agit d'un compromis entre l'utilisation de la seule carte de probabilité objet, et la moyenne géométrique, avec précision supérieure mais moindre rappel.

6.2.2 Filtrage par connexité temporelle

L'aspect temporel de la régularisation utilisée dans la section 6.2.1 ne concerne que la construction des primitives ou des critères utilisés pour la détection. Il n'y a ainsi aucune cohérence temporelle entre les détections ou les cartes de critères finales obtenues pour chaque image.

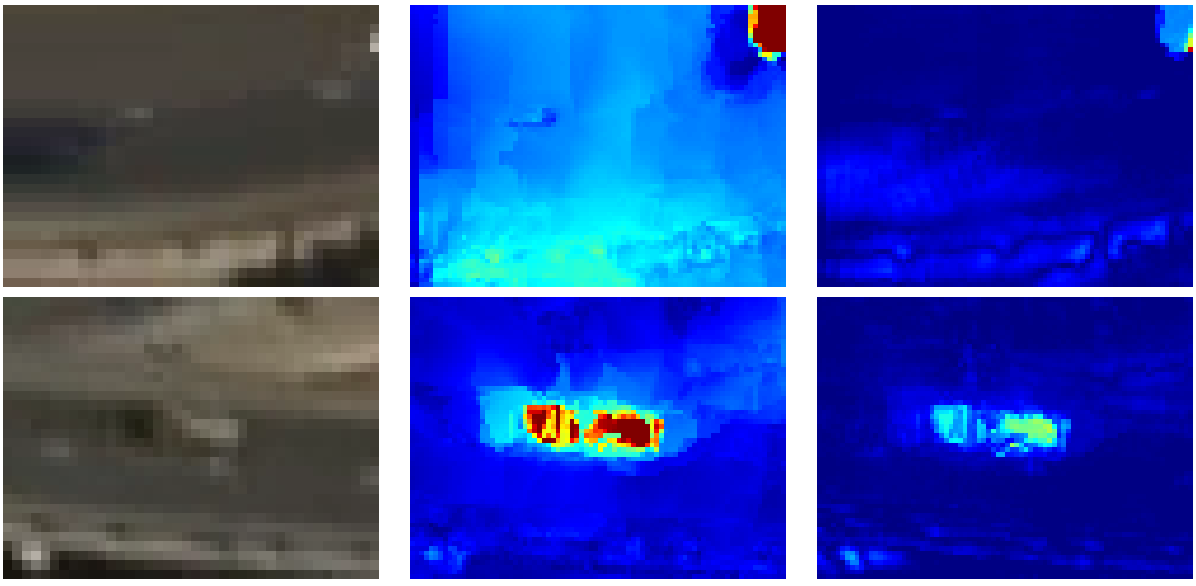
Il s'agit donc de tester l'influence d'une contrainte de connexité temporelle entre ces résultats isolés dans le temps. Un recalage global est nécessaire afin de supprimer les effets du mouvement du capteur. Ce recalage entraîne toutefois des effets de parallaxe qui devront être filtrés, par exemple par le biais du critère spatial sur la norme du flot résiduel, critère introduit dans la section précédente 6.2.1.

Dans le repère affine de référence, nous disposons de différentes images d'étiquettes E_t, une par image t du segment temporel considéré. E_t vaut 0 pour les pixels du fond et



(a) (b)

FIG. 6.7 – 1*ère ligne : (a) image et (b) vérité terrain objet.* 2*ème ligne : objets détectés par seuillage de (a) la probabilité objet (ici obtenue par le processus de classification itératif décrit au chapitre 5) ; (b) la moyenne géométrique des trois critères*



(a) (b) (c)

FIG. 6.8 – *Détail des objets non détectés. (a) Détail de l'image ; (b) image de la norme du flot résiduel associé ; (c) moyenne arithmétique des trois critères.*

prend des valeurs strictement positives pour chaque objet détecté. Ces images d'étiquettes ont été obtenues en trois étapes à partir d'images de critères :

- opérations morphomathématiques afin de ne conserver que les composantes connexes possédant suffisamment de pixels dont la norme du flot résiduel est significative (pour l'échelle considérée, nous avons fixé ce seuil à 1.5 pixels) afin de supprimer les composantes "bruitées" ;
- remplissage de chaque composante connexe par les voisins d'apparence semblable et dont la norme du flot résiduel est supérieure à un seuil de "bruit" (0.3 pixel dans l'exemple choisi). Pour cela, quatre modes couleurs au plus sont extraits de la composante connexe, représentatifs de la composante (sous l'hypothèse que la composante peut être représentée par ces modes). Pour chaque mode, les pixels satisfaisant aux conditions d'apparence et de norme sont rajoutés à la composante. Une condition supplémentaire sur la distance maximale des nouveaux pixels à la composante connexe originale, dépendante de l'échelle de l'image, permet d'éviter un remplissage excessif ;
- des opérations morphomathématiques de fermeture et de suppression des composantes d'aire trop petite (inférieure dans l'exemple choisi à 5 pixels) sont appliquées aux composantes connexes résultant de l'étape précédente afin d'obtenir des composantes plus régulières et de taille suffisante.

Ces images d'étiquettes sont ensuite accumulées dans le repère affine de référence et le score d'accumulation est ainsi donné par :

$$c_{somme} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}(E_t^{aff} > 0)$$

où

- t représente l'indice dans le segment temporel considéré (ici de 15 images) avec N la longueur du segment,
- E_t^{aff} correspond à l'image d'étiquettes après recalage dans un repère commun (celui de l'image centrale du segment temporel).

L'idée est de filtrer les composantes connexes de bruit ne présentant aucune cohérence temporelle, au contraire des véhicules en mouvement. Le déplacement des véhicules résiduel, dans le repère de référence après recalage affine, conduit à des composantes connexes décalées, pour un même véhicule de l'espace réel. Toutefois, la fréquence d'échantillonnage élevée (environ 25 images par seconde, soit un déplacement maximal de 2 mètres par image en supposant une vitesse très élevée de 180 km/h) permet de supposer un recouvrement élevé entre les différentes positions du véhicule dans le repère commun (la longueur d'un véhicule dépasse les 3-4 mètres). Il suffit ensuite de conserver les régions étiquetées sur l'image centrale dont le critère c_{somme} répond à certaines contraintes. Nous avons choisi de conserver celles pour lesquelles la moitié des pixels de la région ont un critère supérieur à 0.5, ce qui permet de tolérer partiellement des erreurs de recalage ainsi que prendre en compte le déplacement de l'objet. En revanche, les régions pour lesquelles la majorité des pixels est incohérente dans le temps sont considérées comme étant du bruit et sont filtrées.

La figure 6.9 présente en première ligne, dans le repère commun, l'image centrale du segment en (a), les étiquettes avant régularisation en (b). Ces étiquettes ne sont obtenues ici qu'à partir des deux critères de différence par rapport à la mosaïque et le critère provenant de la norme du flot résiduel. Le rappel est donc plus élevé au détriment de la précision : l'influence de la contrainte de connexité temporelle sera ainsi plus visible et cela permet d'envisager une approche ne nécessitant pas de classification locale en temps préalable, coûteuse en temps par rapport aux deux autres critères. Le masque de détection est donc

différent de celui obtenu en figure 6.7 (deuxième ligne (b)) qui intégrait la carte de probabilité objet résultat d'une classification locale en temps. Cela se traduit d'une part par de nombreuses étiquettes correspondant aux watermarks, d'autre part par une détection plus complète de certains véhicules (en bas de l'image et dans la partie supérieure, les deux véhicules blancs proches ou encore la voiture grise sur la route horizontale dans la partie supérieure droite). La régularisation temporelle permet de filtrer des étiquettes correspondant à du bruit ainsi qu'une partie des watermarks. Les seules fausses détections après régularisation proviennent des watermarks.

Cette approche sans apprentissage présente plusieurs avantages par rapport à une approche utilisant de l'apprentissage telle que celle développée au chapitre 5. En effet, les résultats ne dépendent pas ici des données d'apprentissage et de leur similarité avec les données test. Il n'y a également pas de processus itératif susceptible de régulariser par trop les cartes de probabilités obtenues et par conséquent de supprimer d'éventuels petits objets. En revanche, l'absence de prise en compte du contexte, qu'il s'agisse de contexte local ou sémantique, ne permettra pas de filtrer les incrustations ni des effets de parallaxe de faible étendue spatiale. Enfin, le filtrage par connexité temporelle décrit dans ce paragraphe peut également être réalisé sur les résultats de l'approche avec apprentissage, ce qui améliorerait les performances de détection associées.

Dans tous les cas, il est possible de contourner le problème des incrustations fixes telles que les watermarks en appliquant auparavant un algorithme de détection de watermarks et en filtrant systématiquement le masque correspondant dans le repère d'origine de la séquence vidéo (les véhicules potentiellement situés au même endroit ne seront en revanche plus détectés).

6.2.3 Graph Cut 3D

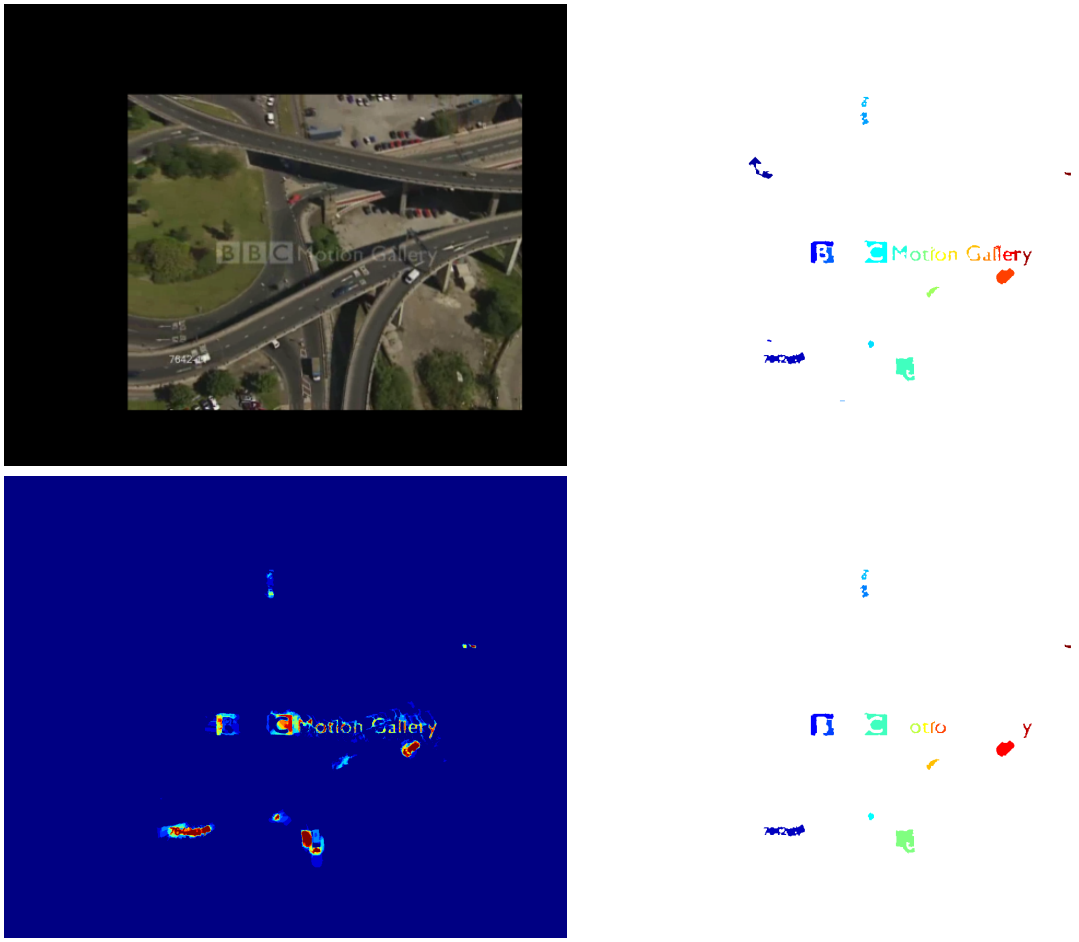
L'approche heuristique présentée au paragraphe précédent ne prend pas en compte l'association des composantes connexes ou étiquettes obtenues à chaque pas de temps. La régularisation temporelle permet de filtrer une partie des fausses alarmes, notamment le bruit temporellement non cohérent. Elle n'apporte en revanche pas d'information sur les trajectoires des différents objets détectés. Une étape supplémentaire de pistage est donc nécessaire pour établir ces trajectoires.

De nombreuses approches traitent ce problème de pistage et sont évoquées à la section 4.2. Nous développons ici une approche de type champ de Markov aléatoire (MRF) en trois dimensions, deux dimensions spatiales image auxquelles s'ajoute le temps, et multi-étiquettes afin de conserver les identités des pistes au cours du temps. L'énergie globale de la segmentation se décompose en :

- des potentiels unaires représentant l'attache aux données ;
- des potentiels de clique représentant les interactions entre pixels voisins spatio-temporels et indiquant la propension de ces derniers à appartenir à la même classe, il s'agit du terme d'*a priori*.

Dans le cas binaire, avec une classe d'objets à détecter par opposition à une classe de fond, les potentiels unaires traduisent simplement une probabilité *a priori* d'appartenance à la classe (de fond ou d'objet). Dans le cas multi-étiquettes, il faut pouvoir imposer un *a priori* d'appartenance à une classe précise, en l'occurrence un véhicule précis afin de conserver leur identité au cours du segment temporel étudié. Une combinaison de deux termes est donc nécessaire afin d'intégrer simultanément l'*a priori* d'appartenance au fond ou à un objet d'une part, et l'appartenance à un véhicule précis d'autre part :

$$V_{unaire}(e,t,\mathbf{x}) = \left((1 - P^t_{obj}(\mathbf{x}))\mathbf{1}_{e\neq 0} + (P^t_{obj}(\mathbf{x}))\mathbf{1}_{e=0}\right) + \mathbf{1}_{e\neq 0}\left(\frac{D_e(\mathbf{x})^2}{100}\right)$$



(a) (b)

FIG. 6.9 – *1ère ligne : (a) image centrale du segment temporel dans le repère de référence (les limites de l'image correspondent à celles de la mosaïque de la figure 6.3 (a) ; (b) étiquettes avant régularisation temporelle. Chaque couleur représente une étiquette différente. 2ème ligne : (a) Somme des masques binaires recalés dans le repère commun (les couleurs chaudes sont associées à des valeurs plus élevées) et (b) étiquettes après régularisation temporelle. Les couleurs correspondent aux étiquettes avant régularisation temporelle (1ère ligne (b)).*

en notant :

- V_{unaire} le potentiel unaire (composante de l'énergie à minimiser) ;
- P^t_{obj} l'image de probabilité d'appartenance à un objet pour l'image t (obtenue par classification, moyenne arithmétique ou géométrique de plusieurs critères etc.) après recalage global ;
- e l'étiquette (0 pour le fond, entre 1 et le nombre d'objets obtenu à l'initialisation, détaillée ci-après, pour les objets) ;
- t le numéro de l'image dans le segment temporel de 15 images considéré ;
- $\mathbf{x}$ la position du pixel considéré dans l'image ;
- $\mathbf{1}_{e\neq 0}$ respectivement $\mathbf{1}_{e=0}$ les fonctions indicatrices d'une étiquette "objet" ou de l'étiquette "fond" ;
- D_e la carte de distance euclidienne à la région (0 au sein de la région) portant l'étiquette e lors de l'initialisation ci-après.

Pour cela, une initialisation est nécessaire. Les composantes connexes étiquetées obtenues après régularisation temporelle telle que présentée à la section 6.2.2, déjà partiellement débruitées, peuvent fournir une telle initialisation. Le terme traduisant l'appartenance à un objet peut être ici binaire (égal à 1 pour les pixels possédant une étiquette objet, 0 sinon) ou réel, entre 0 et 1 selon un score de vraisemblance d'appartenance à un objet (tiré d'une carte de probabilité objet par exemple). Le terme indiquant l'appartenance à un objet spécifique (associé à une étiquette) est dans notre approche traduit par un score de distance à la composante connexe considérée (multiplié par un facteur fixe afin que les deux termes du potentiel unaire soient de même ordre de grandeur).

Pour des raisons de simplicité, les cliques considérées ici sont d'ordre 2. Chaque pixel possède donc, en trois dimensions, 6 voisins : 4 voisins spatiaux et 2 voisins temporels (il est envisageable de considérer les 27-voisinages mais là encore, les 6-voisinages ont été choisis pour des raisons calculatoires). Les potentiels binaires pour une clique d'ordre 2 donnée traduisent d'une part la similarité d'apparence entre les pixels voisins considérés, par le biais de la mesure des différences couleur ; d'autre part la similarité des *a priori* d'appartenance à un objet entre ces mêmes pixels, par la mesure de la différence des probabilités de classe objet :

$$V_{binaire}(e,t,\mathbf{x},e',t',\mathbf{x}') = V^{e=e'}_{binaire} = \left(\|I(\mathbf{x},t) - I(\mathbf{x}',t')\|_2^2\right) + \alpha\left(P^t_{obj}(\mathbf{x}) - P^{t'}_{obj}(\mathbf{x}')\right)^2$$

si $e = e'$ et le "complémentaire" $V_{max} - V^{e=e'}_{binaire}$ si $e \neq e'$, avec $V_{max} = 3 * 255^2 + \alpha$ et :

- I le volume image RGB spatio-temporel correspondant au segment temporel après recalage global ($I(\mathbf{x},t)$ correspond donc au vecteur des coordonnées RGB du pixel situé en $\mathbf{x}$ sur l'image t ;
- $V_{binaire}$ le potentiel unaire (composante de l'énergie à minimiser) pour le couple de points voisins spatio-temporels $\mathbf{x}, t$ et $\mathbf{x}', t'$;
- t (respectivement t') le numéro de l'image dans le segment temporel de 15 images considéré ;
- $\mathbf{x}$ (respectivement $\mathbf{x}'$) la position du pixel considéré dans l'image ;
- $\|.\|_2^2$ la norme L_2 au carré, ici du vecteur couleur RGB ;
- α le facteur de normalisation entre les termes de similarité d'apparence et de différence des probabilités de classe objet ;
- $P^t_{obj}(\mathbf{x})$ (respectivement $P^{t'}_{obj}(\mathbf{x}')$) la probabilité objet pour l'image t (resp. t') au pixel $\mathbf{x}$ (resp. $\mathbf{x}'$).

L'énergie totale d'une configuration donnée (étiquettes) est une somme pondérée des

termes d'attache aux données d'une part, et d'*a priori* d'autre part :

$$E_{totale}(e_i, t_j, \mathbf{x}_k) = \sum_{i,j,k} V_{unaire}(e_i, t_j, \mathbf{x}_k) + \gamma \sum_{i,i',j,j',k,k'} V_{binaire}(e_i, t_j, \mathbf{x}_k, e_{i'}, t_{j'}, \mathbf{x}_{k'})$$

Le paramètre γ contrôle l'influence relative des deux termes. Une valeur de l'ordre de 10^5 correspond ainsi à des termes comparables.

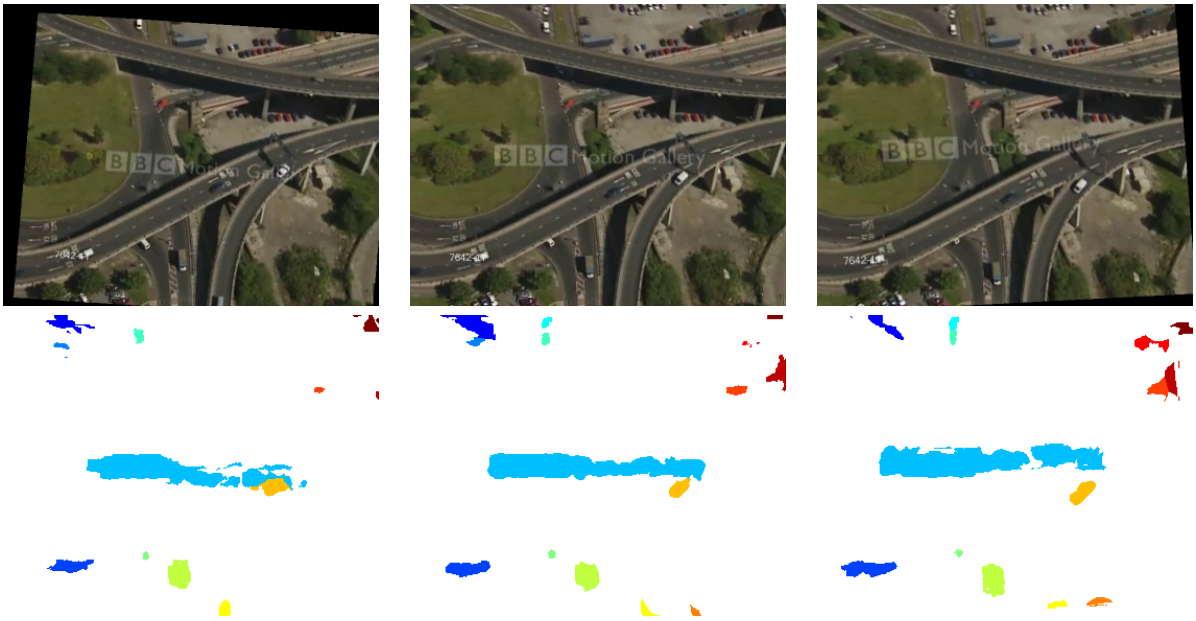
Le choix des potentiels binaires de clique semble logique. En revanche, le choix des potentiels unaires décrits ci-dessus pose plusieurs problèmes. Si le critère de distance évite une trop grande dérive, il ne prend pas en compte la différence temporelle et la dérive de position croissante. Il est possible de corriger cette limitation en modifiant le terme de distance selon la différence temporelle par rapport à l'image centrale du segment. Le seuil de distance maximale à la composante connexe considérée dans l'image centrale serait ainsi linéairement dépendant de la différence temporelle entre l'image centrale et l'image considérée.

Cela amène au second problème du potentiel binaire, à savoir le pouvoir de séparation spatiale. Le cas particulier de deux véhicules d'apparences semblables et situés côte-à-côte illustre cette difficulté. Les potentiels unaires correspondants seront en effet sensiblement égaux pour chacune des classes pour les pixels correspondant aux deux véhicules dans les autres images de la séquence. Les apparences des deux véhicules étant semblables, les potentiels binaires de cliques réunissant des pixels appartenant chacun à un véhicule différent seront peu différents des potentiels de cliques réunissant des pixels appartenant à un même véhicule.

Une solution éventuelle consiste à n'imposer les potentiels unaires que pour l'image centrale comme *a priori* et initialisations des positions des objets, et à modifier les potentiels de clique. Ces derniers devraient alors toujours favoriser l'appartenance à une même classe lorsque les apparences (couleurs, mais possiblement dynamique par l'intermédiaire du flot résiduel) et probabilités de classe objet sont voisins mais handicaper des différences significatives. Le cas de véhicules d'apparence image composite serait alors traité par le biais d'une pondération entre distance image, distance dans l'espace couleur (voire du flot résiduel) et distance entre les probabilités de classe objet. Les poids associés aux potentiels binaires de cliques dans l'image centrale seraient affaiblis afin de faire ressortir les *a priori* de classe. Une autre méthode encore consisterait à intégrer le flot résiduel comme prédiction de la position dans les autres images des objets (et donc comme potentiels unaires), en se rapprochant en cela du principe de pistage par filtre de Kalman par exemple. Les imprécisions et les divergences possibles du cumul de flots résiduels, ainsi que le surcoût calculatoire, sont toutefois des obstacles à cette modification. Un pistage classique initialisé par les composantes connexes de l'image centrale serait alors plus rapide.

La figure 6.10 montre un exemple d'étiquettes spatio-temporelles obtenues sur un ensemble de 15 images consécutives (les régions d'aire inférieure à un seuil, ici 5 pixels, ont été supprimées). Les potentiels unaires ici utilisés proviennent ici non pas de la probabilité objet comme dans la formule ci-dessus, mais directement de la norme du flot optique résiduel (normalisée afin que la probabilité dans la formule soit remplacée par un paramètre de valeurs également comprises entre 0 et 1). Les watermarks (au centre et dans le coin inférieur gauche de l'image), n'ayant pas été filtrées, sont ainsi "détectées".

Il est également intéressant de noter le côté relativement imprécis des régions obtenues. En effet, le compromis entre le terme d'attache aux données traduit par les potentiels unaires d'une part, le terme de régularité traduit par les potentiels binaires de clique d'autre part, ne peut produire directement des régions exactes pour chaque image. Ce problème apparaît surtout aux abords des frontières peu contrastées qui favorisent des configurations locales avec étiquettes identiques par l'intermédiaire des potentiels binaires. Les imprécisions ou erreurs éventuelles de recalage perturbent également la minimisation de l'énergie globale. Une étape de post-traitement fondée sur l'apparence image, plus précise



(a) (b) (c)

FIG. 6.10 – *1ère ligne : (a) 1ère image, (b) image centrale et (c) dernière image d'un segment temporel de 15 images après recalage global affine dans le repère de référence (ici celui de l'image centrale), coupées aux limites de l'image centrale. 2ème ligne : étiquettes d'objets correspondant à ces trois images. Une même couleur représente au cours du temps un même objet.*

(par exemple une approche par contours actifs), doit ainsi être envisagée. Enfin, le réglage du paramètre γ contrôlant l'influence relative des termes unaires et binaires peut se traduire par une régularisation trop importante ou au contraire insuffisante selon la structure locale du volume spatio-temporel.

De plus, les flots optiques estimés sont bruités et plusieurs régions étiquetées ne correspondent à aucun véhicule. Il est donc nécessaire de filtrer les régions dues au bruit.

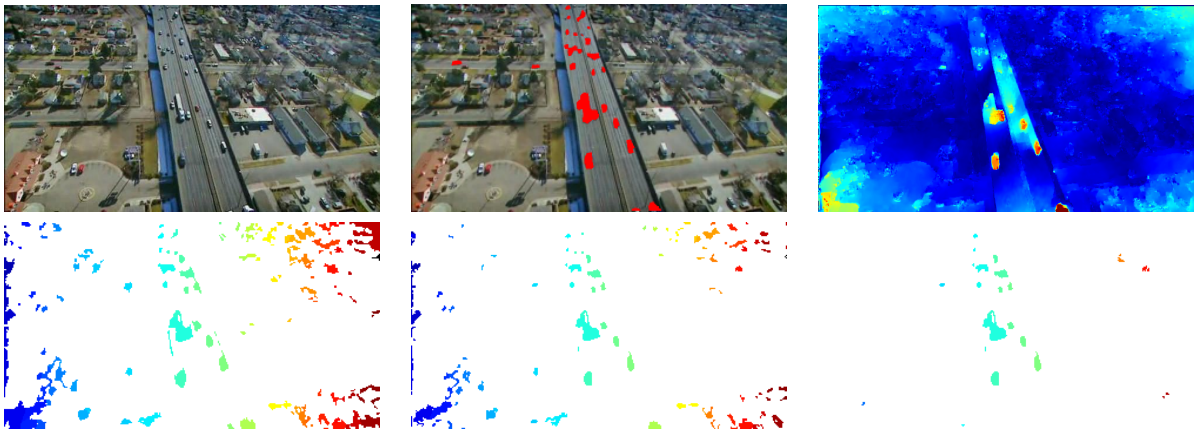
Le coût en temps et en mémoire de la méthode constitue un autre inconvénient. En effet, la structure de voisinage en trois dimensions créée occupe un espace mémoire important, même en ne considèrant que des 6-voisinages. De plus, la minimisation du potentiel global sur l'ensemble de cette structure prend un temps considérable, supérieur à une minute pour un segment temporel de 15 images consécutives.

6.2.4 Étude des pistes obtenues

Une autre limitation de cette méthode concerne sa portée temporelle. L'augmentation de la durée du segment temporel entraîne une explosion du coût de calcul et de mémoire. De plus, les résultats risquent de dériver, pour les images extrêmes du segment notamment. En revanche, les pistes obtenues ou "tracklets" peuvent être utilisées afin de filtrer, selon leur dynamique, leur apparence image et le contexte image des pistes (i.e., pour la détection de véhicules, vérifier l'existence de routes), d'éventuelles fausses détections. Les pistes restantes peuvent ensuite être recollées avec les pistes extraites des segments temporels voisins selon la cohérence de leur apparence et dynamique.

Un filtrage supplémentaire est donc nécessaire après obtention des étiquettes spatio-temporelles. Plusieurs critères d'analyse peuvent ainsi être appliqués aux pistes correspondant à un objet particulier :

- cohérence de l'apparence image et dynamique : couleurs, textures et flots résiduels ;
- analyse de formes : extension spatiale, lien entre extension et dynamique ;
- apparence du contexte spatial.



(a) (b) (c)

FIG. 6.11 – 1*ère ligne : (a) image centrale (b) avec vérité terrain (objets mobiles masqués en rouge) et (c) norme du flot résiduel correspondante.* 2*ème ligne : (a) étiquettes obtenues pour cette image, étiquettes après filtrage (b) de cohérence temporelle puis (c) d'apparence (taille / vitesse) et de contexte (réseau routier suffisamment présent)*

La cohérence de l'apparence vise à conserver des pistes associées à un unique objet et à filtrer les pistes de "bruit", dont l'apparence n'est pas stable. Des métriques adaptées sont par exemple des mesures de corrélation (une éventuelle rotation doit alors être préalablement compensée) ou des mesures d'évolution des modes couleurs et du flot résiduel.

L'analyse de formes est plus délicate, car dans le cadre de détections imprécises, il est possible que plusieurs objets soient fusionnés en une unique piste. La forme correspondante ne remplira alors pas nécessairement les critères correspondants (aire, rapports de dimensions, rapports entre aire et norme du flot résiduel...).

L'apparence du contexte spatial fournit un critère de tri supplémentaire. Ainsi, dans le cadre particulier de la détection d'objets mobiles sur routes, il est nécessaire que le voisinage de l'objet détecté contienne une part de route. Il faut au préalable disposer du réseau routier ou d'un modèle d'apparence de ce réseau et choisir des métriques et des conditions associées. La vérification de ces métriques permettra de confirmer, ou d'infirmer, la piste étudiée comme représentant la trajectoire d'un véhicule.

La figure 6.11 montre un exemple d'application de tels critères sur des étiquettes obtenues par Graph Cut sur un extrait de la séquence "are we destroying planet Earth". Les résultats sont ici particulièrement bruités car les primitives utilisées pour le terme d'attache aux données proviennent ici d'un flot optique après recalage affine. Ce flot résiduel entraîne des détections dues aux effets de parallaxe mais également du bruit (flot original ou imperfections de recalage).

La contrainte de cohérence temporelle permet déjà de supprimer une grande partie des fausses détections. Il subsiste cependant un nombre conséquent d'étiquettes à filtrer.

La figure 6.11 (c) illustre l'intérêt de deux critères supplémentaires. Une première contrainte sur le rapport maximal entre aire de la région étiquetée et la norme moyenne du flot résiduel portée au carré vise à filtrer les régions trop larges, correspondant typiquement à des erreurs de recalage dans les coins de l'image (distorsions radiales). Une deuxième contrainte impose dans le voisinage (distance inférieure à 2 pixels) de la région étiquetée une proportion de pixels de route supérieure à un pourcentage fixé (ici 20%). Le seuil est bas mais permet une certaine flexibilité par rapport aux défauts de détection du réseau routier, et prend en compte la possibilité qu'un véhicule soit situé en bordure de route (donc au maximum $\backsim$ 60% de route en comptant le côté intérieur, l'avant et l'arrière). Une grande majorité de fausses détections sont enlevées. En revanche, des petits véhicules dans la portion supérieure de l'image sont également filtrés à tort. La norme du flot résiduel étant en effet sur ces véhicules très faible, le critère sur le rapport entre aire et norme

TAB. 6.1 – *Précision et rappel (pixels) pour un segment temporel de 15 images de la séquence "arewe" calculés après recalage affine*

	Rappel	Précision
Flot résiduel seuillé à 0.1 pixel en norme	0.84	0.10
Flot résiduel seuillé à 0.2 pixel en norme	0.78	0.18
Flot résiduel seuillé à 0.5 pixel en norme	0.62	0.34
Flot résiduel seuillé à 1 pixel en norme	0.38	0.59
Graph Cut	0.86	0.14
Graph Cut et cohérence temporelle	0.73	0.24
Graph Cut, cohérence temporelle et critère de route	0.67	0.78

au carré n'est pas satisfait. Un relèvement du seuil permettrait de conserver ces objets mais également des fausses détections supplémentaires.

Le tableau 6.1 fournit des pourcentages de précision et de rappel sur un segment de 15 images de la séquence "are we destroying planet Earth". Un seuillage direct de la norme du flot optique résiduel, pour différents seuils correspondant aux quatre premières lignes du tableau, est comparé au Graph Cut décrit précédemment et dont les potentiels unaires sont tirés de cette même norme, avant (cinquième ligne) et après application des contraintes de cohérence temporelle (sixième ligne) puis de contexte routier (dernière ligne). Les différents défauts du flot optique résiduel, notamment dus à la distorsion radiale et à la parallaxe, expliquent les taux plutôt bas de précision pour un seuillage léger (0.1 pixel) du flot et le Graph Cut. Un seuillage plus élevé du flot permet d'obtenir des résultats plus précis, au détriment du rappel (59% de précision pour 38% de rappel pour un seuillage à 1 pixel). Les contraintes supplémentaires présentées ci-dessus permettent d'obtenir de meilleurs résultats : 78% de précision pour 67% de rappel.

6.2.5 Représentation des pistes

Afin de fournir une aide efficace pour l'interprétation de la séquence vidéo, les différentes pistes retenues doivent pouvoir être représentées sous des formes compactes et visuellement accessibles. Plusieurs options sont envisageables :

- choix du repère d'affichage : repère d'origine (évoluant avec la séquence) ou repère commun (après recalage global par rapport à une image de référence) ?
- choix de l'échelle de l'affichage : globale après réalisation d'une mosaïque de fond ; conservant l'échelle d'origine de la séquence (seules les parties "actives" de la séquence sont affichées, avec ou sans recalage) ; ou encore locale (chaque piste est affichée dans un contexte spatial local, ce qui permet de concentrer l'attention sur cette piste au détriment du contexte général) ?
- choix de la représentation des pistes : masques de couleurs ou variations d'intensité (assombrissement du fond par exemple) ; réduits aux centroïdes ou couvrant l'intégralité des régions pistées ?

Les figures 6.12 et 6.13 montrent une possibilité de représentation des objets détectés dans un repère commun (après recalage affine, l'image de référence étant l'image centrale du segment) afin de visualiser plus facilement le mouvement des objets dans un contexte immobile. Les limites du masque ont été choisies de manière à pouvoir représenter l'ensemble des vignettes pour chaque image du segment, ce qui explique que les vignettes ne soient pas centrées (sauf pour l'image centrale si le mouvement de l'objet est approximativement affine). Le choix d'une représentation locale permet de mettre directement l'accent sur l'objet détecté (et permet un gain de mémoire appréciable). Sans image repère (ici

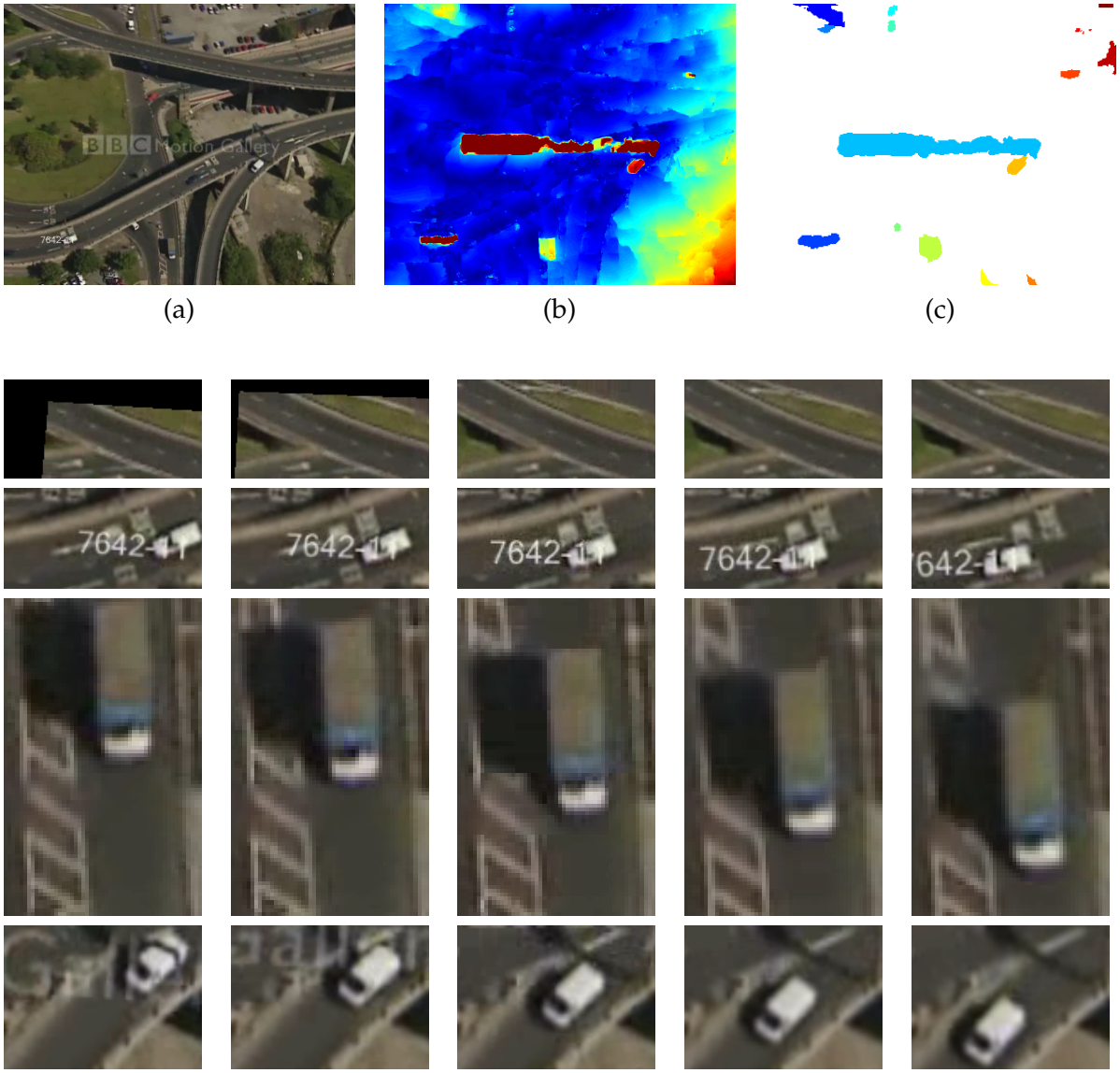


FIG. 6.12 – *Vignettes d'objets détectés par Graph Cut 3D, séquence "BBC". 1ère ligne : (a) image centrale (repère de référence), (b) norme du flot optique résiduel associé et (c) carte d'étiquettes. Lignes 2 à 5 : images 1,4,8 (centrale), 11 et 15 d'un segment de 15 images pour différents objets détectés. La première ligne correspond à une fausse détection, les autres à de vraies détections (parasitées par la watermark pour la deuxième ligne de la figure). Toutes les images représentent les objets dans le repère commun après recalage affine, avec pour image de réference l'image centrale du segment (cf. 1ère ligne).*

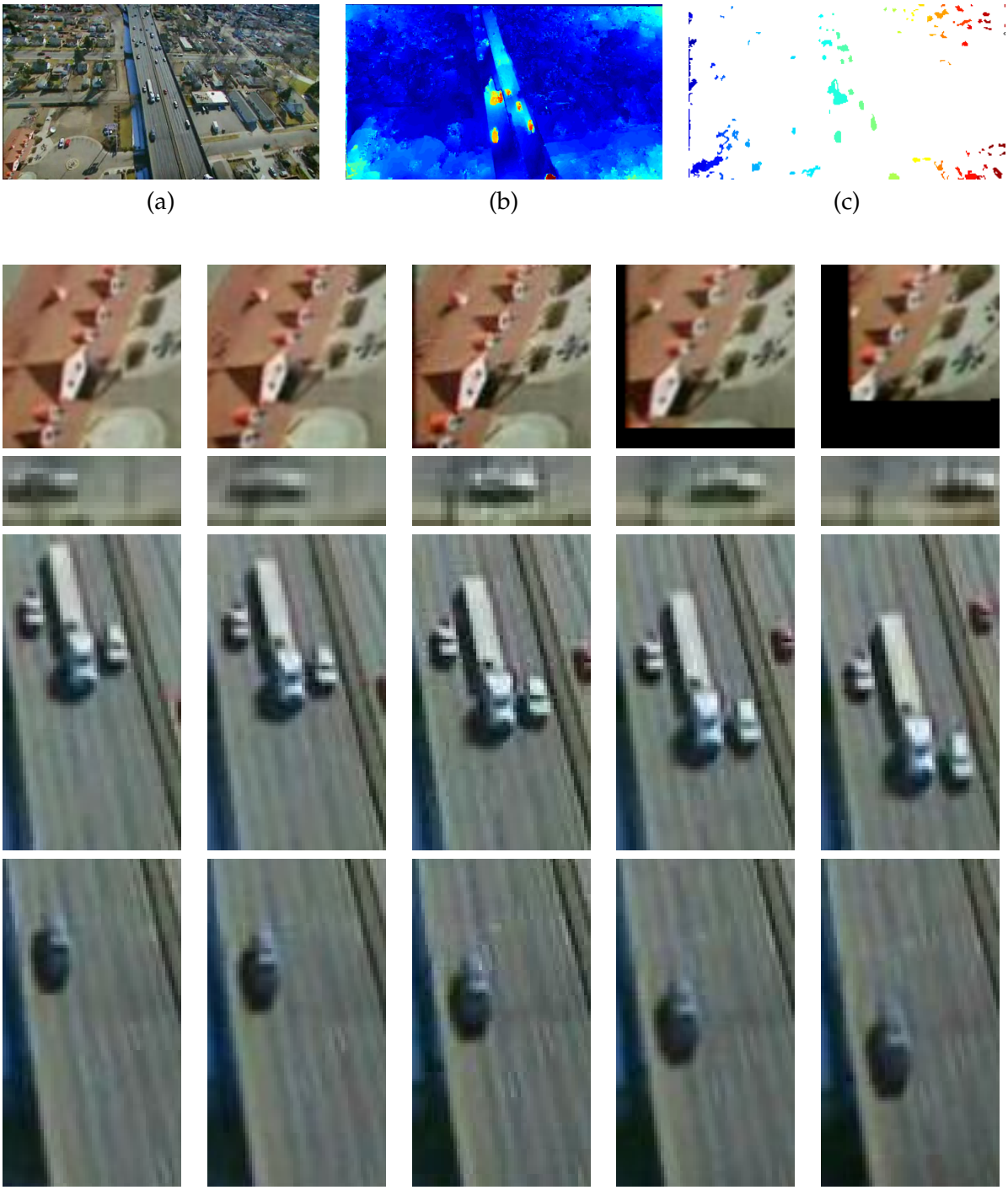


FIG. 6.13 – *Vignettes d'objets détectés par Graph Cut 3D, séquence "are we destroying planet Earth". 1ère ligne : image centrale (repère de référence). Lignes 2 à 5 : images 1,4,8 (centrale), 11 et 15 d'un segment de 15 images pour différents objets détectés. La première ligne correspond à une fausse détection, les autres à de vraies détections (parasitées par la watermark pour la deuxième ligne de la figure). Toutes les images représentent les objets dans le repère commun après recalage affine, avec pour image de réference l'image centrale du segment (cf. 1ère ligne).*

l'image présente en haut de chacune des figures), il peut être en revanche difficile de situer l'objet dans un contexte global (celui de l'image ou d'une mosaïque).

Les vignettes de la séquence "BBC" montrent divers exemples d'objets détectés (les trois dernières lignes). Le premier de ces trois, le véhicule blanc se déplaçant vers la gauche, est associé à la watermark "7642-11" qui possède après recalage un mouvement proche de celui de l'objet. Ils ont ainsi été fusionnés lors de la procédure de Graph Cut. La première ligne correspond à une fausse détection, la portion de route située en haut à gauche de l'image.

Les vignettes de la séquence "are we destroying planet Earth" montrent également des exemples d'objets correctement détectés (trois dernières lignes) et de fausse détection (première ligne). La fausse détection correspond ici à un effet de parallaxe due à la différence de profondeur du toit (situé sur l'image en bas à gauche). En l'occurrence, cette fausse détection peut être filtrée par l'application des critères précédemment cités de contexte (présence suffisante de route) et de rapport taille divisée par la norme au carré du flot résiduel moyen. L'avant-dernière ligne représente un groupe de trois véhicules de vitesse sensiblement égale. Une analyse fondée sur l'apparence image par exemple est donc nécessaire pour isoler chaque véhicule.

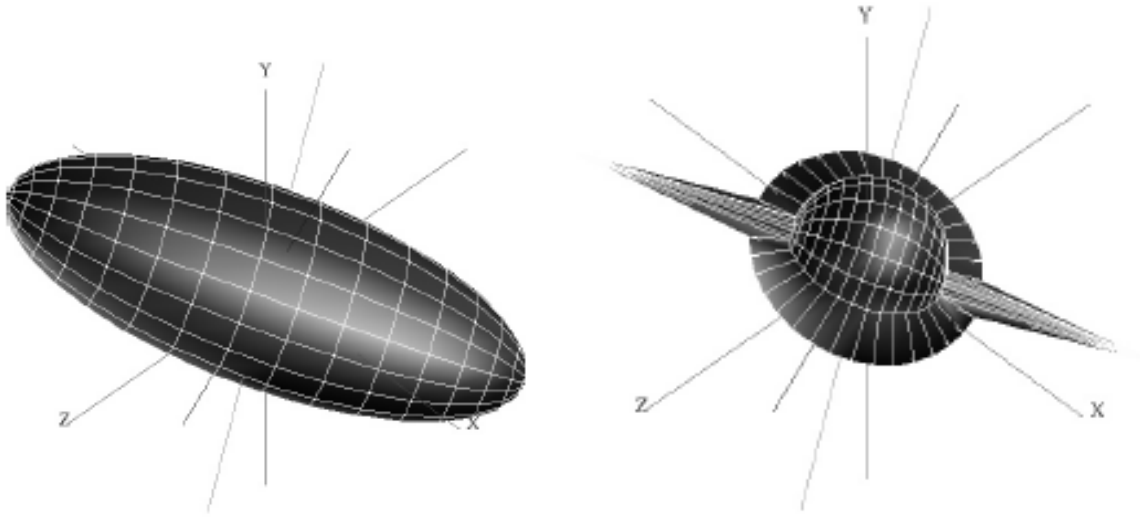
6.3 Approche volumique "directe" : extraction de flots d'activité cohérents par vote tensoriel

Il existe différentes approches analysant directement le volume spatio-temporel dans sa globalité, évoquées à la section 4.2. Nous avons choisi d'étudier plus particulièrement la méthode fondée sur l'utilisation du vote tensoriel. Cette approche défendue par Medioni et al. [279, 232, 280] diffère des approches classiques appliquant des critères successifs de filtrage, critères géométriques, critères de contexte ou critères d'apparence par exemple, afin de filtrer les fausses détections et les structures en trois dimensions. Elle se démarque également des approches de recherche de points ou de régions dites d'intérêt ou de "saillance", extension au cas 3D de la recherche de points d'intérêt sur images fixes.

6.3.1 Principe du vote tensoriel ou "tensor voting"

L'ordre de la variété locale formée par les points $4D$ correspondant aux pixels (x et y représentent la position image et v_x et v_y la vitesse résiduelle associée après recalage global dans un repère de référence) diffère selon la structure physique à laquelle les points sont associés. Ainsi, un filtre sur cet ordre permet d'extraire des structures spatio-temporelles ou des "flots d'activité" correspondant à des ensembles d'objets de mouvements cohérents, notamment des files de véhicules sur une même voie de circulation. Le vote tensoriel constitue un outil pour estimer l'ordre de la variété locale dans un espace de N dimensions, en l'occurrence de 4 dimensions. Cela comporte une double utilité. D'une part, la détection de ces structures permet de filtrer le bruit et les effets de parallaxe. L'approche globale devrait être plus robuste qu'une détection locale en espace et en temps. D'autre part, la forme des résultats est une première étape vers une interprétation sémantique de la scène. En effet, les structures détectées indiquent des groupes d'objets de comportement similaire ainsi que leur extension spatiale et temporelle : les voies de circulation routière ou plus généralement les chemins empruntés par des groupes d'êtres ou d'objets mobiles (piétons, animaux, cyclistes, véhicules...). Ces groupes sont extraits sans nécessiter un regroupement préalable de pistes isolées.

Le calcul tensoriel permet de représenter la structure locale du nuage de points sous une forme compacte associée à une matrice permettant également d'intégrer par addition matricielle les informations de structure des voisins. Cette diffusion des informations de structure, le vote tensoriel, pondère en chaque point "candidat" les tenseurs associés aux



(a) (b)

FIG. 6.14 – *Deux représentations tensorielles possibles d'un tenseur symétrique de rang* 3. *(a) Représentation ellipsoïdale (b) Représentation composite comme sommation d'un "tige", d'un disque et d'une sphère.*

points voisins "électeurs" en tenant compte de la direction de vote reliant "candidat" et "électeurs".

6.3.2 Représentation par tenseur

Les éléments fournis en entrée de la procédure de vote, par exemple des points de l'espace image en deux dimensions, ou des points du volume spatio-temporel en trois dimensions, sont tout d'abord codés sous la forme de tenseurs. Plus généralement, le tenseur code l'ordre de la variété locale et les directions principales en cas d'anisotropie locale des données. La figure 6.14 présente deux méthodes différentes de représentation d'un tenseur $\mathbf{T}$ symétrique de rang 3 : $\mathbf{T} = \sum_{i=1}^{3} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$. avec $\{\mathbf{e}_i\}_{i=1,2,3}$ les vecteurs propres et $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ les valeurs propres correspondantes.

Un ellipsoïde dont les axes principaux sont les vecteurs propres du tenseur (et les valeurs propres les dimensions caractéristiques associées telles que les demi-axes) peut représenter $\mathbf{T}$. Une telle représentation (figure 6.14 (a)) est toutefois difficile à interpréter à partir d'une projection $2D$ statique.

Un autre mode de représentation consiste à décomposer le tenseur comme somme d'une "tige", d'un disque et d'une sphère (figure 6.14 (b)) :

- le "tige" décrit la direction principale du tenseur $\lambda_1 \mathbf{e}_1 \mathbf{e}_1^T$ et sa longueur est proportionnelle à λ_1 ;
- le disque décrit le plan couvert par les vecteurs propres correspondant aux deux plus grandes valeurs propres et de dimension proportionnelle à λ_2 : $\lambda_2 \left(\mathbf{e}_1 \mathbf{e}_1^T + \mathbf{e}_2 \mathbf{e}_2^T\right)$;
- la sphère, de rayon proportionnel à λ_3, traduit le caractère plus ou moins isotrope du tenseur : $\lambda_3 \left(\mathbf{e}_1 \mathbf{e}_1^T + \mathbf{e}_2 \mathbf{e}_2^T + \mathbf{e}_3 \mathbf{e}_3^T\right)$.

Différents cas particuliers de tenseurs décrivant des structures caractéristiques de nuages de points peuvent être déclinés, en restant en dimension 3 pour une plus grande clarté de représentation :

- un tenseur d'ordre 1 avec $\lambda_1 \gg \lambda_2 \simeq \lambda_3$ et $\mathbf{T} \simeq \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T$ décrit une structure locale plane de nuage. $\mathbf{e}_1$ décrit alors l'orientation du vecteur normal à ce plan.

- un tenseur d'ordre 2 avec $\lambda_1 \simeq \lambda_2 \gg \lambda_3$ et $\mathbf{T} \simeq \lambda_1 \left(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_2\mathbf{e}_2^T\right)$ décrit une structure locale "fibreuse" ou selon une ligne. L'orientation de cette ligne est donnée par $\mathbf{e}_3$.
- un tenseur d'ordre 3 avec $\lambda_1 \simeq \lambda_2 \simeq \lambda_3$ et $\mathbf{T} \simeq \lambda_1 \left(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_2\mathbf{e}_2^T + \mathbf{e}_3\mathbf{e}_3^T\right)$ décrit un nuage non organisé selon une direction ou un plan particulier, isotrope. Il s'agit généralement de bruit sans cohérence spatiale ni temporelle.

En trois dimensions, dans le cas de points d'un nuage sans information de structure, ces points sont codés par des tenseurs sphériques tridimensionnels. De la même manière, des points éléments de courbe (respectivement de surface) seraient codés par des tenseurs disques (respectivement "tiges") comme présentés au paragraphe précédent.

Dans le cas général (toujours en dimension 3), le tenseur s'écrit comme une combinaison linéaire de ces trois cas particuliers :

$$\mathbf{T} = (\lambda_1 - \lambda_2)\mathbf{e}_1\mathbf{e}_1^T + (\lambda_2 - \lambda_3)(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_2\mathbf{e}_2^T) + \lambda_3(\mathbf{e}_1\mathbf{e}_1^T + \mathbf{e}_2\mathbf{e}_2^T + \mathbf{e}_3\mathbf{e}_3^T) \tag{6.1}$$

L'**ordre** du tenseur peut alors être défini comme l'indice du plus fort des trois coefficients de pondération correspondants : 1 s'il s'agit de $(\lambda_1 - \lambda_2)$, 2 s'il s'agit de $(\lambda_2 - \lambda_3)$ et 3 pour λ_3. Le caractère plus ou moins marqué de la structure correspondante du nuage de points (plane, "fibreuse" ou isotrope) peut être déduit de la comparaison de ces trois coefficients, par exemple du rapport entre les deux coefficients les plus élevés. Une valeur élevée de ce rapport indiquera une structure claire, au contraire d'une valeur proche de 1.

6.3.3 Vote non linéaire pour la propagation d'information

Les tenseurs associés aux points propagent l'information de structure qu'ils contiennent à leurs voisins, ce qui permet de définir une structure locale. Ainsi, si les points sont situés sur une courbe (respectivement un plan), ils seront représentés par un tenseur adapté (respectivement un tenseur "tige" ou disque). Une seconde phase permet de remplir l'espace (les "trous" du nuage original) en extrapolant par le biais de la représentation tensorielle l'information de structure aux points manquants de l'espace, il s'agit d'un vote tensoriel dense.

Les tenseurs sont découpés en tenseurs élémentaires ("tige", disque et sphère dans le cas 3D) selon l'équation (6.1). Les tenseurs disques et sphères sont obtenus par intégration de tenseurs "tiges", eux-mêmes créés par rotation d'un tenseur "tige" élémentaire. Le vote correspondant à de tels tenseurs (disques ou sphères) est donc obtenu en sommant les votes des tenseurs "tiges" correspondants.

La figure 6.15 illustre la procédure de vote. Un tenseur "tige" (associé à une normale $\mathbf{N}$) en un point "électeur" O vote pour une normale (ou tenseur "tige") $\mathbf{N}_P$ au point récepteur P avec un score V dépendant de la distance curviligne d et de la courbure ρ en suivant un chemin circulaire $2D$ entre les deux points : $V(d,\rho) = \exp\left(-\frac{d^2+c*\rho^2}{\sigma^2}\right)$ où c est un paramètre contrôlant la vitesse de décroissance avec la courbure et σ l'échelle de vote qui détermine la taille effective de la fenêtre de vote. L'orientation de la normale $\mathbf{N}_P$ est orthogonale au chemin circulaire reliant O et P en étant orthogonal en O à $\mathbf{N}$.

6.3.4 Étude des votes

Les votes sont accumulés par simple addition des tenseurs (soit des matrices 3×3 en trois dimensions), opération de faible coût calculatoire. Il suffit ensuite d'extraire les vecteurs et valeurs propres du tenseur résultant. Les composantes "tige", disque et sphère représentent respectivement des structures planes, des courbes ou des intersections de courbes (en trois dimensions). Les scores associés à chaque composante et détaillés au paragraphe 6.3.2 indiquent directement la vraisemblance que le point appartienne à une telle structure ou au contraire ne soit que du bruit.

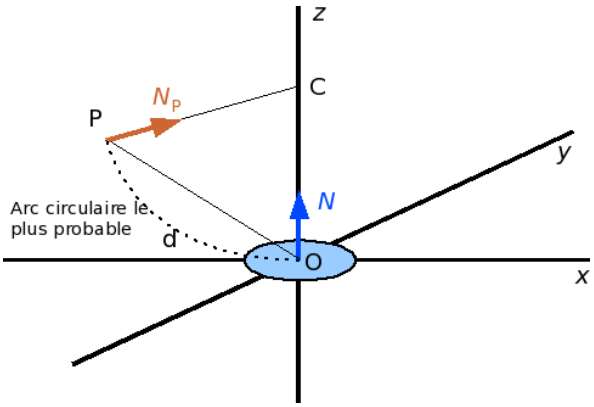


FIG. 6.15 – *Vote du point O (avec normale* **N***) pour le candidat P. Sont représentés l'arc circulaire reliant les deux points et de normale* **N** *en O, ainsi que le centre du cercle C et la normale "élue"* $\mathbf{N}_P$ *en P. d est la distance curviligne associée à l'arc, et la courbure* ρ *se déduit du rayon du cercle.*

6.3.5 Application au cas de l'extraction de flots d'activité en vidéo aérienne

Medioni et al. [279] utilisent l'outil de vote tensoriel afin d'extraire des flots d'activité. Les primitives utilisées sont les coordonnées des points dans l'espace en quatre dimensions (x, y, v_x, v_y) où (x, y) représente la position et (v_x, v_y) le flot résiduel dans un repère mosaïque de référence. Les images ainsi que les flots résiduels sont donc préalablement recalés dans ce modèle, ce qui revient à compenser le mouvement global (estimé par exemple par une affinité).

Dans une première étape, à une échelle fine, seules les structures d'ordre 2 et de score élevé sont conservées. Cela permet de filtrer une partie importante du bruit, souvent associé à un score faible. Une deuxième étape permet de filtrer la parallaxe. En effet, à une échelle plus grossière, la dimension du déplacement d'un véhicule (ou file de véhicules) est bien supérieure à ses propres dimensions. La structure correspondante forme donc une "tige" d'ordre 1. En revanche, les bâtiments de petite taille sont réduits à un point isolé et ceux de grande taille produisent une structure planaire d'ordre 2.

Ces deux étapes permettent donc de filtrer une grande partie du bruit et de la parallaxe. Nous avons effectué plusieurs tests sur les différentes séquences d'évaluation. La figure 6.16 illustre les différentes étapes du processus, pour un segment temporel ici de 100 images :

- le recalage global dans un repère commun de référence associé à l'une des images du segment, ici la première image, et la construction d'une mosaïque du fond (image (a) de la première ligne). Une telle représentation est déjà utile car elle agrandit le champ de vue et permet par conséquent d'appréhender le contexte à plus grande échelle. Elle peut également être utilisée comme critère de détection d'objets en mouvement comme précisé à la section 6.2.1.

- les flots résiduels sont également recalés dans le repère de la mosaïque, avec compensation du mouvement global (donc notamment du facteur d'échelle et de la rotation). L'image (a) sur la deuxième ligne affiche les normes des flots résiduels après recalage et après seuillage, seuls les points associés à un flot résiduel significatif (ici de norme supérieure à 1 pixel) sont ainsi conservés. Ces normes sont affichées simultanément pour l'ensemble des points des 100 images satisfaisant cette condition de seuil, le temps a été ainsi marginalisé dans la visualisation. On distingue déjà sur cette image des flots d'activité dans la partie droite de l'image : un vertical, un en courbe, ainsi que deux trajectoires correspondants aux véhicules (un blanc, l'autre bleu et gris, visibles sur la mosaïque) situés sur la route comportant des "zébra". Les autres régions

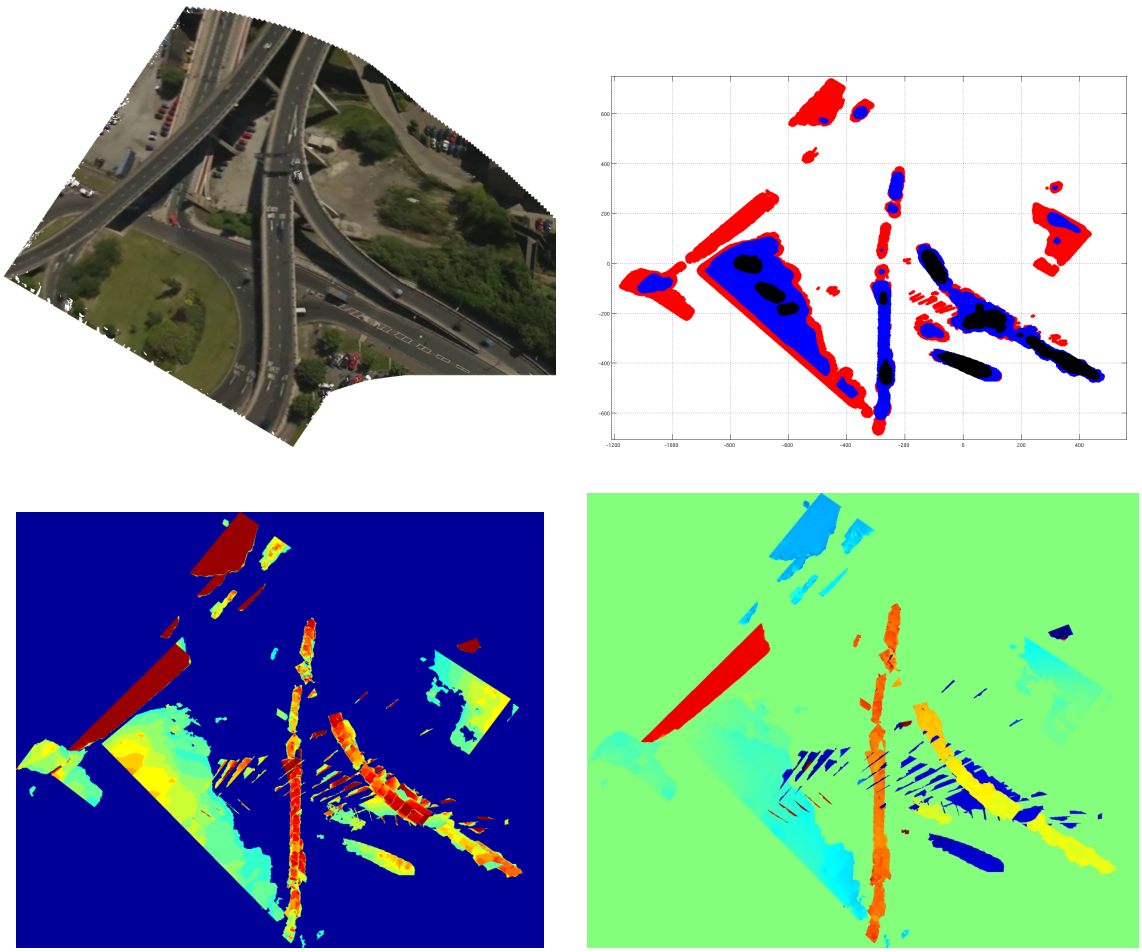
affichées correspondent à des restes de watermarks (éléments fins saccadés au centre de l'image) ou sont dus à des erreurs d'estimation du flot résiduel (la compensation affine du mouvement global n'est ici pas parfaite et crée des régions de résidus dans les coins des images).
L'image (b) de la deuxième ligne affiche les phases correspondantes. On remarque sur cette image la cohérence des phases des différents flots d'activité. Les phases des régions de watermark et de bruit sont également cohérentes.

– le filtrage des points par l'ordre du tenseur associé après l'étape de vote tensoriel dans l'espace en quatre dimensions (x, y, v_x, v_y) normalisé. Cette étape utilise l'exécutable Windows disponible sur le site de Medioni [NDT]. Sont représentés sur la figure (b) de la première ligne les points d'ordre 2 dont les scores (i.e., les coefficients de pondération définis au paragraphe 6.3.2) sont supérieurs à, respectivement 100 (en rouge), 200 (en bleu) et 500 (en noir) (le maximum étant ici de l'ordre de 2000). Ces résultats appellent plusieurs remarques. Tout d'abord, la majorité des éléments de bruit ont été filtrés car d'ordre différent de 2 : ces points n'étaient pas associés à une structure cohérente. De plus, une partie importante des fausses alarmes dues aux erreurs de recalage (les coins de l'image) ont un score faible, ici inférieur à 200.
En revanche, les pixels situés au centre de la mosaïque, décrivant la trajectoire d'un véhicule, obtiennent également des scores peu élevés. Cela s'explique par un manque de données, en effet, les points correspondants sur les images d'origine du segment étaient proches de la watermark centrale (supprimée sur la mosaïque par régularisation temporelle) et les portions du véhicule situées sur la watermark n'ont pas été retenues (grâce à un masque de la watermark centrale) afin de ne pas fausser l'estimation de structures. Enfin, plusieurs zones associées aux coins sont cohérentes et ont des scores élevés. Le passage à une échelle plus grossière afin de filtrer ces zones apparaît donc comme nécessaire.

La projection dans un espace de dimension supérieure (ici 4) et l'utilisation de l'outil de vote tensoriel afin d'extraire des structures ou des 7"flots" d'activité cohérents se révèle ainsi être une approche globale intéressante pour filtrer le bruit, mais les défauts constatés à partir de séquences et flots résiduels bruités (ainsi que les résultats obtenus en ne filtrant pas auparavant les watermarks) soulignent plusieurs limites.

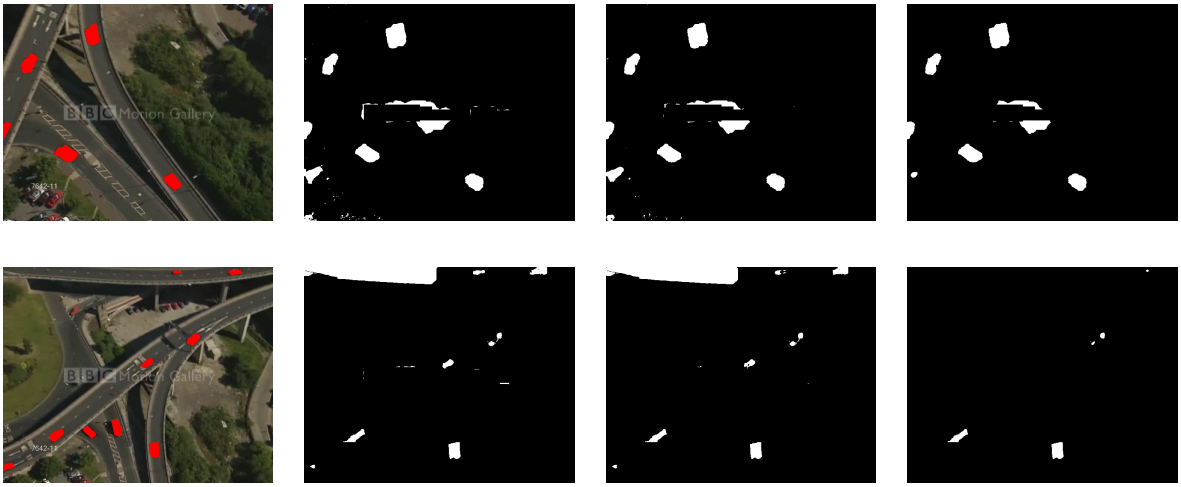
D'une part, la pertinence des résultats dépend de la précision des primitives, ici les flots résiduels après recalage, et de leur densité. Il est plus difficile d'obtenir des tenseurs d'ordre non ambigu (avec un score élevé par rapport aux autres coefficients) à partir d'un nombre limité d'échantillons. Un échantillonnage temporel à une fréquence faible (par exemple une image par seconde) afin d'alléger la quantité de données à traiter serait ainsi contre-productif. Les pourcentages de précision et de rappel obtenus sur 20 images d'un segment de 100 images consécutives avec un échantillon toutes les 5 images montrent ainsi une amélioration de la précision associée à une diminution du taux de rappel lorsque le seuillage sur le score augmente. Le filtrage ne conservant que les points d'ordre 2 ne change que peu les pourcentages. En effet, les quelques éléments de bruit filtrés par cette sélection sont composés de peu de pixels, et leur suppression n'intervient donc que peu dans les métriques calculées sur l'ensemble des pixels détectés sur les 20 images.

Sur ce segment temporel, peu de véhicules sont présents. Ceci couplé à l'échantillonnage temporel conduit à des structures plus lâches dans l'espace à 4 dimensions. Le filtrage du bruit est ainsi plus difficile. Les défauts de recalage sont également importants, notamment dans les coins des images (cf. figure 6.16) et la précision correspondant à un simple seuillage du flot est donc faible, aux alentours de 30%. L'ajout du critère de score permet d'améliorer la précision au détriment du rappel, mais même un seuil élevé du score fournit une précision limitée, environ 55%. Les structures cohérentes provenant



(a) (b)

FIG. 6.16 – *1ère ligne : (a) Mosaique de fond réalisée à partir d'un segment temporel de* 100 *images consécutives ; (b) après vote tensoriel, représentation des points d'ordre* 2 *de scores respectivement supérieurs à* 100 *(rouge),* 200 *(bleu) et* 500 *(noir). 2ème ligne, (a) et (b) normes et phases des flots résiduels des pixels de ces* 100 *images après recalage dans le repère de référence de la mosaïque. Seuls les pixels pour lesquels les normes associées étaient supérieures à un seuil (ici* 2 *pixels) ont été conservés.*



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

FIG. 6.17 – *Filtrage des masques binaires de détection sur deux images du segment temporel par vote tensoriel. Le repère ici utilisé est celui de la séquence originale, sans recalage. (a) Image de la séquence originale avec masque rouge de vérité terrain pour la classe "objets mobiles" ; (b) masque obtenu par simple seuillage de la norme du flot résiduel à* 1.5 *pixels ; (c) masque correspondant aux points extraits d'ordre* 2 *dans l'espace en* 4 *dimensions* (x, y, v_x, v_y) *après recalage affine ; (d) masque correspondant aux points extraits d'ordre* 2 *et dont le score est supérieur à* 200 *dans ce même espace.*

TAB. 6.2 – *Précision et rappel (pixelliques) pour le filtrage par vote tensoriel sur un segment temporel de* 100 *images, avec un pas en temps de* 5 *images, de la séquence "BBC"*

	Rappel	Précision
Flot seuillé à 1.5	0.69	0.31
Pixels d'ordre deux	0.68	0.32
Pixels d'ordre deux et de score ≥ 200	0.64	0.38
Pixels d'ordre deux et de score ≥ 500	0.45	0.44
Pixels d'ordre deux et de score ≥ 800	0.23	0.51
Pixels d'ordre deux et de score ≥ 1000	0.09	0.55

des erreurs de recalage doivent donc être filtrées à une échelle plus grossière lors d'une étape supplémentaire. En revanche, le seuillage de la norme du flot résiduel à 1.5 pixels, pour des raisons de coût de calcul, supprime déjà une portion des véhicules. Il est donc envisageable de fournir comme points d'entrée à l'algorithme de vote tensoriel non pas un simple seuillage du flot résiduel, mais les résultats d'un algorithme de classification ou de détection plus élaboré, contenant encore des fausses alarmes mais avec un rappel plus proche de 100%.

En supposant toutefois que les échantillons soient denses dans le temps et non excessivement bruités, se pose alors le problème du coût calculatoire. Même en ne conservant que les points associés à une norme de flot résiduel "significative" (il faut alors définir un seuil tel que 1 ou 2 pixels), le nombre d'échantillons est de l'ordre du million pour un segment temporel "court" de 100 images (soit 4 secondes à une fréquence de 25Hz). Les traitements d'un tel nombre de points dans un espace en dimension 4, détection des voisins et vote de tenseurs sphères par intégration de tenseurs "tiges" (obtenus par de multiples rotations dans l'espace 4D d'un tenseur "tige" original), demandent ainsi plusieurs dizaines de minutes pour traiter quelques centaines de milliers de points (correspondants à un segment temporel de 100 images). L'obtention de pistes correctement identifiées pour chacun des objets nécessite de plus des traitements supplémentaires tels que le recollement de flots d'activité (séparés à cause d'occultations par exemple) et la séparation des différents véhicules et pistes associées contenues au sein d'un même flot d'activité.

6.4 DISCUSSION

Les différentes approches globales en temps étudiées, qu'il s'agisse d'approches "semi-" globales par régularisation temporelle ou de considérer véritablement le volume spatio-temporel dans son ensemble, permettent de combler, du moins partiellement, les lacunes des approches purement locales en temps.

L'apport peut être sur le plan des performances de détection, avec une précision accrue pour un rappel équivalent (ou, de façon équivalente, des taux d'erreurs réduits). Ainsi, la régularisation temporelle de cartes de détection binaires ou à valeurs réelles permet de filtrer une partie des détections incohérentes. Si les données d'origine (les images de la séquence vidéo) et les primitives qui en dérivent telles que les flots optiques ne présentent pas un comportement instable voire chaotique au cours du temps, l'application d'une phase de régularisation temporelle en pré-traitement sur ces primitives conduit également à des résultats plus précis.

La segmentation de flots d'activité, par exemple par le biais de l'outil de vote tensoriel présenté à la section 6.3, permet d'accéder directement à une interprétation sémantique "haut niveau" de la séquence sans nécessiter la construction de pistes d'objets isolés, application intermédiaire pouvant être jugée accessoire selon la finalité d'application. Ainsi, le

Tab. 6.3 – *Comparaison des approches globales en temps*

Approche	robustesse au bruit et parallaxe	pistes	gestion des occultations	temps de calcul
Filtrage par connexité temporelle	partielle	non	non	faible
Graph-Cut 3D	partielle	oui	partielle	élevé
Vote tensoriel	oui	non	oui (post-traitement)	élevé

contrôle de la densité du trafic ne suppose pas forcément de disposer des pistes de chaque véhicule.

Les approches globales sont également plus robustes aux occultations : les parties masquées peuvent être extrapolées à partir des pistes précédant et suivant l'occultation, après appariement. Hormis quelques cas extrêmes, tels que par exemple la présence de véhicules d'apparences identiques (mêmes marque et modèle, même couleur, même vitesse) au passage d'un tunnel, il est généralement possible de recoller les pistes correspondant à un unique objet mobile. L'application d'hypothèses supplémentaires est parfois nécessaire afin de définir le choix le plus probable en cas d'ambiguïté (dans l'exemple précédent, l'hypothèse d'un dépassement sera considérée peu probable si la durée d'occultation est courte et les vitesses des deux objets avant occultation sont similaires).

En revanche, la complexité des objets manipulés dans des approches globales s'accompagne de plusieurs difficultés. D'une part, l'apport d'une dimension supplémentaire conduit à des structures plus lourdes et généralement plus coûteuses à manipuler. D'autre part, le recalage nécessaire afin de pouvoir exploiter la cohérence temporelle peut créer des artefacts supplémentaires : erreurs de recalage ou apparition de zones de flot résiduel si le modèle choisi (souvent affine) de mouvement global traduit ce dernier de manière imparfaite.

Le tableau 6.3 résume plusieurs caractéristiques des différentes approches développées. La gestion des occultations n'est dans tous les cas pas immédiate. La structure des résultats du graph-cut $3D$ et du vote tensoriel, respectivement des étiquettes spatio-temporelles et des flots d'activité, permet de combler des "trous" dus à des occultations plus ou moins étendues dans le temps. La structure globale des flots d'activité est toutefois plus à même de réassembler des portions de flots que les étiquettes spatio-temporelles du graph-cut, plus localisées en temps. Parmi les trois méthodes présentées, seul le graph-cut fournit des pistes, chaque piste étant associée à une étiquette spatio-temporelle. Il est toutefois possible d'extraire des pistes à partir des flots d'activité mais cela nécessite une étape supplémentaire de traitement. De même, les résultats de l'heuristique de cohérence temporelle peuvent être fournis en entrée d'un algorithme de pistage. Enfin, cette heuristique présente par rapport aux deux autres approches des facilités d'implémentation et un temps de calcul réduit.

6.5 Conclusion et perspectives

Les différentes approches globales en temps présentent plusieurs avantages par rapport à des approches purement locales. Outre une précision accrue des détections, l'intégration de la cohérence temporelle aide à la constitution de pistes sans nécessiter une phase de pistage dédiée.

Un recalage préalable est nécessaire afin de compenser le mouvement global. Ce prétraitement des données est susceptible de générer des erreurs supplémentaires et donc de

dégrader les performances ; il est donc utile de disposer d'un ou de plusieurs critères de qualité afin d'estimer ces erreurs et d'en tenir compte au sein de l'algorithme de détection.

Si l'ensemble des méthodes décrites n'est pas exhaustif, il semble cependant difficile de sélectionner une approche ne présentant aucun inconvénient. Les méthodes de régularisation temporelle sont rapides mais sont sensibles à des erreurs de recalage et peuvent dériver. L'intégration de contraintes de cohérence temporelle et d'*a priori* de détection (obtenus par classification ou par application directe de critères sur un ensemble de primitives extraites de la séquence, par exemple en utilisant un graph-cut dans l'espace-temps), se heurte à des difficultés de pondération des termes de l'énergie minimisée. Les méthodes volumiques dont le but est d'extraire des "tubes" ou "flots" d'activité sont généralement coûteuses en mémoire et en temps de calcul et dépendent de la densité d'activité.

Quelle que soit l'approche choisie, il semble souhaitable d'intégrer des critères de contexte et d'apparence spatio-temporelle en complément de contraintes de cohérence et de contraintes géométriques afin de filtrer des fausses alarmes (parallaxe, bruit). Une telle intégration est en revanche susceptible de dégrader les performances si les critères correspondants sont de faible qualité. Cela implique de relâcher chaque critère en fonction de mesures de qualité associées afin de ne supprimer aucun objet mobile réel, quitte à supprimer peu de fausses détections. Le point de fonctionnement est ainsi déplacé mais l'évolution des taux d'erreur ou de précision et de rappel est difficile à contrôler à partir d'une telle relaxation.

Perspectives Une évaluation complète sur plusieurs séquences vidéo est nécessaire afin de mesurer quantitativement l'impact de chaque facteur de variation sur les résultats de détection : qualité du flot optique estimé et capacité du modèle de mouvement global à compenser ce dernier ; fréquence de l'échantillonnage temporel ; densité des objets à détecter. La robustesse vis-à-vis de la parallaxe et des variations d'échelle devrait être également mesurée de manière quantitative par le biais de mesures d'erreur ou de précision et de rappel sur des séquences comportant de tels effets en proportions variables. Les essais effectués sur différentes séquences montrent en effet des performances de détection, en précision et en rappel estimés visuellement à partir des images de résultats, variables selon la difficulté des séquences.

La constitution de pistes à partir des résultats des différentes méthodes a été également évoquée mais doit faire l'objet d'une évaluation indépendante. Les résultats intermédiaires tels que les flots d'activité, les images de détection à chaque pas de temps ou encore les étiquettes spatio-temporelles nécessitent en effet des traitements supplémentaires dans un but de constitution de pistes.

Les approches travaillant directement sur le volume spatio-temporel ou celles de régularisation temporelle peuvent être combinées afin de fournir des résultats (cartes de détection, pistes d'objets, flots d'activité, voire modèles d'apparence et de dynamique des objets) de meilleure qualité, avec moins d'erreurs de détection, des pistes exactes et robustes aux occultations... Il est ainsi tentant d'incorporer des contraintes de régularité spatiale et temporelle des images et des flots optiques, avec ou sans recalage, ainsi que des critères de forme et de dynamique, non pas comme post-traitements permettant de filtrer des pistes incohérentes mais dans la formulation même d'une fonctionnelle de segmentation du volume spatio-temporel, sur le principe de [203] avec inclusion de contraintes de régularité. Une partie de ces contraintes, traduisant localement en temps la cohérence d'apparence et de flot, sont d'ailleurs incluses dans la formulation du Graph-Cut mais des critères d'apparence spatio-temporelle plus complexes nécesssitent un traitement supplémentaire.

La représentation des différents types de résultats mentionnés mérite une étude plus complète. Cet aspect interactif impliquant une part de subjectivité, il semble pertinent de proposer différents formats de représentation à l'interprète : dans le repère original de la séquence ou dans un repère commun, intégrant une mosaïque afin de bénéficier d'un champ de vision élargi, un affichage des pistes, la possibilité d'effectuer des requêtes sur l'apparence ou la dynamique des pistes... L'interaction est également un moyen d'intégrer un retour de pertinence dans le cadre de procédures d'apprentissage.

Conclusion

Les travaux décrits dans cette thèse s'orientent autour du filtrage de segments informatifs dans des séquence vidéo et visent spécifiquement l'analyse de données issues de capteurs aériens. L'interprétation manuelle par un opérateur humain de vidéos aériennes dans une optique de renseignement se heurte au volume sans cesse grandissant des données disponibles. Une assistance algorithmique fondée sur diverses modalités d'indexation est ainsi envisagée dans l'objectif de repérer les "segments d'intérêt" et d'éviter une inspection intégrale fastidieuse de la vidéo.

La problématique a été abordée sous deux angles d'analyse complémentaires : l'étude des *conditions de prise de vue ou CPDV*, celle du *contenu dynamique*, plus particulièrement la *détection d'activité*. Ces deux caractérisations de séquences vidéo sont à replacer dans un cadre plus large de systèmes d'aide à l'indexation qui comporte une part importante d'interactions entre l'interprète humain et des traitements algorithmiques.

Les différents axes d'étude abordés sont complémentaires mais suffisamment distincts pour que les états de l'art correspondants soient traités séparément. Le chapitre 1 cible plusieurs objectifs. Il présente un aperçu des différentes chaînes de traitement et les solutions existantes de systèmes d'aide à l'indexation (et de recherche) de séquences vidéo. Cette présentation des systèmes dans leur ensemble vise également à apporter un éclairage d'ensemble sur la problématique d'aide à l'indexation soulevée, en soulignant notamment les interactions entre les différentes briques d'indexation. Trois éléments principaux y sont développés : les interfaces homme-machine ; les éléments d'interprétation sémantique manipulant des concepts de haut niveau tels que des scénarios ou compositions spécifiques d'activités ; des éléments de traitement plus bas niveau, produisant des sorties déjà interprétables mais moins complexes d'un point de vue sémantique, telles que des pistes d'objets.

Première partie : conditions de prise de vue La partie 1 regroupe les chapitres 2 et 3. Elle s'est attachée plus particulièrement à l'étude des *conditions de prise de vue*. Cette dernière fournit plusieurs modalités d'indexation d'une séquence vidéo.

La recherche de segments temporels peut être directe, par seuillage de critères de qualité ou sélection de classes particulières de mouvement global, ou plus indirecte, en contribuant à construire des objets spatio-temporels de résumé par exemple. Cependant, les critères choisis peuvent être ambigus selon le contenu de la séquence. Ainsi, la présence de détails, une incidence rasante, des structures en trois dimensions sont autant d'éléments qui perturbent l'analyse des CPDV. Un découpage possible des conditions de prise de vue distingue leur état en un moment précis d'une part, leur évolution d'autre part.

Nous avons choisi d'étudier plus particulièrement des critères de *qualité image*, notamment l'évaluation du flou de bougé ou de mise au point. Ces déformations du contenu image sont en effet un critère important pour la sélection des segments vidéo peu voire non exploitables car trop dégradés. Trois approches d'estimation du flou ont été ainsi décrites et évaluées dans le chapitre 2 sur plusieurs images classiques ainsi que sur des images tirées de séquences vidéo. L'utilisation de mesures de qualité comme modalités d'indexation a été également abordée. Il apparaît toutefois des difficultés dues à l'évolution du contenu

de la séquence, ce qui perturbe la construction d'une mesure de qualité robuste au contenu.

L'*étude du mouvement global*, directement lié au mouvement de la caméra par rapport à la scène observée, relève également des conditions de prise de vue. Nous avons construit et évalué dans le chapitre 3 une classification du mouvement global afin de fournir une modalité d'indexation supplémentaire de séquences vidéo.

La fiabilité d'une telle classification dépend de plusieurs facteurs, dont la qualité de l'estimation du mouvement image et le modèle de recalage utilisé. La comparaison avec une évaluation manuelle souligne les ambiguïtés possibles entre des mouvements proches mais de type différent (tels que rotations et translations). Elle montre également la subjectivité inhérente à la définition de classes de mouvement complexes cherchant à traduire l'intention de l'opérateur caméra. Une interface homme-machine permettant de définir de telles classes complexes à partir de classes élémentaires moins sujettes à ambiguïté est donc souhaitable.

Deuxième partie : détection d'activité Le contenu statique et dynamique d'une séquence vidéo ne doit pas être oublié. Si par exemple un zoom suivi d'une stabilisation de la caméra signifie un intérêt de l'opérateur caméra, il n'apporte aucune information sur la nature des possibles objets ou structures d'intérêt concernés. En revanche, il peut guider l'interprète vers les segments temporels pertinents de par leur échelle, dans lesquels les détails sont observables avec une grande précision, dans un but d'identification par exemple. Le recalage dans un repère commun permet aussi un tri spatial, par localisation d'une région sur une mosaïque et extraction des segments couvrant cette région. Le filtrage de segments moins informatifs, trop flous ou rendus difficilement exploitables de par leurs mouvements d'ensemble trop rapides ou oscillants, constitue une autre application des conditions de prise de vue.

La détection d'activité repose notamment sur des primitives de mouvement telles que le flot optique résiduel. Mais la présence de bruit de mesure, les effets de parallaxe et d'éventuelles variations de qualité image entraînent de nombreuses fausses alarmes voire un défaut de détection dans le cadre d'une approche purement fondée sur ces primitives.

Un état de l'art des différentes méthodes de détection de mouvement a été réalisé au chapitre 4. Nous y distinguons approches locales en temps, adaptées à une analyse "en ligne" de la séquence vidéo, et approches globales introduisant un délai et souvent plus coûteuses en ressources mémoire et en temps de calcul mais généralement plus précises.

Nous avons construit dans le chapitre 5 une approche locale en temps en plusieurs étapes, alliant classification locale, apprentissage itératif du contexte spatial et inclusion de contraintes de connaissance *a priori*. Cette approche couple les primitives de mouvement issues du flot optique avec des primitives d'apparence dans un cadre d'apprentissage local. Elle compense partiellement les erreurs d'algorithmes purement fondés sur l'utilisation du flot mais la dérive importante du contenu nécessite une mise à jour des classifieurs sur la base de nouvelles données de référence afin d'éviter un écart trop marqué entre base d'apprentissage et base de test (le reste de la séquence).

L'introduction d'informations de contexte à plusieurs niveaux, contexte local, régional ou sémantique, améliore les performances de détection mais introduit de nouvelles contraintes : qualité de la détection de structures pour l'application de critères sémantiques (dans l'application présentée, la qualité du réseau routier obtenu) ; robustesse aux objets de faibles dimensions avec le risque de rejeter les détections associées comme du bruit.

La fragilité des résultats de détection provenant d'une analyse locale en temps souligne la nécessité d'une approche plus globale. Le chapitre 6 décrit plusieurs approches globales sans recherche d'exhaustivité. Ainsi, l'ajout *a posteriori* de contraintes de cohérence

temporelle peut permettre de filtrer les fausses alarmes non cohérentes : il s'agit d'une régularisation temporelle de faible coût calculatoire mais sujette à des dérives (et donc limitée à des segments temporels de faible durée).

Le couplage de ces contraintes de cohérence avec l'initialisation de détections d'objets mobiles en utilisant un algorithme de type graph-cut en espace-temps permet d'aller plus loin en créant des étiquettes cohérentes dans le temps.

Un autre type d'approche, étudiant directement le volume spatio-temporel, intègre directement les contraintes de cohérence temporelle en recherchant des nappes spatio-temporelles représentant des ensembles de trajectoires d'objets en mouvement. La complexité des objets manipulés et la nécessité de disposer de l'ensemble de la séquence imposent en revanche un travail hors ligne.

Perspectives

Les différentes modalités d'indexation étudiées dans cette thèse représentent chacune un domaine de recherche à part entière. Les travaux réalisés ont révélé plusieurs limites qui représentent autant de perspectives directes évoquées au sein de chaque chapitre. En considérant l'aspect système plutôt que chaque modalité d'indexation prise séparément, plusieurs perspectives plus générales peuvent être dégagées.

- Une étape préliminaire de stabilisation est nécessaire pour la plupart des approches globales, ce qui crée des sources d'erreur supplémentaires. Il faut alors tirer parti de la richesse du volume spatio-temporel, de critères d'erreur et de contraintes de régularité pour limiter l'impact de ces erreurs.

- Les interactions homme-machine constituent un pilier de tout système d'indexation et de recherche pour plusieurs raisons : retour de pertinence, définition de requêtes complexes par combinaison de critères élémentaires résultant de différentes briques (pistage, mosaïcage, classification du mouvement apparent, cartes de profondeur...) ou encore l'aspect navigation avec création de résumés interactifs. Ces IHM ont été abordées au chapitre 1 mais sous la seule forme d'état de l'art. Les fonctionnalités présentées sont nombreuses, il faut donc choisir les plus appropriées au contexte d'application considéré : navigation, recherche, description de la scène...

- Les différents axes d'indexation ont été abordés séparément mais une collaboration entre plusieurs de ces axes pourrait les renforcer mutuellement en réduisant l'incertitude par un éclairage complémentaire. Ainsi, certains mouvements dominants, susceptibles d'indiquer un intérêt de l'opérateur caméra pour une région présentant de l'activité, pourraient orienter le module de détection d'activité.

- L'analyse du contenu statique pourrait être développée. En particulier, la caractérisation de l'environnement (tissu urbain, campagne, forêt, montagne, mer...) apporterait un éclairage contextuel non négligeable. L'identification précise de l'environnement reste toutefois une tâche difficile, de par une grande similarité visuelle entre différents concepts d'environnement et une diversité importante au sein de chaque concept. La conception de primitives spécifiques ou un apprentissage automatique intégrant d'importantes bases de données constituent des pistes possibles.
  La détection, la classification voire l'identification d'objets statiques peuvent compléter l'analyse de la scène, dans un but de compréhension sémantique "haut niveau". La cohérence temporelle de l'objet analysé, un flux vidéo, permet un gain de temps non négligeable par la propagation des étiquettes au cours du temps.

- Si la présence d'un interprète humain reste indispensable pour préciser les scéna-

rios recherchés, le "fossé sémantique" entre les données image et une interprétation de haut niveau sémantique est encore présent, notamment au niveau de la jonction entre activités élémentaires et scénarios complexes de situation. Il s'agit d'une piste supplémentaire à développer, notamment par apprentissage ou par le biais d'outils de logique spatio-temporelle.

Annexes

Flot optique

A

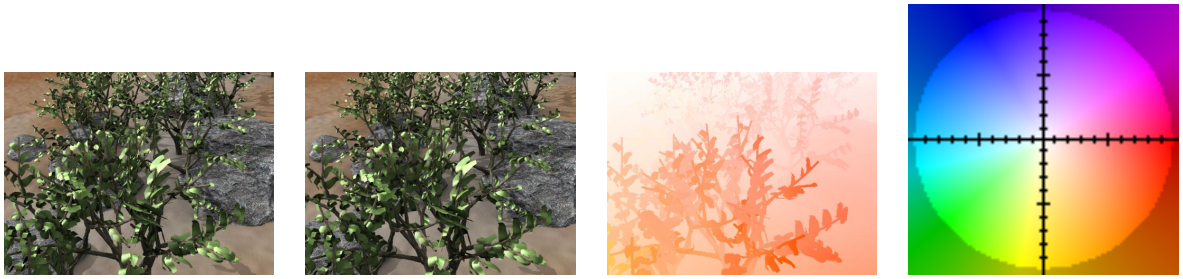
A.1 Introduction

Le champ de mouvement est le mouvement apparent d'objets, surfaces ou arêtes causé par le déplacement relatif de l'observateur par rapport à la scène. Le flot optique, qui peut différer du champ de mouvement, correspond au mouvement des pixels représentant ces différents éléments d'une image à la suivante, il s'agit du mouvement apparent des motifs d'illumination dans l'image. Le concept de flot optique a été étudié dans les années 40 et défini par le psychologue James J. Gibson en 1950 dans une étude sur la vision humaine [83].

L'estimation du flot optique fait partie de traitements de l'image dits de bas niveau. De nombreuses applications comprennent l'analyse de mouvements de fluides en phy-sique expérimentale [55], la compression de séquences d'images vidéo par compensation de mouvement [141], ou des phases de traitement des images de plus haut niveau, comme la reconstruction de scènes tridimensionnelles [274] ainsi que le mouvement en trois dimensions d'objets et de l'observateur par rapport à la scène observée. Le flot optique est également utilisé pour la détection et pistage d'objets [283], l'extraction de plan dominant, la détection de mouvement ou encore l'odométrie visuelle voire plus généralement l'aide à la navigation de robots [34]. L'estimation du flot optique présente également une application dans les problèmes de correspondance, dans lesquels il s'agit d'obtenir une fonction d'appariement suffisamment lisse associant les primitives d'une image aux structures d'une autre.

Le flot optique est un outil particulièrement important dans le cadre de notre étude. En effet, il est nécessaire pour un recalage préalable d'une séquence ou de segments vidéo. Le recalage en lui-même a de multiples applications telles que la constitution de mosaïques, le lissage temporel de primitives (régularisation du flot optique par exemple) ou la détermination du mouvement dominant. Mais le flot optique apporte également de précieuses informations sur le contenu dynamique de la séquence. Après compensation du mouvement dominant, le flot optique résiduel est une primitive essentielle pour la détection des objets mobiles, mais les structures en trois dimensions sont également associées à un flot résiduel non nul (effet de parallaxe).

Les séquences vidéo aériennes apportent plusieurs défis pour l'estimation du flot optique : régions homogènes, changement de luminosité, déplacements parfois de grande amplitude, qualité image variable (flou et défauts de compression), objets de petite taille... Ces difficultés représentent autant de critères d'évaluation et de choix de méthode pour l'estimation du flot optique. Il est d'ailleurs nécessaire de disposer de séquences diverses présentant un éventail des difficultés citées pour pouvoir évaluer les forces et faiblesses de chaque méthode et / ou implémentation.



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

FIG. A.1 – *Exemple de représentation couleur de flot sur la séquence synthétique Grove de Middlebury [15] : (a) et (b) couple d'images (c) flot (vérité terrain) (d) encodage couleur de l'amplitude et direction du flot*

A.2 Définition du flot optique et problème d'ouverture

Le problème d'estimation du flot optique peut être traduit mathématiquement sous la forme de systèmes d'équations reliant généralement intensité des pixels d'un couple d'images successives. Plusieurs hypothèses sont nécessaires afin de garantir l'unicité de la solution. Une première hypothèse couramment utilisée dans les méthodes d'estimation du flot optique considère que la luminosité apparente d'un point donné est conservée au cours du temps, au moins pour de courtes périodes. Il s'agit de l'hypothèse d'illumination constante : les intensités des pixels d'une image sont conservées d'une image à l'autre en suivant le flot optique correspondant. En notant $(x;y)$ les coordonnées d'un point suivi dans la séquence d'images, $(x(t);y(t);t)$ la trajectoire $2D$ du point au cours du temps dans la séquence et $I(x(t);y(t);t)$ l'intensité lumineuse au point $(x(t);y(t))$ sur l'image au temps t, cette hypothèse peut s'écrire ainsi :

$$I(x(t);y(t);t) = constante. \tag{A.1}$$

En dérivant cette expression par rapport au temps, on obtient l'équation suivante, dite contrainte du flot optique :

$$I_x \frac{dx}{dt} + I_y \frac{dy}{dt} + I_t = 0 \tag{A.2}$$

où I_x, I_y, I_t sont respectivement les dérivées partielles de l'image $I(x,y,t)$ par rapport à x, y et t. Le vecteur $(\frac{dx}{dt}, \frac{dy}{dt})$ représente la vitesse du point $(x;y)$ sur le plan image au temps t. Le champ de déplacement correspondant à l'ensemble des vitesses pour tous les points de l'image forme le flot optique. On peut réécrire (A.1) en prenant un pas de temps égal à 1 :

$$I(x(t+u(t));y(t+v(t));t+1) - I(x(t);y(t);t) = 0. \tag{A.3}$$

où $(u(t);v(t))$ représente le déplacement du point $(x;y)$ de t à $t+1$. L'équation (A.2) correspond alors à une linéarisation de Taylor d'ordre 1 au point $(x(t);y(t);t)$. Cette approximation est raisonnable si le champ de déplacement varie de manière suffisamment lisse et pour des déplacements réduits, de l'ordre du pixel ou inférieurs (cf. section A.4.2). Il s'agit de la base des méthodes fondées sur l'utilisation du gradient. Toutefois, cette hypothèse seule ne suffit pas à garantir l'unicité du flot optique car elle ne permet de calculer que la composante du flot dans la direction locale du gradient de l'intensité de l'image. En effet, ce dernier comporte deux composantes et sa détermination demande donc deux équations indépendantes : il s'agit du problème d'ouverture [249, 223]. Il est donc nécessaire de poser une hypothèse complémentaire. Il peut s'agir d'une contrainte de régularité globale ou de contraintes locales sur le modèle du flot. Des méthodes non fondées sur la conservation du gradient de luminance peuvent également être utilisées.

A.3 Calcul du flot optique

A.3.1 Méthodes fondées sur une analyse fréquentielle

Les méthodes utilisant le domaine fréquentiel se fondent sur l'emploi de filtres ajustés en vitesse. Les filtres utilisés dépendent de l'orientation dans le domaine de Fourier des images temporellement variables (volumes vidéo). Ces méthodes sont capables d'estimer le mouvement d'objets difficiles à appréhender par des méthodes d'appariement (cf. A.3.2) ou par primitives, tel que le mouvement de motifs aléatoires de points. Ainsi, Adelson et Bergen [4] proposent-ils plusieurs méthodes d'estimation du flot optique en calculant l'orientation spatio-temporelle du volume vidéo, par exemple en extrayant l'énergie spatio-temporelle à l'aide de filtres de Gabor. Jähne [115] a montré que la détection de l'orientation spatio-temporelle revenait à analyser les valeurs propres du tenseur d'inertie du volume vidéo. Ces différentes méthodes présentent des similarités avec des modèles de perception humaine, en intégrant la réponse de divers filtres spatio-temporels [89, 257].

D'autres méthodes se concentrent sur l'étude de la phase dans le domaine spatio-temporel. La vitesse est définie par Fleet et Jepson [76] comme le mouvement instantané des lignes de niveau de la phase dans les sorties de filtres de Gabor passe-bande qui décomposent le signal d'entrée suivant l'échelle, la vitesse et l'orientation. La phase est plus robuste aux variations d'illumination que les dérivées de l'intensité ou les méthodes fondées sur l'utilisation de l'énergie spatio-temporelle. En revanche, ces méthodes nécessitent un nombre important de filtres afin de recouvrir l'espace de Fourier.

A.3.2 Méthodes d'appariement par région

Ces méthodes sont particulièrement adaptées aux cas où la qualité du signal est faible ou lorsque le support temporel est réduit. Les méthodes différentielles ou fréquentielles sont alors moins efficaces. Les primitives saillantes telles que des coins ou jonctions en T, sont rares et les correspondances peuvent être difficiles à établir en cas de multiples instances. Les méthodes d'appariement par région définissent la vitesse comme étant le déplacement $\mathbf{u} = (d_x, d_y)$ permettant d'associer des régions de l'image à des instants différents. Pour une position $\mathbf{x} = (x, y)$, le meilleur déplacement optimise une quantité de comparaison sur $\mathbf{u}$, maximisation d'une mesure de similarité telle que la corrélation croisée normalisée (NCC) :

$$NCC_{1,2}(\mathbf{x};\mathbf{u}) = \frac{1}{(2n-1)^2} \sum_{j=-n}^{n} \sum_{i=-n}^{n} \frac{[I_1(\mathbf{x}+(i,j)) - m_1)(I_2(\mathbf{x}+\mathbf{u}+(i,j)) - m_2)}{\sigma_1 \sigma_2} \tag{A.4}$$

où m_1 et m_2, σ_1 et σ_2 représentent respectivement les moyennes et écarts-types de I_1 et I_2 sur des fenêtres de taille $(2n+1) \times (2n+1)$ respectivement centrées en $\mathbf{x}$ pour I_1 et $\mathbf{x}+\mathbf{u}$ pour I_2 ; ou minimiser une mesure de distance telle que la somme des différences carrées (SSD) :

$$\begin{aligned} SSD_{1,2}(\mathbf{x};\mathbf{u}) &= \sum_{j=-n}^{n} \sum_{i=-n}^{n} W(i,j)[I_1(\mathbf{x}+(i,j)) - I_2(\mathbf{x}+\mathbf{u}+(i,j))]^2 \\ &= W(\mathbf{x} * [I_1(\mathbf{x}) - I_2(\mathbf{x}+\mathbf{u})]^2 \end{aligned} \tag{A.5}$$

où W représente une fonction de pondération définie sur une fenêtre $2D$ de taille $(2n+1) \times (2n+1)$ et $\mathbf{u}$ prend des valeurs entières. Ces méthodes comprennent notamment [8], dans laquelle Anandan utilise une pyramide laplacienne et une stratégie d'appariement "coarse-to-fine" fondée sur une minimisation SSD. L'approche choisie par Singh [225] cherche également à minimiser un critère SSD et procède aussi en deux temps. La première étape consiste à calculer une probabilité de distribution pour $\mathbf{u}$ par l'intermédiaire de plusieurs filtres passe-bande adjacents de l'image, $\mathbf{u}$ correspondant à la moyenne

de cette distribution. La seconde étape lisse le champ de déplacement par l'utilisation de contraintes de voisinage.

A.3.3 Méthodes fondées sur l'utilisation du gradient

Critère quadratique Une estimation simple consiste à minimiser un critère des moindres carrés à partir de l'équation de contrainte du flot optique (A.2) en supposant un modèle localement constant de mouvement, selon l'approche de Lucas et Kanade [156] :

$$E(\mathbf{u}) = \sum_{\mathbf{x}} W(\mathbf{x})[\mathbf{u} \cdot \nabla I(\mathbf{x}, t) + I_t(\mathbf{x}, t)]^2 \tag{A.6}$$

où W une fonction de pondération sur un voisinage de $\mathbf{x}$, par exemple de type gaussien. Le champ de vitesses $\mathbf{u}$ obtenu est celui qui minimise $E(\mathbf{u})$. Lorsque l'image n'est pas assez texturée (ou si le support de W est trop local), le système aux dérivées partielles obtenu à partir de (A.6) ne permet pas de déterminer de manière unique $\mathbf{u}$ (problème d'ouverture).

Estimation itérative Dans le cadre de déplacements réduits, pour lesquels une approximation de Taylor de (A.3) est raisonnable, un schéma itératif peut être utilisé. Ainsi, une estimation du mouvement permet d'effectuer un recalage approximatif pour appliquer de nouveau la méthode d'estimation aux images recalées afin de trouver le mouvement résiduel dans une optimisation de type Gauss-Newton. Les itérations s'arrêtent à l'obtention d'un mouvement résiduel suffisamment faible. L'utilisation d'un schéma coarse-to-fine [25] permet de résoudre le problème de repliement temporel ainsi que d'accélérer le processus, le risque étant de propager une mauvaise initialisation aux niveaux les plus grossiers de la pyramide.

Modèles de mouvement local Les différents modèles présentés jusqu'à maintenant reposent sur l'hypothèse d'un flot localement homogène. Cette hypothèse est valable dans certains cas tels que le déplacement d'objets rigides dans un plan orthogonal à l'axe optique et sans variation de profondeur (déplacement de véhicules observé en visée nadir sur un terrain à altitude constante par exemple). Toutefois, en présence d'un mouvement caméra complexe ou de variations de profondeur relative entre la scène et l'objectif, cette hypothèse n'est plus vérifiée. D'autres modèles sont alors plus adaptés. Il peut s'agir d'un modèle affine, paramétrique tel que l'homographie [25], ou pouvant s'exprimer comme combinaison linéaire de champs de déplacement élémentaires, par exemple définis par apprentissage [75].

Lissage global Afin de résoudre le problème d'ouverture A.2, des hypothèses complémentaires sont nécessaires. Il s'agit souvent d'une hypothèse de régularisation avec l'introduction d'une contrainte de régularité spatiale du flot pour compléter la contrainte du flot optique (A.2). Cette hypothèse a été introduite par Horn et Schunk [104] en 1980 :

$$E(\mathbf{u}, t) = \int_{x,y} [\mathbf{u} \cdot \nabla I(x, y, t) + I_t(x, y, t)]^2 + \lambda \left(\|\nabla u_1\|^2 + \|\nabla u_2\|^2 \right) dxdy \tag{A.7}$$

avec $\mathbf{u} = (u_1, u_2)$. Cette méthode permet la propagation de l'information sur de grandes distances dans l'image, ce qui permet par exemple de remplir à partir des bords le flot sur des zones homogènes de l'image, au contraire de méthodes locales sous-dimensionnées. En revanche, ces méthodes globales sont plus coûteuses que des méthodes locales. L'ajustement du paramètre de régularisation λ contrôlant l'importance du lissage est également délicate et dépend de l'application. Un état de l'art des méthodes utilisant un terme de régularisation a été réalisé en 2006 par Weickert et al. [260].

Modèle TVL1 Parmi ces dernières, les modèles variationnels utilisant un terme de données L^1 et une régularisation par variation totale (TV) produisent des résultats précis et conservent les arêtes. Chacun des deux termes peut être amélioré : les contraintes de conservation de données [186] et le terme de régularisation. Werlberger et al. [262] proposent de remplacer le terme de régularisation TV isotrope par un terme anisotrope, guidé par les contours de l'image et qui utilise une fonction de pénalisation de type Huber [110]. L'énergie à minimiser du modèle TVL^1 (après linéarisation du terme de données) peut s'écrire ainsi :

$$E(\mathbf{u},t) = \int_{x,y} \mid \mathbf{u} \cdot \nabla I(x,y,t) + I_t(x,y,t) \mid + \lambda \left(\mid \nabla u_1 \mid + \mid \nabla u_2 \mid \right) dxdy \tag{A.8}$$

avec $\mathbf{u} = (u_1, u_2)$. L'utilisation de la norme L^1 favorise les solutions constantes par morceaux dans les régions faiblement texturées et conduit à un effet "en marches d'escalier". La fonction de régularisation Huber permet de réduire cet effet. Le choix d'un modèle anisotrope guidé par les gradients de l'image permet lui de réduire l'effet de lissage sur les contours. Avec de plus l'introduction de variables auxiliaires $\mathbf{v} = (v_1, v_2)$ afin de faciliter la minimisation de la fonctionnelle, le nouveau modèle peut s'écrire comme la minimisation sur $\mathbf{u}$ et $\mathbf{v}$ de :

$$E(\mathbf{u},\mathbf{v},t) = \int_{x,y} \mid \mathbf{v} \cdot \nabla I(x,y,t) + I_t(x,y,t) \mid + \lambda \sum_{d=1}^{2} [\mid \mathbf{q}_d \mid_\epsilon + \frac{1}{2\theta}(u_d - v_d)^2] dxdy \tag{A.9}$$

avec θ une constante positive faible, $\mathbf{q}_d = D^{1/2} \nabla u_d$ et $\mid \mathbf{q}_d \mid_\epsilon = \frac{|\mathbf{q}_d|^2}{2\epsilon}$ si $\mid \mathbf{q}_d \mid \leq \epsilon$, $= \mid \mathbf{q}_d \mid - \frac{\epsilon}{2}$ sinon. D est un tenseur de diffusion symétrique et défini positif faisant intervenir le gradient de l'image.

- Le terme $\int_{x,y} \sum_{d=1}^{2} \frac{1}{2\theta}(u_d - v_d)^2 dxdy$ est un terme de couplage qui assure que le vecteur de variables auxiliaires $\mathbf{v}$ sera proche de la solution recherchée $\mathbf{u}$.
- Le terme $\sum_{d=1}^{2} \mid \mathbf{q}_d \mid_\epsilon dxdy$ correspond au terme de régularisation anisotropique ($\mid \cdot \mid_\epsilon$ étant la norme Huber et $\mathbf{q}_d$ la fonction anisotrope du gradient de l'image.
- Le dernier terme $\int_{x,y} \mid \mathbf{v} \cdot \nabla I(x,y,t) + I_t(x,y,t) \mid dxdy$ est le terme de données correspondant à la contrainte de flot optique ou d'intensité constante, ici avec une norme L^1.

La dualisation permet de transformer le problème en un couple de minimisations convexes alternées sur $\mathbf{u}$ et $\mathbf{v}$.

Modélisation de mouvement par couches La modélisation du mouvement image comme superposition de différentes couches permet de contourner les difficultés inhérentes aux méthodes de régression par région. En effet, ces dernières font intervenir un compromis entre le nombre de contraintes pour l'estimation des paramètres d'une part, la difficulté de représenter une région par un unique modèle paramétrique : pour des tailles de support plus importantes, le modèle sera plus contraint mais le risque de mouvements multiples mal représentés par un modèle unique, augmente également. Cette difficulté est d'autant plus marquée aux alentours des frontières d'occultations qui présentent généralement des différences de profondeur non négligeables. La modélisation explicite de la scène en différentes couches de mouvements indépendants peut être exprimée au travers de différents modèles [23]. Citons notamment les modèles de mélange probabilistes [114], en utilisant l'algorithme EM (espérance - maximisation) [65] pour l'estimation des paramètres.

A.4 Implémentation

Un certain nombre d'approches algorithmiques sont détaillées dans [15], notamment les algorithmes d'optimisation continue ou discrète, le choix des estimateurs ou encore l'emploi de stratégies coarse-to-fine. Nous nous efforçons ici de donner un aperçu des ces approches.

A.4.1 Optimisation

Lorsque l'estimation du flot optique s'écrit comme minimisation d'une énergie E_{global}, comportant ici un terme de données voire un terme d'a priori (continuité ou caractère lisse du flot), la méthode d'optimisation peut être continue ou discrète.

Optimisation continue Les deux méthodes principales d'optimisation continue sont la descente de gradient d'une part, les approches extrémales ou variationnelles d'autre part. Si $\mathbf{f}$ représente la concaténation des composantes horizontales et verticales de l'ensemble des pixels, le but de la descente de gradient est d'optimiser E_{global} par rapport à $\mathbf{f}$. L'algorithme le plus simple est la descente dans la direction de la plus forte pente [14], dans lequel les pas successifs de $\mathbf{f}$ sont pris dans la direction de l'opposé du gradient $-\frac{\partial E_{global}}{\partial \mathbf{f}}$. Le choix du pas de descente n'est pas unique : il peut être adaptatif, la norme du pas étant diminuée ou augmentée si l'énergie augmente ou diminue respectivement ; il peut aussi dépendre des dérivées de l'énergie relativement à $\mathbf{f}$ [30]. Les algorithmes de descente de gradient ne modélisent pas les couplages entre inconnues et convergent donc lentement. L'utilisation d'un modèle de second ordre de ces couplages par la matrice hessienne $\frac{\partial^2 E_{global}}{\partial f_i \partial f_j}$ permet d'accélérer la convergence. Plusieurs algorithmes tels que la méthode de Newton, Gauss-Newton, Levenberg-Marquardt [14], en tirent parti. Toutefois, ils nécessitent l'estimation et l'inversion de la matrice hessienne et ne sont donc applicables qu'à des systèmes réduits. Ils recherchent alors un unique vecteur de flot [156, 135] ou des modèles paramétriques [25] par blocs.

Les approches variationnelles supposent que l'énergie peut être écrite sous la forme :

$$E_{global} = \int\int E\left(u(x,y), v(x,y), x, y, u_x, u_y, v_x, v_y\right) dxdy \tag{A.10}$$

où u_x,u_y,v_x et v_y représentent les dérivées partielles de u et v par rapport à x et y. u et v sont ainsi considérées comme des fonctions de x et y plutôt que comme paramètres inconnus. Plusieurs fonctions satisfont cette formulation, dont [104, 47, 45, 175, 284]. L'interprétation de (A.10) conduit aux équations d'Euler-Lagrange. Dans le cas général, ces équations ne sont pas linéaires et sont résolues par une méthode itérative, par exemple par linéarisation de Taylor.
Une autre approche [239] encore consiste à découpler les termes de données et d'a priori en faisant intervenir deux ensembles de flots intermédiaires $u_{donnees}, v_{donnees}$ et $u_{apriori}, v_{apriori}$ liés par un terme de distance :

$$\begin{aligned} E_{global} = \quad & E_{donnees}(u_{donnees}, v_{donnees}) + \lambda E_{apriori}(u_{apriori}, v_{apriori}) \\ & + \gamma \left(\|u_{donnees} - u_{apriori}\|^2 + \|v_{donnees} - v_{apriori}\|^2 \right) \end{aligned} \tag{A.11}$$

L'optimisation se fait en deux étapes, en considérant dans un premier temps $(u_{apriori}, v_{apriori})$ fixes et optimisant l'énergie sur $(u_{donnees}, v_{donnees})$, puis en optimisant sur $(u_{apriori}, v_{apriori})$ à $(u_{donnees}, v_{donnees})$ fixés.
Des algorithmes d'optimisation convexe continue tels que [219] sont également utilisables dans le but de calculer le flot optique.

Optimisation discrète Les méthodes d'optimisation discrète réalisent une approximation de l'espace continu des solutions par un problème simplifié. La recherche au sein de l'espace d'état peut ainsi être plus complète, au prix d'une perte de précision. Les approximations généralement effectuées limitent la capacité des algorithmes discrets à éviter les minima locaux et doivent donc être contrôlées.

Les algorithmes de fusion constituent le premier ensemble de méthodes discrètes. Le principe de ces algorithmes est de combiner les forces de différents algorithmes standards

de calcul de flot, tels que Lucas-Kanade [156] ou Horn et Schunk [104], parfois en exemplaires multiples avec variation des paramètres. L'obtention du flot final revient alors à choisir en chaque pixel le meilleur flot en ce point. Ainsi, Lempitsky et al. [139] utilisent une série d'optimisations par graph-cut binaire pour remplacer des sous-ensembles du flot estimé par l'une des solutions candidates, ce qui améliore progressivement le résultat. Trobin et al. [240] emploient également plusieurs étapes de fusion en résolvant à chaque itération un problème d'optimisation continue et en seuillant le résultat.

Les méthodes de reparamétrisation dynamique forment le second groupe de méthodes. Ils effectuent un balayage approximatif de l'espace d'états, en discrétisant à la fois l'espace et l'ensemble des flots possibles à chaque pixel. Une deuxième étape améliore ensuite la précision de l'espace d'état à partir du résultat obtenu. Black et Anandan [29] utilisent le recuit simulé sur un espace d'état adaptatif selon la forme local de la fonction d'énergie. D'autres méthodes partent d'un espace d'état à une échelle grossière avant de raffiner les résultats dans une approache coarse-to-fine. Ainsi, Glocker et al. [85] font évoluer la densité spatiale et la précision des flots selon l'incertitude de la solution à l'itération précédente. Lei et Yang [138] font dépendre l'allocation spatiale d'une hiérarchie de segmentations, avec un unique flot possible pour un segment donné à un niveau de précision donné, avec un raffinement du flot dans un voisinage de la solution trouvée au niveau de précision inférieur. Cooke [54] alterne itérativement entre l'estimation du mouvement horizontal et vertical respectivement, en considérant une composante fixée pour chacune des deux étapes à chaque itération.

Les résultats obtenus par une méthode d'optimisation discrète peuvent ensuite être encore raffinés par une optimisation continue.

A.4.2 Traitement hiérarchique multi-échelles

L'utilisation d'un cadre pyramidal multi-échelles permet de traiter les cas présentant des déplacements rapides, pour lesquels l'approximation (A.2) n'est plus valide. Un tel cadre permet de plus d'accélérer grandement les calculs. L'image est décomposée en plusieurs résolutions sous la forme de pyramide gaussienne ou laplacienne [21, 70]. Une estimation initiale du flot est calculée au niveau le plus grossier puis projetée sur un niveau de résolution plus fin avant d'être itérativement raffinée. L'estimation finale du flot est obtenue au niveau de résolution le plus fin.

A.4.3 Estimateurs robustes

L'estimation de mouvement peut être rendue plus robuste en choisissant un estimateur autre que celui des moindres carrés. En effet, ce dernier n'est optimal que lorsque les erreurs de contrainte de gradient :

$$e(\mathbf{x}) \equiv \mathbf{u} \cdot \nabla I(\mathbf{x}, t) + I_t(\mathbf{x}, t) \tag{A.12}$$

suivent une distribution gaussienne de moyenne nulle et que les erreurs correspondant à différentes contraintes sont indépendantes et identiquement distribuées (IID). Les changements d'orientation de surface, les réflections spéculaires, des ombres non constantes trahissent l'hypothèse d'illumination constante (A.1). Les variations abruptes de profondeur contredisent de plus le modèle de mouvement constant, notamment aux frontières d'occultations. Il est donc préférable de remplacer l'estimateur quadratique de (A.6) par un estimateur plus robuste limitant l'influence des contraintes présentant des erreurs significatives :

$$E(\mathbf{u}) = \sum_{\mathbf{x}} W(\mathbf{x}) \rho(e(\mathbf{x}), \sigma) \tag{A.13}$$

où σ représente le ou les paramètres de l'estimateur robuste. L'estimateur de Geman-McLure [81], $\rho(e,\sigma) = e^2/(e^2+\sigma^2)$ ou l'utilisation de la norme L^1 en donnent des exemples. Les méthodes de moyenne adaptative avec diffusion non linéaire, dont celles fondées sur l'utilisation d'une diffusivité à variation totale A.3.3 fournissent un exemple plus récent.

A.4.4 Formulations probabilistes

Au contraire des estimateurs des moindres carrés ou estimateurs robustes (A.4.3), une formulation probabiliste permet d'apporter des informations sur les intervalles de vraisemblance sur l'estimation du flot ainsi que d'incorporer de l'information a priori sur la distribution du flot : l'information sur le flot estimé à des instants précédents peut ainsi être propagée à l'instant actuel. Le choix du modèle de bruit (sur l'image ou ses dérivées) conduit à différents estimateurs, tels que les moindres carrés totaux (TLS) [258] pour un bruit additif, isotrope et IID, ou des modèles plus complexes [173].

A.4.5 Illumination et couleur

L'hypothèse d'illumination constante peut ne pas être vérifiée, lors de changements météo, d'heure ou d'éclairage. Il s'agit d'un phénomène typique des séquences vidéo en extérieur, aériennes ou non, qui doit être maîtrisé. Il faut alors pouvoir suivre d'autres primitives que l'intensité, moins ou pas sensibles aux changements d'illumination et / ou de contraste. Il peut s'agir d'arêtes de l'image, d'orientation de textures. Des dérivées d'ordre supérieur (ordre 2) peuvent être utilisées [171, 242] mais sont plus bruitées que l'intensité et nécessitent un mouvement sans déformation de premier ordre telle qu'une rotation. Une autre piste consiste à étudier la phase dans le domaine spatio-temporel, cf. section A.3.1.

Dans le cadre de variations de luminosité importantes, il peut être préférable de modéliser directement ces variations. Les modèles utilisés, génériques ou inspirés de la physique [96], peuvent être spécifiques à des objets ou des variations d'éclairage [92] et de pose [31].

La prise en compte des différents canaux d'une image (par exemple rouge, vert et bleu pour une image couleur) nécessite d'adapter le terme de fidélité aux données correspondant à (A.2) fondé sur l'intensité seule. Une approche simple est de traduire cette équation pour chaque canal avant d'effectuer une sommation sur les canaux [177, 158, 87]. Il est aussi possible de considérer des espaces couleur différents tels que HSV et de pondérer différemment chaque canal [284].

A.4.6 Cohérence temporelle

L'inclusion de contraintes de cohérence temporelle peut permettre d'améliorer la qualité du flot obtenu et de supprimer une partie des outliers. Il peut s'agir d'un calcul incrémental du flot [29] ou de contrainte de symétrie temporelle [261]. La régularisation temporelle peut toutefois être contre-productive dans le cas de mouvements complexes et présentant une grande amplitude (ce qui est le cas d'une grande partie des séquences d'évaluation Middlebury, cf. A.5)

A.5 Évaluation

A.5.1 Méthodologie

Afin d'évaluer la précision ainsi que les points d'achoppement des différentes méthodes d'estimation de flot optique, plusieurs éléments sont nécessaires, détaillés dans [15] :

– une base de séquences vidéo complexes et présentant des difficultés complémentaires (occultations, mouvements rapides, variations de profondeur, faible contraste...) avec les vérités terrain correspondantes. L'ensemble de données Middlebury [15] et le site web associé `http://vision.middlebury.edu/flow/` est l'une des références les plus récentes.

– des mesures de précision du flot : vectorielle ("endpoint") ou angulaire [19] :

$$AE(\mathbf{u}, \mathbf{u_{gt}}) = \arccos\left(\frac{\mathbf{u} \cdot \mathbf{u_{gt}}}{\mid \mathbf{u} \mid\mid \mathbf{u_{gt}} \mid}\right) \tag{A.14}$$

où $\mathbf{u}$ et $\mathbf{u_{gt}}$ représentent respectivement le flot à évaluer et le flot "vérité terrain"

– des métriques d'agrégation des mesures précédentes : moyennes et écarts-types sur l'ensemble de l'image, métriques par quantiles...

A.5.2 Comparaison

Les résultats récents de l'évaluation sur les données Middlebury [15] montrent qu'aucune méthode ne montre d'excellentes performances sur l'ensemble des cas A.5.1. L'ensemble des méthodes listées dans le tableau A.1 est loin d'être exhaustif en regard de l'ensemble de l'existant. La sélection a été guidée à la fois par un choix chronologique, des premières méthodes telles que Horn et Schunk [104] jusqu'aux plus récentes dont l'anisotrope Huber-L^1 [262] ou encore Zimmer et al. [284], et par un effort de représentation de différents choix selon les critères exposés (raffinements, choix des normes pour le terme de données et de régularisation, optimisation), inspirés par [15].

Le spectre de toutes les difficultés vidéo n'est de plus pas couvert : types de matériaux avec réflectances diverses, changements de luminosité, effets atmosphériques et transparence, durée plus importante...

Complexité Dans l'ensemble, les algorithmes les plus performants utilisent plus de raffinements sur les termes de données et d'a priori : pondération spatiale, anisotropie, robustesse vis-à-vis de la luminosité par normalisation du terme de données ou utilisation de primitives intermédiaires. Ces raffinements permettent de mieux appréhender la structure locale de l'image et du déplacement associé, qu'il s'agisse de la richesse des textures, de la conservation des gradients ou de tolérance par rapport à des changements de luminosité (éclairage, nuages, ombres...)

Choix de la fonction de pénalisation La norme L^1 est beaucoup utilisée, notamment pour le terme de données. Cela est moins le cas pour le terme d'a priori, pour lequel d'autres fonctions plus tronquées sont également utilisées.

Optimisation continue Les algorithmes de descente de gradient obtiennent des scores relativement mauvais. Les approches variationnelles sont mieux réparties sur l'ensemble des résultats mais utilisent également des énergies plus complexes.

Optimisation discrète Les méthodes ne sont pas très performantes mais les énergies utilisées sont simples, et les résultats seraient possiblement meilleurs avec des énergies plus sophistiquées. Une optimisation continue subséquente est nécessaire pour améliorer la précision des résultats vectoriels.

TAB. A.1 – *Classification d'algorithmes de flot minimisant une énergie selon les termes de données et d'a priori, le type d'optimisation utilisés*

Algorithme	Terme de données				Terme d'a priori				Optimisation		Autres	
	Norme L1	Autre fonction de pénalité robuste	Gradient ou autres primitives	Robustesse à la luminosité	Norme L1 / Variation totale	Autre fonction de pénalité robuste	Pondération spatiale	Pondération anisotrope	Continue	Discrète	Gestion de l'occultation	Images couleur
Zimmer et al. (09) [284]	X		X	X		X	X	X	X			X
Anis. Huber-L^1 (09) [262]	X			X		X	X	X	X			
Lei et Yang (09) [138]				X	X		X		X	X	X	X
Trobin et al. (08) [240]	X				X					X		X
F-TVL1 (08) [259]	X				X				X			
Fusion (08) [139]		X		X		X	X		X	X		X
MRF dynamique (08) [85]	X				X					X		
Seitz et Baker (09) [219]	X			X	X		X		X			X
Graph Cuts (08)[54]	X				X					X		X
Papenberg et al. (06) [186]		X	X			X	X		X			
Fleet et al. (00) [75]		X							X			
Bergen et al. (92) [25]									X			
Battiti et al. (91) [21]									X			
Black et Anandan (91) [29]		X				X			X			
Horn et Schunk (81) [104]									X			

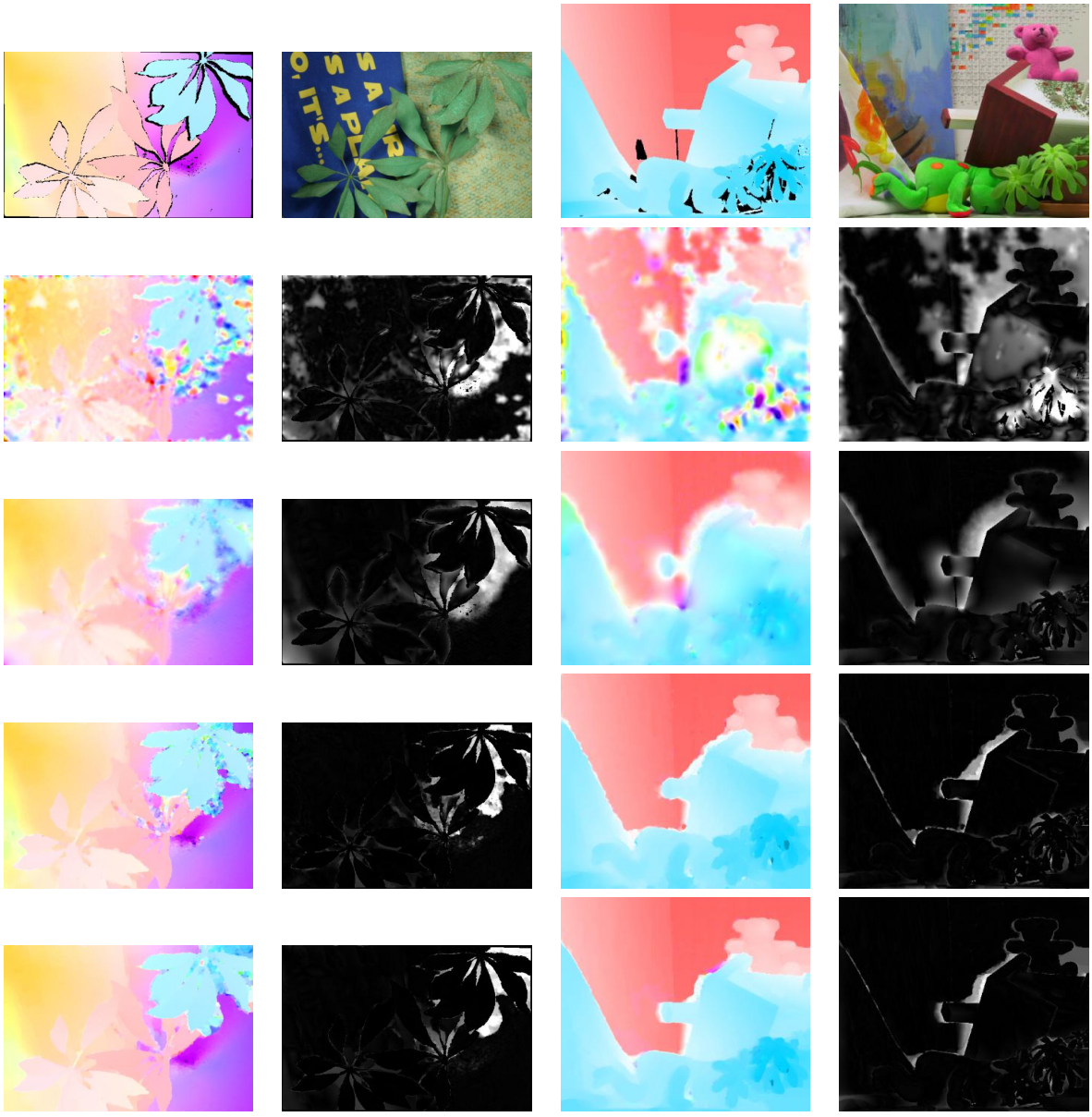


FIG. A.2 – *1ère ligne, flot (vérité terrain) et 1ère image : (a) et (b) Schefflera, (c) et (d) Teddy ; 2ème, 3ème, 4ème et 5ème ligne : flot et erreur vectorielle correspondante, (a) et (b) Schefflera, (c) et (d) Teddy. 2ème ligne : Lucas-Kanade [156] avec implémentation pyramidale ; 3ème ligne : Horn et Schunk [104] ; 4ème ligne : Flot TVL1 [259] ; 5ème ligne : Huber L1 anisotrope [262]*

Autres composantes L'influence de certaines composantes est difficile à estimer en raison du faible nombre de méthodes utilisant ces derniers : apprentissage, prise en compte des occultations, ou calcul à partir de plus de 2 frames et l'utilisation de l'information couleur qui ne semble pas apporter d'amélioration siginificative sur la base des résultats disponibles.

A.5.3 Conclusion

L'évaluation d'algorithmes de flot optique est une tâche complexe qui nécessite de tenir compte de plusieurs considérations. Un ensemble de données de référence est nécessaire afin d'effectuer une comparaison pertinente. Les algorithmes sont susceptibles d'être particulièrement adaptés à tout ou partie des données toutefois. Ainsi, des séquences telles que Yosemite de Lynn Quam [100] privilégient notamment un lissage temporel qui se révèle dégrader les performances pour d'autres séquences moins régulières. La rigidité des séquences influe également sur les résultats selon les méthodes utilisées. La présence de discontinuités complexes ainsi que de mouvements de large amplitude influent également sur les performances des méthodes. Il n'existe pas d'algorithme particulièrement performant sur l'ensemble des cas présentés. L'évaluation elle-même n'est pas immédiate et selon le contexte, différentes métriques peuvent être utilisées. Ainsi, des algorithmes peuvent fournir une estimation de moindre qualité selon le critère de l'erreur angulaire moyenne, mais présenter une interpolation proche de l'image intermédiaire réelle.

Qui plus est, ces cas ne couvrent pas l'ensemble des séquences possibles. De nouveaux matériaux présentant différents types de réflection ou transparence, l'ajout d'effets atmosphériques ou de changements importants de luminosité constituent autant de nouvelles conditions de difficulté, non seulement pour les algorithmes de flot optique mais aussi pour l'établissement d'une vérité terrain. L'utilisation de séquences synthétiques semble pertinente dans l'objectif de contrôler plus aisément le niveau des différents effets ou sources de complexité. Afin de couvrir au mieux l'éventail des conditions réelles possibles, il importe de créer un ensemble de séquences de référence important et diversifié, faiblement corrélées. Cela permet d'éviter de ne concentrer les efforts que sur un unique point (par exemple les déplacements de grande amplitude ou les occultations) voire suivant les approches une optimmisation manuelle afin d'améliorer les performances sur une unique séquence (ou groupe de séquences voisines). Une telle approche facilite la compréhension des facteurs limitant les performances des algorithmes selon les difficultés particulières associées aux séquences afin d'orienter les recherches sur les points les plus problématiques.

Dans le cadre d'applications temps réel, la complexité algorithmique ainsi que la facilité de parallélisation sont d'autres éléments à prendre en compte. Les algorithmes les plus performants utilisent ainsi souvent une pondération spatiale et anisotrope des termes de données ou régularisation et incorporent une certaine robustesse aux changements de luminosité par le biais de normalisation du terme de données ou prise en compte des différents canaux de l'image. Ces raffinements divers pèsent sur le temps de calcul et peuvent être limitants selon le cadre d'application, en présence d'un volume important de données par exemple.

L'ajustement des différents paramètres utilisés pour un algorithme est un autre problème encore. Par exemple, le compromis entre un flot lisse mais qui supprime potentiellement de très fins détails et une recherche de précision fournissant un flot bruité est difficile à apprécier. Les paramètres associés peuvent être réglés manuellement si le temps de calcul est proche du temps réel ou ajustés automatiquement par minimisation d'une énergie, au détriment d'une flexibilité interactive.

Apprentissage automatique (Machine learning)

B

B.1 Introduction

Les tâches de détection, reconnaissance et identification visent des applications diverses, qu'il s'agisse de classification, recherche d'images par leur contenu, de pistage ou encore de vidéosurveillance. Plusieurs approches visent à effectuer ces tâches de manière automatique afin d'éviter une prise en charge manuelle fastidieuse et lente. A partir de données d'apprentissage, exemples positifs (contenant des instances de classes à reconnaître) ou négatifs (fond et objets n'appartenant pas à la classe recherchée), ces approches doivent pouvoir reconnaître de nouvelles instances des classes apprises sur de nouvelles données, les données test. Les données d'apprentissage ne représentent qu'une faible partie de l'ensemble des cas possibles (négatifs ou positifs), l'apprentissage automatique doit donc inférer un modèle robuste capable de performances élevées sur des cas d'application beaucoup plus nombreux et variés.

Les approches d'apprentissage automatique sont multiples. Il est possible de les classer selon le type de sortie voulu :

- l'apprentissage par renforcement consiste à modéliser le comportement d'un agent de façon à maximiser une récompense globale en fonction de ses actions au sein d'un environnement donné.
- la transduction vise à prédire des sorties de données d'entrée à partir d'un ensemble de données d'entrée et de sortie fournis pour apprentissage.
- les approches de type "learning to learn", dans lesquelles il s'agit d'apprendre le biais même d'induction tel que les hypothèses de marge maximale (utilisées dans les SVM), ou l'appartenance à une même classe de voisins (dans les K-plus proches voisins).

La disponibilité d'étiquettes associées aux données d'apprentissage est un autre critère permettant de séparer les différentes approches, supervisées, semi-supervisées ou non supervisées :

- lorsque chaque donnée est associée à une étiquette, il s'agit d'apprentissage supervisé. L'algorithme doit alors fournir une fonction capable d'associer à de nouvelles données les étiquettes correctes.
- l'apprentissage non supervisé ne dispose que de données sans étiquettes. Le but est alors plutôt de modéliser cet ensemble de données, par exemple en les regroupant en sous-ensembles cohérents ou "clusters".
- l'apprentissage semi-supervisé produit une fonction ou un classificateur à partir de données partiellement étiquetées.

B.2 Apprentissage supervisé

Dans le contexte d'interprétation de séquences vidéo aériennes, les approches d'apprentissage supervisé voire semi-supervisé sont pertinentes. Il importe en effet d'obtenir en sortie d'algorithme des résultats interprétables sur le plan sémantique, à partir de données d'apprentissage étiquetées. Ces données peuvent être simplement des instances d'objets spécifiques (bâtiments, véhicules, piétons, animaux...) dans un cadre de détection, reconnaissance ou identification d'objet. Une interprétation sémantique humaine de la scène tirera alors parti des objets automatiquement détectés. Mais l'apprentissage automatique peut également concerner, sur les données complexes que sont les séquences vidéo, des actions voire des scénarios mêlant actions, points de repère et interactions entre différentes entités.

Les principales méthodes d'apprentissage supervisé regroupent notamment l'apprentissage par arbre de décision, les réseaux de neurones artificiels, les machines à vecteurs de support (SVM) ainsi que les réseaux bayésiens. Il faut rajouter à cela les différentes approches d'ensemble qui combinent ou agrègent des classifieurs souvent primitifs afin de fournir des classifieurs plus élaborés. Les plus connus sont le "bagging", le "boosting" et les "random forests" ou forêts aléatoires.

- les arbres de décision également connus sous le nom de CART (Classification And Regression Trees) [42] regroupent les arbres de classification, associant des étiquettes de classes aux données d'entrée ; et de régression, pour lesquels la sortie peut être continnue. Un arbre de décision choisit à chaque sommet une variable et ,lorsque cette variable est continue, un seuil de coupure associé, en maximisant un critère donné. Le critère utilisé caractérise la pureté (ou le gain en pureté) lors du passage du sommet à segmenter vers les feuilles ainsi produites. Les critères les plus utilisés sont l'entropie de Shannon et le coefficient de Gini ainsi que leurs variantes.
  Les avantages des arbres de décision sont multiples. Outre un principe simple, ils sont capables d'intégrer simultanément des données numériques et de catégorie ou classe sans normalisation préalable. La logique booléenne permet d'interpréter facilement sous la forme d'arbre les résultats par rapport aux différentes dimensions des données. La création de l'arbre à partir d'ensembles importants de données est enfin rapide et permet de sélectionner les variables discriminantes en présence d'un nombre important de variables.
  Toutefois, il est souvent nécessaire d'élaguer les arbres de décision obtenus afin d'éviter un effet de surapprentissage. Certains concepts tels que le ou exclusif ou le multiplexage sont également difficiles à traduire par le biais d'arbres de décision. Des structures multidimensionnelles complexes correspondant à une classe d'objet dans l'espace de variables produisent ainsi des arbres particulièrement complexes.
  Les arbres de décision peuvent être combinés au sein d'approches telles que forêts aléatoires [41] ou arbres renforcés [95]. Les arbres à décision alternante "ADTrees" [77] constituent une généralisation des arbres de décision en structurant l'ensemble des hypothèses ou classifieurs faibles. Une branche "fille" d'un nœud dépend donc de l'hypothèse correspondante.

- Les méthodes d'ensemble construisent un ensemble fini de modèles avant de les agréger en un unique modèle. Dietterich montre leur utilité dans différents problèmes d'apprentissage [66]. Ces méthodes permettent en effet d'améliorer significativement la précision de classifieurs originaux dans le cadre d'un apprentissage supervisé. Les approches les plus populaires sont le "boosting", le "bagging" et les "random forests".

  Le "boosting" consiste généralement à apprendre de manière itérative un ensemble de classifieurs faibles sur un jeu de données d'apprentissage et les combiner afin

de former un classifieur fort présentant de meilleures performances de classification. Les classifieurs faibles sont pondérés avant ajout selon leur précision. Les échantillons d'apprentissage sont également repondérés à chaque itération : ceux qui ont été mal classés sont associés à un poids plus important et inversement. Lors de la phase de test, l'ensemble des classifieurs faibles retenus sont appliqués aux nouvelles données et leurs résultats pondérés selon les poids obtenus lors de la phase d'apprentissage au sein du classifieur fort final.
Les principales différences entre les divers algorithmes de boosting concernent la méthode de pondération des échantillons d'apprentissage et des classifieurs faibles. Le premier algorithme, Adaboost [78], reste très utilisé. Il est sensible à des données bruitées et outliers mais montre une certaine robustesse au surapprentissage. Il lui est impossible de gérer directement les problèmes à classes multiples, il est alors nécessaire de découper le problème en classifications "un contre un" ou "un contre tous". De nombreuses variantes de boosting existent, parmi lesquelles LogitBoost et GentleBoost [79], Linear Programming Boosting (LPBoost) [64], Real Adaboost [142] ou encore le Quadratic Boosting [192].
Le "online boosting" [150, 88] permet d'intégrer l'information contenue dans de nouveaux échantillons en adaptant en "temps réel" le classifieur. D'autres primitives peuvent ainsi être incluses au sein du classifieur final afin de mieux percevoir des changements d'apparence d'objets ou de scènes.

Le "bagging", ou "bootstrap aggregating", a été proposé par Breiman en 1994 [40]. Chaque modèle participe avec un vote de poids identique. Chaque modèle de l'ensemble est entraîné à partir d'un sous-ensemble des données d'apprentissage obtenu par bootstrap (échantillonnage avec répétitions d'échantillons possibles) afin de rendre compte de la variance du modèle. Oza présente une extension du bagging et boosting à l'apprentissage en ligne [182].

Les "random forests" ou forêts aléatoires [41] conjuguent les arbres de décision aléatoires avec le bagging afin d'obtenir une précision de classification élevée. Chaque arbre est entraîné sur un sous-ensemble obtenu par bootstrap des données d'apprentissage. La sortie de l'algorithme correspond au mode de l'ensemble des sorties des différents arbres. Ces forêts présentent de nombreux avantages, dont une très grande précision, un apprentissage rapide, peut traiter de grandes bases de données et un nombre élevé de variables. Elles peuvent également être adaptées au cas de variables non étiquetées [222]. En revanche, elles peuvent présenter des problèmes de surapprentissage.

– Les réseaux de neurones artificiels s'inspirent de la structure des réseaux de neurones biologiques. Ils réalisent les calculs par le biais de groupes interconnectés de neurones artificiels. Ils permettent de modéliser des relations complexes non linéaires entre entrées et sorties, de découvrir des motifs reliant les données ou encore d'appréhender la statistique d'une distribution jointe inconnue à partir de variables observées.
Chaque neurone est associé à une fonction de transfert pouvant accepter plusieurs entrées et renvoyant une sortie selon des règles précises telles que sommation ou seuillage par exemple. Des pondérations, appelées "poids synaptiques" sont attribuées à chaque neurone et modulent l'efficacité ou prise en compte de l'information transmise par un neurone à d'autres neurones. Afin de présenter plus de flexibilité, il est nécessaire que ces poids puissent évoluer en fonction de nouvelles données. Le modèle du perceptron [212] permet de modifier la valeur des poids en fonction des activités des neurones associés en suivant la règle de Hebb. Les premiers réseaux de neurones artificiels étaient limités à la résolution de problèmes linéaires, mais

l'apparition du perceptron multi-couches en [215] a permis de traiter également des problèmes non linéaires.
Les réseaux de neurones artificiels sont utilisés dans le cadre de plusieurs applications telles que la reconnaissance de motif, approximation de fonction inconnue ou connue mais d'évaluation exacte complexe... Au contraire des arbres de décision, ils ne fournissent pas de justification aisément interprétable et se comportent plus tels des "boîtes noires". Ils nécessitent également des bases de données dont l'importance croît avec la complexité du problème et sont plus adaptés à certains types de problèmes (notamment les problèmes de classification sur domaines convexes). Un état de l'art des applications des réseaux est réalisé dans [184], et le cas plus précis de la reconnaissance de motif est détaillé dans [209] ou [28].

– Les machines à vecteurs de support (SVM) sont une généralisation des classifieurs linéaires. Développées dans les années 1990 et découlant de la théorie statistique de l'apprentissage de Vapnik-Chervonenkis [244], ils présentent les avantages de pouvoir travailler avec des données de grandes dimensions, d'utiliser peu d'hyper paramètres (valeurs fixées avant l'apprentissage pour régler la sensibilité de l'algorithme), de fournir de bons résultats en pratique.
Les SVM reposent sur deux idées principales, qui permettent de traiter des problèmes de discrimination non-linéaire, et de reformuler le problème de classement comme un problème d'optimisation quadratique. La première idée est la notion de marge maximale. La marge est la distance entre la frontière de séparation et les échantillons les plus proches, appelés vecteurs supports. Dans les SVM, la frontière de séparation est choisie comme celle qui maximise la marge [99]. Le problème revient à trouver la frontière optimale à partir d'un ensemble d'apprentissage. Il faut pour cela redéfinir le problème en problème d'optimisation quadratique, résoluble par des algorithmes connus.
Les données ne sont souvent pas linéairement séparables. Il est alors nécessaire de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension (possiblement de dimension infinie), dans lequel il sera possible de séparer linéairement les données. Ceci est réalisé grâce à une fonction noyau [218] respectant certaines conditions mathématiques et ne nécessitant pas la connaissance explicite de la transformation à appliquer pour le changement d'espace. Les fonctions noyau permettent de transformer un produit scalaire coûteux dans un espace de grande dimension, en une simple évaluation ponctuelle d'une fonction.

– Les réseaux bayésiens sont des modèles graphiques probabilistes qui permettent de représenter un ensemble de variables aléatoires ainsi que leurs relations conditionnelles par le biais d'un graphe orienté non cyclique [191]. Un réseau bayésien comporte d'une part un graphe, qui représente la structure du modèle, d'autre part les tables de probabilités des variables conditionnellement aux autres variables et correspondent aux arêtes du graphe. Le graphe et les tables de probabilités peuvent être définis *a priori* ou appris.
Les réseaux peuvent être utilisés pour calculer la distribution *a posteriori* de variables à partir de la connaissance d'autres variables, il s'agit de l'inférence probabiliste [90]. Les méthodes d'inférence exacte étant de complexité exponentielle en fonction de la largeur arborescente du graphe [32] et donc coûteuses, il existe des méthodes approximatives dont par exemple la "loopy belief propagation" [170], la simulation "Monte-Carlo Markov Chain" (MCMC) [84] ou encore l' "importance sampling" [86].
Les champs de Markov [123] ressemblent aux réseaux bayésiens dans le mode de représentation des dépendances. Il s'agit de graphes non orientés qui ne peuvent donc représenter des dépendances induites au contraire des réseaux bayésiens. En revanche, ils peuvent traiter des dépendances cycliques. La section 4.1 décrit plus

précisément les modèles graphiques ainsi que diverses extensions des champs de Markov.

B.3 Descripteurs

Les descripteurs utilisés dépendent des caractéristiques des objets ou actions à reconnaître ou classer. Ces descripteurs doivent être robustes aux variations d'apparence de la classe tout en conservant un pouvoir séparateur suffisant par rapport au fond. Parmi les différents descripteurs, un premier sous-ensemble est fondé sur l'utilisation de primitives de bas niveau : descripteurs de Haar [247], histogrammes d'orientation [60] dont dérivent les SIFT ("Scale-invariant feature transform") [154], patrons binaires locaux pour la classification de textures [178]. Ces primitives sont rapides à calculer de par leur simplicité et sont souvent combinées par renforcement ou boosting [150]. Un second groupe de descripteurs englobe les dictionnaires de patchs [237, 5]. Les objets sont décrits à partir de plusieurs patchs représentatifs qui seront ensuite recherchés au sein des données test avant décision (par majorité, cohérence géométrique des patchs...) Le choix des patchs et la construction des dictionnaires n'est toutefois pas évident [118]. Yilmaz et al. présentent dans [276] un inventaire de descripteurs géométriques et de méthodes de pistage d'objets associées.

Bibliographie

[NDT] Code et manuel de calcul tensoriel en n dimensions. `http://iris.usc.edu/people/medioni/ndtensorvoting.html`. (Cité page 140.)

[Tre] Trecvid. `http://www-nlpir.nist.gov/projects/trecvid/`. (Cité pages 12, 44 et 68.)

[1] (2010). Ovire. `http://www.daedalus.gr/prdinfo1.html`. (Cité page 10.)

[2] Abdollahian, G. and Delp, E. (2007a). Analysis of unstructured video based on camera motion. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6506, page 19. (Cité page 45.)

[3] Abdollahian, G. and Delp, E. (2007b). Finding regions of interest in home videos based on camera motion. In *ICIP*, volume 4, pages IV–545. IEEE. (Cité page 45.)

[4] Adelson, E. and Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2) :284–299. (Cité page 155.)

[5] Agarwal, A. and Triggs, B. (2006). Hyperfeatures–multilevel local coding for visual recognition. *Computer Vision–ECCV 2006*, pages 30–43. (Cité page 169.)

[6] Ali, S., Reilly, V., and Shah, M. (2007). Motion and appearance contexts for tracking and re-acquiring targets in aerial videos. In *Computer Vision and Pattern Recognition*, volume 2. Citeseer. (Cité page 114.)

[7] Almeida, M. and Almeida, L. (2008). Blind deblurring of natural images. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1261–1264. IEEE. (Cité page 22.)

[8] Anandan, P. (1989). A computational framework and an algorithm for the measurement of visual motion. *IJCV*, 2(3) :283–310. (Cité page 155.)

[9] Aoki, K. (2009). Shot-Change Detections and Shot-Change Effects Recognition based on Motions and Their Estimation Reliabilities. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, volume 4, pages 194–198. IEEE. (Cité page 44.)

[10] Aubert, O. and Prié, Y. (2007). Advene : an open-source framework for integrating and visualising audiovisual metadata. In *Proceedings of the 15th international conference on Multimedia*, pages 1005–1008. ACM. (Cité page 10.)

[11] Avgerinakis, K., Briassouli, A., and Kompatsiaris, I. (2010). Real time illumination invariant motion change detection. In *Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams*, pages 75–80. ACM. (Cité page 44.)

[12] Badrinarayanan, V., Perez, P., Le Clerc, F., Oisel, L., et al. (2007). Probabilistic color and adaptive multi-feature tracking with dynamically switched priority between cues. In *In Proc. ICCV, Rio de Janeiro*. Citeseer. (Cité page 17.)

[13] Bain, M. and Sammut, C. (1999). A Framework for Behavioural Cloning. In *Machine Intelligence 15, Intelligent Agents [St. Catherine's College, Oxford, July 1995]*, pages 103–129. Oxford University. (Cité page 10.)

[14] Baker, S. and Matthews, I. (2004). Lucas-kanade 20 years on : A unifying framework. *IJCV*, 56(3) :221–255. (Cité page 158.)

[15] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., and Szeliski, R. (2011). A database and evaluation methodology for optical flow. *IJCV*, 92(1) :1–31. (Cité pages 68, 154, 157, 160 et 161.)

[16] Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., and Serra, G. (2011). Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, pages 1–24. (Cité page 16.)

[17] Bar, L., Berkels, B., Rumpf, M., and Sapiro, G. (2007). A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *ICCV*, pages 1–8. IEEE. (Cité page 24.)

[18] Bar-Shalom, Y. and Fortmann, T. (1988). *Tracking and data association*, volume 179. Academic Pr. (Cité page 17.)

[19] Barron, J., Fleet, D., and Beauchemin, S. (1994). Performance of optical flow techniques. *IJCV*, 12(1) :43–77. (Cité page 161.)

[20] Bastan, M., Cam, H., Gudukbay, U., and Ulusoy, O. (2010). Bilvideo-7 : An mpeg-7-compatible video indexing and retrieval system. *IEEE MultiMedia*, 17(3) :62–73. (Cité page 10.)

[21] Battiti, R., Amaldi, E., and Koch, C. (1991). Computing optical flow across multiple scales : an adaptive coarse-to-fine strategy. *IJCV*, 6(2) :133–145. (Cité pages 159 et 162.)

[22] Bazzani, L., Bloisi, D., and Murino, V. (2009). A comparison of multi-hypothesis kalman filter and particle filter for multi-target tracking. In *Performance Evaluation of Tracking and Surveillance workshop at CVPR 2009*, pages 47–54, Miami, Florida. (Cité page 72.)

[23] Beauchemin, S. and Barron, J. (1995). The computation of optical flow. *ACM Computing Surveys (CSUR)*, 27(3) :433–466. (Cité page 157.)

[24] Benezeth, Y., Jodoin, P., Emile, B., Laurent, H., and Rosenberger, C. (2008). Review and evaluation of commonly-implemented background subtraction algorithms. In *ICPR*, pages 1–4. IEEE. (Cité page 66.)

[25] Bergen, J., Anandan, P., Hanna, K., and Hingorani, R. (1992). Hierarchical model-based motion estimation. In *Computer Vision ?ECCV'92*, pages 237–252. Springer. (Cité pages 2, 156, 158 et 162.)

[26] Bescós, J. (2004). Real-time shot change detection over online MPEG-2 video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(4) :475–484. (Cité page 44.)

[27] Bhatti, A. and Nahavandi, S. (2009). Wavelets/multiwavelets in stereo correspondence estimation : a comparative study. In *Digital Image Computing : Techniques and Application Conference*. (Cité page 71.)

[28] Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford university press. (Cité page 168.)

[29] Black, M. and Anandan, P. (1991). Robust dynamic motion estimation over time. In *CVPR*, pages 296–302. (Cité pages 159, 160 et 162.)

[30] Black, M. and Anandan, P. (1996). The robust estimation of multiple motions : Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1) :75–104. (Cité page 158.)

[31] Black, M. and Jepson, A. (1998). Eigentracking : Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, 26(1) :63–84. (Cité page 160.)

[32] Bodlaender, H. (1997). Treewidth : Algorithmic techniques and results. *Mathematical Foundations of Computer Science 1997*, pages 19–36. (Cité page 168.)

[33] Boiman, O. and Irani, M. (2005). Detecting irregularities in images and in video. In *ICCV*, volume 1, pages 462–469. IEEE. (Cité page 14.)

[34] Bonin-Font, F., Ortiz, A., and Oliver, G. (2008). Visual navigation for mobile robots : a survey. *Journal of Intelligent and Robotic Systems*, 53(3) :263–296. (Cité page 153.)

[35] Borth, D., Ulges, A., Schulze, C., and Breuel, T. (2008). Keyframe extraction for video tagging and summarization. In *Proc. Informatiktage*, pages 45–48. (Cité page 15.)

[36] Boutellier, J. and Silvén, O. (2006). Panoramas from partially blurred video. *Advances in Machine Vision, Image Processing, and Pattern Analysis*, pages 300–307. (Cité page 39.)

[37] Bouthemy, P., Gelgon, M., and Ganansia, F. (1999). A unified approach to shot change detection and camera motion characterization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(7) :1030–1044. (Cité page 44.)

[38] Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9) :1124–1137. (Cité page 112.)

[39] Brand, M. and Kettnaker, V. (2000). Discovery and segmentation of activities in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8) :844–851. (Cité page 13.)

[40] Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2) :123–140. (Cité page 167.)

[41] Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32. (Cité pages 166 et 167.)

[42] Breiman, L., Friedman, J., Olshen, R., Stone, C., Breiman, L., Hoeffding, W., Serfling, R., Friedman, J., Hall, O., Buhlmann, P., et al. (1984). Classification and regression trees. *Ann. Math. Statist.*, 19 :293–325. (Cité pages 83 et 166.)

[43] Brostow, G., Fauqueur, J., and Cipolla, R. (2009). Semantic object classes in video : A high-definition ground truth database. *Pattern Recognition Letters*, 30(2) :88–97. (Cité page 69.)

[44] Brostow, G., Shotton, J., Fauqueur, J., and Cipolla, R. (2008). Segmentation and recognition using structure from motion point clouds. *Lecture Notes in Computer Science*, 5302(PT 1) :44–57. (Cité pages 3 et 69.)

[45] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. *ECCV*, pages 25–36. (Cité page 158.)

[46] Brox, T., Rousson, M., Deriche, R., and Weickert, J. (2010). Colour, texture, and motion in level set based segmentation and tracking. *IVC*, 28(3) :376–390. (Cité pages 66 et 68.)

[47] Bruhn, A., Weickert, J., and Schnörr, C. (2005). Lucas/Kanade meets Horn/Schunck : Combining local and global optic flow methods. *IJCV*, 61(3) :211–231. (Cité page 158.)

[48] Calderara, S., Melli, R., Prati, A., and Cucchiara, R. (2006). Reliable background suppression for complex scenes. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 211–214. ACM. (Cité page 73.)

[49] Cannons, K. and Wildes, R. (2007). Spatiotemporal oriented energy features for visual tracking. *ACCV*, pages 532–543. (Cité page 68.)

[50] Capel, D. (2004). *Image mosaicing and super-resolution*. Springer-Verlag New York Inc. (Cité page 16.)

[51] Chan, M., Hoogs, A., Sun, Z., Schmiederer, J., Bhotika, R., and Doretto, G. (2006). Event recognition with fragmented object tracks. *Pattern Recognition*, 1 :412–416. (Cité page 13.)

[52] Christel, M. et al. (2007). Examining user interactions with video retrieval systems. In *Proceedings of International Society for Optical Engineering Conference (SPIE)*, volume 6506. Citeseer. (Cité page 12.)

[53] Cluff, S., Morse, B., Duchaineau, M., and Cohen, J. (2009). GPU-accelerated hierarchical dense correspondence for real-time aerial video processing. In *Motion and Video Computing, 2009. WMVC'09. Workshop on*, pages 1–8. IEEE. (Cité pages 16 et 71.)

[54] Cooke, T. (2008). Two Applications of Graph-Cuts to Image Processing. In *DICTA*, pages 498–504. (Cité pages 159 et 162.)

[55] Corpetti, T., Heitz, D., Arroyo, G., Memin, E., and Santa-Cruz, A. (2006). Fluid experimental flow estimation based on an optical-flow scheme. *Experiments in fluids*, 40(1) :80–97. (Cité page 153.)

[56] Coulange, B. and Moisan, L. (2010). An aliasing detection algorithm based on suspicious colocalizations of fourier coefficients. In *ICIP*, pages 2013–2016. IEEE. (Cité page 21.)

[57] Csurka, G. and Bouthemy, P. (1999). Direct identification of moving objects and background from 2D motion models. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 566–571. IEEE. (Cité page 45.)

[58] Czyz, J. (2006). Object detection in video via particle filters. *Pattern Recognition*, 1 :820–823. (Cité pages 17 et 72.)

[59] Dai, S. and Wu, Y. (2008). Estimating space-variant motion blur without deblurring. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 661–664. IEEE. (Cité pages 24 et 26.)

[60] Dalal, N. and Triggs, B. (2005). Histogram of oriented gradient object detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, volume 2, page 4. (Cité page 169.)

[61] Davey, S., Rutten, M., and Cheung, B. (2008). A comparison of detection performance for several track-before-detect algorithms. *EURASIP Journal on Advances in Signal Processing*, 2008 :41. (Cité page 16.)

[62] de Haan, G., Piguillet, H., and Post, F. (2010). Spatial Navigation for Context-Aware Video Surveillance. *Computer Graphics and Applications, IEEE*, 30(5) :20–31. (Cité page 12.)

[63] Delong, A., Osokin, A., Isack, H., and Boykov, Y. (2010). Fast approximate energy minimization with label costs. In *CVPR*, pages 2173–2180. IEEE. (Cité page 72.)

[64] Demiriz, A., Bennett, K., and Shawe-Taylor, J. (2002). Linear programming boosting via column generation. *Machine Learning*, 46(1) :225–254. (Cité page 167.)

[65] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38. (Cité page 157.)

[66] Dietterich, T. (2000). Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15. (Cité page 166.)

[67] Dimitrova, N., Zhang, H., Shahraray, B., Sezan, I., Huang, T., and Zakhor, A. (2002). Applications of video-content analysis and retrieval. *Multimedia, IEEE*, 9(3) :42–55. (Cité page 9.)

[68] Ding, Y. and Fan, G. (2009). Sports video mining via multichannel segmental hidden markov models. *Multimedia, IEEE Transactions on*, 11(7) :1301–1309. (Cité page 14.)

[69] Ebrahimi Moghaddam, M. and Jamzad, M. (2007). Motion blur identification in noisy images using mathematical models and statistical measures. *Pattern recognition*, 40(7) :1946–1957. (Cité page 26.)

[70] Enkelmann, W. (1988). Investigations of multigrid algorithms for the estimation of optical flow fields in image sequences. *Computer Vision, Graphics, and Image Processing*, 43(2) :150–177. (Cité page 159.)

[71] Fan, J., Xu, J., and Wu, Y. (2009). Context-aware tracking of small targets in video. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7445, page 7. (Cité pages 3 et 70.)

[72] Fang, X., Luo, B., He, B., and Wu, H. (2010). Feature based multi-resolution registration of blurred images for image mosaic. *International Journal of CAD/CAM*, 9(1). (Cité page 39.)

[73] Farsiu, S., Elad, M., and Milanfar, P. (2006). Video-to-video dynamic super-resolution for grayscale and color sequences. *EURASIP Journal on Applied Signal Processing*, 2006 :232–232. (Cité page 2.)

[74] Filipovych, R. and Ribeiro, E. (2009). Adaptive tuboid shapes for action recognition. *Advances in Visual Computing*, pages 367–376. (Cité page 75.)

[75] Fleet, D., Black, M., Yacoob, Y., and Jepson, A. (2000). Design and use of linear models for image motion analysis. *IJCV*, 36(3) :171–193. (Cité pages 156 et 162.)

[76] Fleet, D. and Jepson, A. (1990). Computation of component image velocity from local phase information. *IJCV*, 5(1) :77–104. (Cité page 155.)

[77] Freund, Y. and Mason, L. (1999). The alternating decision tree algorithm. *ICML99*, pages 124–133. (Cité page 166.)

[78] Freund, Y. and Schapire, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer. (Cité page 167.)

[79] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Special invited paper. additive logistic regression : A statistical view of boosting. *Annals of statistics*, pages 337–374. (Cité pages 83 et 167.)

[80] Gaborski, R., Vaingankar, V., Chaoji, V., and Teredesai, A. (2002). VENUS : A System for Novelty Detection in Video Streams with Learning. (Cité page 14.)

[81] Geman, S., McClure, D., and for Intelligent Control Systems (US), C. (1987). *Statistical methods for tomographic image reconstruction*. Center for Intelligent Control Systems. (Cité pages 47 et 160.)

[82] Georgeson, M., May, K., Freeman, T., and Hesse, G. (2007). From filters to features : Scale–space analysis of edge and blur coding in human vision. *Journal of vision*, 7(13). (Cité pages 22 et 23.)

[83] Gibson, J. (1950). *The perception of the visual world*. Houghton Mifflin. (Cité page 153.)

[84] Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). Markov chain monte carlo in practice. (Cité page 168.)

[85] Glocker, B., Paragios, N., Komodakis, N., Tziritas, G., and Navab, N. (2008). Optical flow estimation with uncertainties through dynamic MRFs. In *CVPR*, pages 1–8. (Cité pages 159 et 162.)

[86] Glynn, P. and Iglehart, D. (1989). Importance sampling for stochastic simulations. *Management Science*, pages 1367–1392. (Cité page 168.)

[87] Golland, P. and Bruckstein, A. (1997). Motion from Color. *CVIU*, 68(3) :346–362. (Cité page 160.)

[88] Grabner, H. and Bischof, H. (2006). Online boosting and vision. In *CVPR*, volume 1, pages 260–267. (Cité page 167.)

[89] Grzywacz, N. and Yuille, A. (1990). A model for the estimate of local velocity by cells in the visual cortex. *Proceedings of the Royal Society of London B*, 239 :129–161. (Cité page 155.)

[90] Guo, H. and Hsu, W. (2002). A survey of algorithms for real-time bayesian network inference. In *In In the joint AAAI-02/KDD-02/UAI-02 workshop on Real-Time Decision Support and Diagnosis Systems*. Citeseer. (Cité page 168.)

[91] Güting, R. and Schneider, M. (2005). *Moving objects databases*. Morgan Kaufmann Pub. (Cité page 13.)

[92] Hager, G. and Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination. *PAMI*, 20(10) :1025–1039. (Cité page 160.)

[93] Harasse, S., Bonnaud, L., and Desvignes, M. (2007). Content and illumination invariant blur measures for realtime video processing. *cit*, pages 551–556. (Cité pages 2, 22 et 23.)

[94] Hartley, R. and Zisserman, A. (2003). Multiple view geometry in computer vision. (Cité page 2.)

[95] Hastie, T., Tibshirani, R., and Friedman, J. (2009). Boosting and additive trees. *The Elements of Statistical Learning*, pages 1–51. (Cité page 166.)

[96] Haussecker, H. and Fleet, D. (2001). Estimating optical flow with physical models of brightness variation. *PAMI*, 23(6) :661–673. (Cité page 160.)

[97] He, X., King, O., Ma, W., Li, M., and Zhang, H. (2003). Learning a semantic space from user's relevance feedback for image retrieval. *Circuits and Systems for Video Technology*, 13(1) :39–48. (Cité page 10.)

[98] He, X., Zemel, R., and Carreira-Perpinan, M. (2004). Multiscale conditional random fields for image labeling. In *CVPR*, pages 695–703. IEEE Computer Society. (Cité page 70.)

[99] Hearst, M., Dumais, S., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4) :18–28. (Cité page 168.)

[100] Heeger, D. (1987). Model for the extraction of image flow. *JOSA A*, 4(8) :1455–1471. (Cité page 164.)

[101] Heitz, G. and Koller, D. (2008). Learning spatial context : Using stuff to find things. In *ECCV : Part I*, pages 30–43. Springer-Verlag. (Cité pages 3 et 70.)

[102] Hoogs, A., Bush, S., Brooksby, G., Perera, A., Dausch, M., and Krahnstoever, N. (2008). Detecting semantic group activities using relational clustering. In *IEEE Workshop on Motion and video Computing*, pages 1–8. IEEE. (Cité page 13.)

[103] Hoogs, A., Chan, M., Bhotika, R., and Schmiederer, J. (2005). Recognizing complex behaviors in aerial video. In *Proc. International Conference on Intelligence Analysis*. (Cité page 13.)

[104] Horn, B. and Schunck, B. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3) :185–203. (Cité pages 156, 158, 159, 161, 162 et 163.)

[105] Hsu, P. and Chen, B. (2008). Blurred image detection and classification. In *Proceedings of the 14th international conference on Advances in multimedia modeling*, pages 277–286. Springer-Verlag. (Cité pages 22 et 23.)

[106] Hu, H. and De Haan, G. (2006). Low cost robust blur estimator. In *Image Processing, 2006 IEEE International Conference on*, pages 617–620. IEEE. (Cité pages 22, 23 et 24.)

[107] Hu, H. and De Haan, G. (2007). Adaptive image restoration based on local robust blur estimation. In *Proceedings of the 9th international conference on Advanced concepts for intelligent vision systems*, pages 461–472. Springer-Verlag. (Cité page 22.)

[108] Huang, C., Lee, H., and Chen, C. (2008a). Shot change detection via local keypoint matching. *Multimedia, IEEE Transactions on*, 10(6) :1097–1108. (Cité page 44.)

[109] Huang, T., Dagli, C., Rajaram, S., Chang, E., Mandel, M., Poliner, G., and Ellis, D. (2008b). Active learning for interactive multimedia retrieval. *Proceedings of the IEEE*, 96(4) :648–667. (Cité page 14.)

[110] Huber, P. (1973). Robust regression : asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5) :799–821. (Cité page 157.)

[111] Huiskes, M. and Lew, M. (2008). Performance evaluation of relevance feedback methods. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 239–248. ACM. (Cité page 10.)

[112] Irani, M. and Anandan, P. (1998). A unified approach to moving object detection in 2d and 3d scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(6) :577–589. (Cité page 71.)

[113] Jansen, M., Heeren, W., and van Dijk, B. (2008). Videotrees : Improving video surrogate presentation using hierarchy. In *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, pages 560–567. IEEE. (Cité page 16.)

[114] Jepson, A. and Black, M. (1993). Mixture models for optical flow computation. In *CVPR*, pages 760–761. IEEE. (Cité page 157.)

[115] Jähne, B. (1990). Motion determination in space-time images. *ECCV*, pages 161–173. (Cité page 155.)

[116] Jiang, J. and Tu, Z. (2009). Efficient scale space auto-context for image segmentation and labeling. In *CVPR*. (Cité pages 3 et 70.)

[117] Johnson, R., Sasiadek, J., and Zalewski, J. (2003). Kalman Filter Enhancement for UAV Navigation. *SIMULATION SERIES*, 35(1) :267–272. (Cité pages 17 et 72.)

[118] Jurie, F. and Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 604–610. IEEE. (Cité page 169.)

[119] Kaplan, L. and Nasrabadi, N. (2009). Block Wiener-based image registration for moving target indication. *Image and Vision Computing*, 27(6) :694–703. (Cité page 16.)

[120] Kapler, T. and Wright, W. (2005). Geotime information visualization. *Information Visualization*, 4(2) :136–146. (Cité page 14.)

[121] Karpenko, A. and Aarabi, P. (2008). Tiny videos : Non-parametric content-based video retrieval and recognition. In *Tenth IEEE International Symposium on Multimedia*, pages 619–624. IEEE. (Cité page 15.)

[122] Kender, J. and Yeo, B. (2000). On the structure and analysis of home videos. In *Proceedings of the Asian Conference on Computer Vision*. Citeseer. (Cité page 44.)

[123] Kindermann, R. and Snell, J. (1980). *Markov random fields and their applications*, volume 1. Amer Mathematical Society. (Cité page 168.)

[124] Kluckner, S., Mauthner, T., Roth, P., and Bischof, H. (2010). Semantic classification in aerial imagery by integrating appearance and height information. *Computer Vision–ACCV 2009*, pages 477–488. (Cité page 68.)

[125] Kokaram, A., Rea, N., Dahyot, R., Tekalp, M., Bouthemy, P., Gros, P., and Sezan, I. (2006). Browsing sports video : trends in sports-related indexing and retrieval work. *Signal Processing Magazine, IEEE*, 23(2) :47–58. (Cité pages 12 et 14.)

[126] Kolekar, M., Palaniappan, K., Sengupta, S., and Seetharaman, G. (2009). Semantic concept mining based on hierarchical event detection for soccer video indexing. *Journal of Multimedia*, 4(5) :298–312. (Cité page 14.)

[127] Koprinska, I. and Carrato, S. (2001). Temporal video segmentation : A survey. *Signal processing : Image communication*, 16(5) :477–500. (Cité page 16.)

[128] Kumar, S. and Hebert, M. (2003). Discriminative random fields : A discriminative framework for contextual interaction in classification. In *ICCV*, volume 2, pages 1150–1157. (Cité page 69.)

[129] Lacoste, C., Fablet, R., Bouthemy, P., and Yao, J. (2002). Video summarization using a statistical approach. *Congrés Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*. (Cité page 45.)

[130] Ladicky, L., Russell, C., Kohli, P., and Torr, P. (2009). Associative hierarchical crfs for object class image segmentation. In *ICCV*, pages 739–746. IEEE. (Cité page 116.)

[131] Ladicky, L., Sturgess, P., Alahari, K., Russell, C., and Torr, P. (2010). What, where and how many ? combining object detectors and crfs. *ECCV*, pages 424–437. (Cité pages 3 et 69.)

[132] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289. Citeseer. (Cité page 69.)

[133] Lagendijk, R. and Biemond, J. (1999). Basic methods for image restoration and Identification. (Cité page 22.)

[134] Lan, D., Ma, Y., and Zhang, H. (2003). A novel motion-based representation for video mining. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–469. IEEE. (Cité page 44.)

[135] Le Besnerais, G. and Champagnat, F. (2005). Dense optical flow by iterative local window registration. In *ICIP*, volume 1, pages I–137. (Cité page 158.)

[136] Lee, C., Wang, S., Jiao, F., Schuurmans, D., and Greiner, R. (2007). Learning to model spatial dependency : Semi-supervised discriminative random fields. *Advances in Neural Information Processing Systems*, 19 :793. (Cité page 70.)

[137] Lefèvre, S., Holler, J., and Vincent, N. (2003). A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, 9(1) :73–98. (Cité pages 16 et 44.)

[138] Lei, C. and Yang, Y. (2009). Optical flow estimation on coarse-to-fine region-trees using discrete optimization. In *ICCV*, pages 1562–1569. (Cité pages 159 et 162.)

[139] Lempitsky, V., Roth, S., and Rother, C. (2008). FusionFlow : Discrete-continuous optimization for optical flow estimation. In *CVPR*, pages 1–8. (Cité pages 159 et 162.)

[140] Levin, A. (2007). Blind motion deblurring using image statistics. *Advances in Neural Information Processing Systems*, 19 :841. (Cité page 24.)

[141] Lin, S., Shi, Y., and Zhang, Y. (1997). An optical flow based motion compensation algorithm for very low bit-rate video coding. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 4, pages 2869–2872. IEEE. (Cité page 153.)

[142] Lin, W., Oakes, M., and Tait, J. (2008). Real adaboost for large vocabulary image classification. In *International Workshop on Content-Based Multimedia Indexing*, pages 192–199. IEEE. (Cité page 167.)

[143] Lin, W., Sun, M., Li, H., and Hu, H. (2011). A new shot change detection method using information from motion estimation. *Advances in Multimedia Information Processing-PCM 2010*, pages 264–275. (Cité page 44.)

[144] Lin, Y. and Medioni, G. (2007). Map-enhanced UAV image sequence registration and synchronization of multiple image sequences. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE. (Cité pages 2 et 16.)

[145] Lindeberg, T. (1996). Scale-space : A framework for handling image structures at multiple scales. *CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH-REPORTS-CERN*, pages 27–38. (Cité page 22.)

[146] Liu, C., Yuen, J., and Torralba, A. (2009a). Nonparametric scene parsing : Label transfer via dense scene alignment. In *CVPR*, pages 1972–1979. IEEE. (Cité page 68.)

[147] Liu, C., Yuen, P., and Qiu, G. (2009b). Object motion detection using information theoretic spatio-temporal saliency. *Pattern Recognition*, 42(11) :2897–2906. (Cité pages 17 et 73.)

[148] Liu, F. and Gleicher, M. (2009). Learning color and locality cues for moving object detection and segmentation. In *CVPR*, pages 320–327. IEEE. (Cité page 66.)

[149] Liu, R., Li, Z., and Jia, J. (2008). Image partial blur detection and classification. In *CVPR*, pages 1–8. IEEE. (Cité pages 2, 22, 23 et 24.)

[150] Liu, X. and Yu, T. (2007). Gradient feature selection for online boosting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE. (Cité pages 167 et 169.)

[151] Lokhande, R., Arya, K., and Gupta, P. (2006). Identification of parameters and restoration of motion blurred images. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 301–305. ACM. (Cité page 26.)

[152] Loveland, R., Rosten, E., and Porter, R. (2008). Improving multiple target tracking in structured environments using velocity priors. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6969, page 14. Citeseer. (Cité page 17.)

[153] Lowe, D. (1999). Object recognition from local scale-invariant features. In *ICCV*, volume 2, pages 1150–1157. Ieee. (Cité page 80.)

[154] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2) :91–110. (Cité pages 16 et 169.)

[155] Létienne, A., Champagnat, F., Kulcsàr, C., Le Besnerais, G., and DeLesegno, P. (2008). Fast super-resolution on moving objects in video sequences. In *EUSIPCO European Signal Processing Conference*. (Cité page 39.)

[156] Lucas, B., Kanade, T., et al. (1981). An iterative image registration technique with an application to stereo vision. In *International joint conference on artificial intelligence*, volume 3, pages 674–679. Citeseer. (Cité pages 156, 158, 159 et 163.)

[157] Magarey, J. and Kingsbury, N. (1998). Motion estimation using a complex-valued wavelet transform. *Signal Processing, IEEE Transactions on*, 46(4) :1069–1084. (Cité page 71.)

[158] Markandey, V. and Flinchbaugh, B. (1990). Multispectral constraints for optical flow computation. In *ICCV*, pages 38–41. (Cité page 160.)

[159] Marquis-Bolduc, M., Deschênes, F., and Pan, W. (2008). Combining apparent motion and perspective as visual cues for content-based camera motion indexing. *Pattern Recognition*, 41(2) :445–457. (Cité page 2.)

[160] Matthies, L., Kanade, T., and Szeliski, R. (1989). Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3(3) :209–238. (Cité page 2.)

[161] Mattivi, R. and Shao, L. (2011). Robust spatio-temporal features for human action recognition. In Lin, W., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E., and Wang, H., editors, *Multimedia Analysis, Processing and Communications*, volume 346 of *Studies in Computational Intelligence*, pages 351–367. Springer Berlin - Heidelberg. (Cité page 74.)

[162] Mazor, E., Averbuch, A., Bar-Shalom, Y., and Dayan, J. (1998). Interacting multiple model methods in target tracking : a survey. *Aerospace and Electronic Systems, IEEE Transactions on*, 34(1) :103–123. (Cité page 17.)

[163] McCallum, A., Freitag, D., and Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. In *ICML*, pages 591–598. Citeseer. (Cité page 69.)

[164] Mikolajczyk, K. and Schmid, C. (2005). A performance evaluation of local descriptors. *PAMI*, pages 1615–1630. (Cité page 80.)

[165] Min, C., Yu, Q., and Medioni, G. (2006). Multi-layer Mosaics in the Presence of Motion and Depth Effects. *Pattern Recognition*, 1 :992–995. (Cité pages 2 et 15.)

[166] Min, Z. (2009). MPEG-7 Feature Based Shot Change Detection for Scenery Video. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, volume 6, pages 388–391. IEEE. (Cité page 44.)

[167] Möller, B., Garcia, R., and Posch, S. (2007). Towards objective quality assessment of image registration results. In *Proceedings of the 2nd International Conference on Computer Vision Theory and Applications, VISAPP*. Citeseer. (Cité page 16.)

[168] Money, A. and Agius, H. (2008). Video summarisation : A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2) :121–143. (Cité page 16.)

[169] Muñoz-Salinas, R., Aguirre, E., García-Silvente, M., and Gonzalez, A. (2008). A multiple object tracking approach that combines colour and depth information using a confidence measure. *Pattern Recognition Letters*, 29(10) :1504 – 1514. (Cité page 17.)

[170] Murphy, K., Weiss, Y., and Jordan, M. (1999). Loopy belief propagation for approximate inference : An empirical study. In *Proceedings of Uncertainty in AI*, volume 9, pages 467–475. Citeseer. (Cité page 168.)

[171] Nagel, H. (1987). On the estimation of optical flow : Relations between different approaches and some new results. *Artificial Intelligence*, 33(3) :299–324. (Cité page 160.)

[172] Nam, J. and Tewfik, A. (1999). Dynamic video summarization and visualization. In *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*, pages 53–56. ACM. (Cité pages 45 et 57.)

[173] Nestares, O. and Fleet, D. (2003). Error-in-variables likelihood functions for motion estimation. In *ICIP*, volume 3, pages III–77. IEEE. (Cité page 160.)

[174] Nikitidis, S., Zafeiriou, S., and Pitas, I. (2008). Camera motion estimation using a novel online vector field model in particle filters. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(8) :1028–1039. (Cité page 2.)

[175] Nir, T., Bruckstein, A., and Kimmel, R. (2008). Over-parameterized variational optical flow. *International Journal of Computer IJCV*, 76(2) :205–216. (Cité page 158.)

[176] Odobez, J. and Bouthemy, P. (1995). Robust multiresolution estimation of parametric motion models. *Journal of visual communication and image representation*, 6(4) :348–365. (Cité pages 45, 46 et 82.)

[177] Ohta, N. (1990). Optical flow detection by color images. *NEC research & development*, (97) :78–84. (Cité page 160.)

[178] Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, pages 971–987. (Cité page 169.)

[179] Ommer, B., Mader, T., and Buhmann, J. (2009). Seeing the objects behind the dots : Recognition in videos from a moving camera. *IJCV*, 83(1) :57–71. (Cité page 68.)

[180] Over, P., Awad, G., Fiscus, J., Michel, M., Smeaton, A., and Kraaij, W. (2010). TRECVID 2009-goals, tasks, data, evaluation mechanisms and metrics. (Cité page 44.)

[181] Over, P., Smeaton, A., and Awad, G. (2008). The trecvid 2008 BBC rushes summarization evaluation. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 1–20. ACM. (Cité page 12.)

[182] Oza, N. (2001). Online bagging and boosting. In *Systems, man and cybernetics, 2005 IEEE international conference on*, volume 3, pages 2340–2345. IEEE. (Cité page 167.)

[183] Ozcanli, O., Tamrakar, A., and Kimia, B. (2006). Augmenting shape with appearance in vehicle category recognition. (Cité page 16.)

[184] Paliwal, M. and Kumar, U. (2009). Neural networks and statistical techniques : A review of applications. *Expert Systems with Applications*, 36(1) :2–17. (Cité page 168.)

[185] Pan, W. and Deschenes, F. (2006). Interpreting camera operations in the context of content-based video indexing and retrieval. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, pages 7–7. IEEE. (Cité pages 2 et 3.)

[186] Papenberg, N., Bruhn, A., Brox, T., Didas, S., and Weickert, J. (2006). Highly accurate optic flow computation with theoretically justified warping. *IJCV*, 67(2) :141–158. (Cité pages 157 et 162.)

[187] Park, S., Park, M., and Kang, M. (2003). Super-resolution image reconstruction : a technical overview. *Signal Processing Magazine, IEEE*, 20(3) :21–36. (Cité page 16.)

[188] Parks, D. and Fels, S. (2008). Evaluation of background subtraction algorithms with post-processing. In *Advanced Video and Signal Based Surveillance*, pages 192–199. IEEE. (Cité page 73.)

[189] Patel, D. and Chaudhuri, S. (2009). Performance analysis for image super-resolution using blur as a cue. In *Advances in Pattern Recognition, 2009. ICAPR'09. Seventh International Conference on*, pages 73–76. IEEE. (Cité page 39.)

[190] Patino, L., Bremond, F., and Thonnat, M. (2010). Activity discovery from video employing soft computing relations. In *IJCNN*, pages 1–8. (Cité page 14.)

[191] Pearl, J. (1998). Bayesian networks. In *The handbook of brain theory and neural networks*, pages 149–153. MIT Press. (Cité page 168.)

[192] Pham, T. and Smeulders, A. (2008). Quadratic boosting. *Pattern Recognition*, 41(1) :331–341. (Cité page 167.)

[193] Piccardi, M. (2004). Background subtraction techniques : a review. In *ICSMC*, volume 4, pages 3099–3104. Ieee. (Cité page 66.)

[194] Pollard, E., Pannetier, B., and Rombaut, M. (2011). Hybrid algorithms for multitarget tracking using mht and gm-cphd. *IEEE Transactions on Aerospace and Electronic Systems*, 47(2) :832–847. (Cité page 72.)

[195] Porter, R., Harvey, N., and Theiler, J. (2009a). A change detection approach to moving object detection in low fame-rate video. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7341, page 23. (Cité page 16.)

[196] Porter, R., Ruggiero, C., and Morrison, J. (2009b). A framework for activity detection in wide-area motion imagery. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7341, page 19. (Cité page 13.)

[197] Pritch, Y., Ratovitch, S., Hendel, A., and Peleg, S. (2009). Clustered synopsis of surveillance video. In *Advanced Video and Signal Based Surveillance, 2009*, pages 195–200. IEEE. (Cité page 75.)

[198] Pritch, Y., Rav-Acha, A., and Peleg, S. (2008). Nonchronological video synopsis and indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1971–1984. (Cité pages 16, 39, 68, 74 et 75.)

[199] Privett, G. and Kent, P. (2005). Automated image registration with arachnid. In *SPIE*, volume 5809, pages 186–196. (Cité page 71.)

[200] Qi, B., Ghazal, M., and Amer, A. (2008). Robust global motion estimation oriented to video object segmentation. *Image Processing, IEEE Transactions on*, 17(6) :958–967. (Cité page 2.)

[201] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., and Belongie, S. (2007). Objects in context. In *ICCV*, pages 1–8. IEEE. (Cité page 70.)

[202] Rajan, D. and Chaudhuri, S. (2002). Super-resolution imaging using blur as a cue. *Super-resolution imaging*, pages 107–129. (Cité page 39.)

[203] Ranchin, F., Chambolle, A., and Dibos, F. (2007). Total variation minimization and graph cuts for moving objects segmentation. *Scale Space and Variational Methods in Computer Vision*, pages 743–753. (Cité page 144.)

[204] Rapantzikos, K., Avrithis, Y., and Kollias, S. (2009a). Dense saliency-based spatiotemporal feature points for action recognition. (Cité page 17.)

[205] Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., and Kollias, S. (2009b). Spatiotemporal saliency for video classification. *Signal Processing : Image Communication*, 24(7) :557–571. (Cité page 73.)

[206] Rav-Acha, A., Pritch, Y., and Peleg, S. (2006). Making a long video short : Dynamic video synopsis. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 435–441. IEEE. (Cité page 58.)

[207] Reibman, A. and Suthaharan, S. (2008). A no-reference spatial aliasing measure for digital image resizing. In *ICIP*, pages 1184–1187. IEEE. (Cité page 21.)

[208] Ren, W., Singh, S., Singh, M., and Zhu, Y. (2009). State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition*, 42(2) :267–282. (Cité page 17.)

[209] Ripley, B. (1996). Pattern recognition and neural networks. (Cité page 168.)

[210] Rochefort, G., Champagnat, F., Le Besnerais, G., and Giovannelli, J. (2006). An improved observation model for super-resolution under affine motion. *Image Processing, IEEE Transactions on*, 15(11) :3325–3337. (Cité page 39.)

[211] Rodriguez, M. (2010). CRAM : Compact representation of actions in movies. In *CVPR*, pages 3328–3335. IEEE. (Cité page 16.)

[212] Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6) :386. (Cité page 167.)

[213] Rosten, E., Loveland, R., and Hickman, M. (2009). Automatic creation of urban velocity fields from aerial video. *Arxiv preprint arXiv :0912.1310*. (Cité page 15.)

[214] Rui, Y., Huang, T., Ortega, M., and Mehrotra, S. (1998). Relevance feedback : A power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5) :644–655. (Cité page 14.)

[215] Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In *Parallel distributed processing : explorations in the microstructure of cognition, vol. 1*, pages 318–362. MIT Press. (Cité page 168.)

[216] Saleemi, I., Shafique, K., and Shah, M. (2009). Probabilistic modeling of scene dynamics for applications in visual surveillance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(8) :1472–1485. (Cité page 14.)

[217] Salgian, G., Bergen, J., Samarasekera, S., and Kumar, R. (2006). *Moving target indication from a moving camera in the presence of strong parallax*. Citeseer. (Cité pages 66 et 67.)

[218] Scholkopf, B. and Smola, A. (2001). Learning with kernels : Support vector machines, regularization, optimization, and beyond. (Cité page 168.)

[219] Seitz, S. and Baker, S. (2009). Filter flow. In *ICCV*, pages 143–150. (Cité pages 158 et 162.)

[220] Sheikh, Y., Javed, O., and Kanade, T. (2009). Background subtraction for freely moving cameras. In *ICCV*, pages 1219–1225. IEEE. (Cité page 67.)

[221] Shi, J. and Tomasi, C. (1994). Good features to track. In *CVPR*, pages 593–600. IEEE. (Cité page 71.)

[222] Shi, T. and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1) :118–138. (Cité page 167.)

[223] Shimojo, S., Silverman, G. H., and Nakayama, K. (1989). Occlusion and the solution to the aperture problem for motion. *Vision Research*, 29(5) :619 – 626. (Cité page 154.)

[224] Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8. IEEE. (Cité page 68.)

[225] Singh, A. (1991). *Optic flow computation : a unified perspective*. IEEE Computer Society. (Cité page 155.)

[226] Smoliar, S. and Zhang, H. (1994). Content based video indexing and retrieval. *Multimedia, IEEE*, 1(2) :62–72. (Cité page 44.)

[227] Strat, T., Arambel, P., Antone, M., Rago, C., and Landan, H. (2004). A Multiple-Hypothesis Tracking of Multiple Ground Targets from Aerial Video with Dynamic Sensor Control. In *Fusion 2004 : Seventh International Conference on Information Fusion ; Stockholm*. International Society of Information Fusion, ONERA-DTIM, BP 72, 29 Av. de la Division Leclerc, Chatillon, 92320, France,. (Cité page 17.)

[228] Su, C., Liao, H., Tyan, H., Lin, C., Chen, D., and Fan, K. (2007). Motion flow-based video retrieval. *Multimedia, IEEE Transactions on*, 9(6) :1193–1201. (Cité page 44.)

[229] Sun, J., Zhang, W., Tang, X., and Shum, H. (2006). Background cut. *ECCV*, pages 628–641. (Cité page 74.)

[230] Swears, E., Hoogs, A., and Perera, A. (2008). Learning Motion Patterns in Surveillance Video using HMM Clustering. In *Proceedings of the 2008 IEEE Workshop on Motion and video Computing*, pages 1–8. IEEE Computer Society. (Cité page 17.)

[231] Szeliski, R. (2006). Image alignment and stitching : A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2(1) :1–104. (Cité pages 2, 15 et 16.)

[232] Tang, C., Medioni, G., and Lee, M. (2001). N-dimensional tensor voting, application to epipolar geometry estimation. *PAMI*, 23(8) :829–844. (Cité pages 73 et 136.)

[233] Thomas, J. and Cook, K. (2005). Illuminating the path : The research and development agenda for visual analytics. *IEEE Computer Society*. (Cité page 15.)

[234] Tian, Y., Hampapur, A., Brown, L., Feris, R., Lu, M., Senior, A., Shu, C., and Zhai, Y. (2009). Event Detection, Query, and Retrieval for Video Surveillance. *Artificial intelligence for maximizing content based image retrieval*. (Cité page 12.)

[235] Tiburzi, F. and Bescos, J. (2007). Camera Motion Analysis in On-line MPEG Sequences. In *WIAMIS*, pages 42–42. IEEE. (Cité page 44.)

[236] Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE. (Cité page 94.)

[237] Torralba, A., Murphy, K., and Freeman, W. (2007). Sharing visual features for multiclass and multiview object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(5) :854–869. (Cité page 169.)

[238] Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. (2000). Bundle adjustment ?a modern synthesis. *Vision algorithms : theory and practice*, pages 153–177. (Cité page 16.)

[239] Trobin, W., Pock, T., Cremers, D., and Bischof, H. (2008a). An unbiased second-order prior for high-accuracy motion estimation. *Pattern Recognition*, pages 396–405. (Cité page 158.)

[240] Trobin, W., Pock, T., Cremers, D., and Bischof, H. (2008b). Continuous energy minimization via repeated binary fusion. *ECCV*, pages 677–690. (Cité pages 159 et 162.)

[241] Tu, Z. (2008). Auto-context and its application to high-level vision tasks. In *CVPR*, pages 1–8. IEEE. (Cité pages 3, 70 et 100.)

[242] Uras, S., Girosi, F., Verri, A., and Torre, V. (1988). A computational approach to motion perception. *Biological Cybernetics*, 60(2) :79–87. (Cité page 160.)

[243] Urvoy, M., Cammas, N., Pateux, S., Déforges, O., Babel, M., and Pressigout, M. (2009). Motion tubes for the representation of image sequences. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 105–108. IEEE. (Cité page 16.)

[244] Vapnik, V. (1995). The nature of statistical learning theory. (Cité page 168.)

[245] Veeraraghavan, H., Schrater, P., and Papanikolopoulos, N. (2006). Robust target detection and tracking through integration of motion, color, and geometry. *Computer Vision and Image Understanding*, 103(2) :121–138. (Cité page 17.)

[246] Venkatesh Babu, R., Suresh, S., and Perkis, A. (2007). No-reference jpeg-image quality assessment using gap-rbf. *Signal Processing*, 87(6) :1493–1503. (Cité page 27.)

[247] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE Comput. Soc. (Cité page 169.)

[248] Waizenegger, W., Feldmann, I., and Schreer, O. (2008). Semantic annotation and retrieval of unedited video based on extraction of 3d camera motion. In *CBMI*, pages 265–271. IEEE. (Cité pages 3 et 44.)

[249] Wallach, H. (1935). Über visuell wahrgenommene bewegungsrichtung. *Psychological Research*, 20 :325–380. (Cité page 154.)

[250] Wang, H., Ullah, M. M., Klaser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. (Cité page 74.)

[251] Wang, T., Mei, T., Hua, X., Liu, X., and Zhou, H. (2007). Video collage : A novel presentation of video sequence. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1479–1482. IEEE. (Cité page 16.)

[252] Wang, X., Tian, B., Liang, C., and Shi, D. (2008a). Blind image quality assessment for measuring image blur. In *2008 Congress on Image and Signal Processing*, pages 467–470. IEEE. (Cité pages 2, 22 et 23.)

[253] Wang, Y., Fevig, R., and Schultz, R. (2008b). Super-resolution mosaicking of uav surveillance video. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 345–348. IEEE. (Cité pages 2 et 16.)

[254] Wang, Z., Bovik, A., and Evan, B. (2000). Blind measurement of blocking artifacts in images. In *Image Processing*, volume 3, pages 981–984. IEEE. (Cité page 21.)

[255] Wang, Z. and Li, Q. (2007). Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, 24(12) :B61–B69. (Cité page 2.)

[256] Wang, Z., Sheikh, H., and Bovik, A. (2002). No-reference perceptual quality assessment of jpeg compressed images. In *Image Processing*, volume 1, pages I–477. IEEE. (Cité page 27.)

[257] Watson, A. and Ahumada, A. (1985). Model of human visual-motion sensing. *Optical Society of America, Journal, A : Optics and Image Science*, 2 :322–342. (Cité page 155.)

[258] Weber, J. and Malik, J. (1995). Robust computation of optical flow in a multi-scale differential framework. *IJCV*, 14(1) :67–81. (Cité page 160.)

[259] Wedel, A., Pock, T., Braun, J., Franke, U., and Cremers, D. (2008). Duality TV-L1 flow with fundamental matrix prior. In *IVCNZ*, pages 1–6. (Cité pages 162 et 163.)

[260] Weickert, J., Bruhn, A., Brox, T., and Papenberg, N. (2006). A survey on variational optic flow methods for small displacements. *Mathematical models for registration and applications to medical imaging*, pages 103–136. (Cité page 156.)

[261] Weickert, J. and Schnörr, C. (2001). Variational optic flow computation with a spatio-temporal smoothness constraint. *JMIV*, 14(3) :245–255. (Cité page 160.)

[262] Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., and Bischof, H. (2009). Anisotropic huber-l1 optical flow. In *BMVC*. Citeseer. (Cité pages 157, 161, 162 et 163.)

[263] Wildemuth, B., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., and Gruss, R. (2003). How fast is too fast ? : evaluating fast forward surrogates for digital video. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 221–230. IEEE Computer Society. (Cité page 15.)

[264] Willert, V., Schmuedderich, J., Eggert, J., Goerick, C., and Koerner, E. (2008). Probabilistic optical flow estimation for large pixel displacements utilizing egomotion flow compensation. In *BMVC*, pages 695–704. (Cité page 71.)

[265] Winkler, S. (2005). *Digital video quality : vision models and metrics*. Wiley. (Cité pages 2, 16 et 21.)

[266] Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A. (1997). Pfinder : Real-time tracking of the human body. *PAMI*, 19(7) :780–785. (Cité page 73.)

[267] Wu, S., Lin, W., Xie, S., Lu, Z., Ong, E., and Yao, S. (2009). Blind blur assessment for vision-based applications. *Journal of Visual Communication and Image Representation*, 20(4) :231–241. (Cité pages 22 et 23.)

[268] Wu, S., Lu, Z., Ong, E., and Lin, W. (2007). Blind image blur identification in cepstrum domain. In *Computer Communications and Networks, 2007. ICCCN 2007. Proceedings of 16th International Conference on*, pages 1166–1171. IEEE. (Cité pages 22 et 23.)

[269] Wu, Y. and Fan, J. (2009). Contextual flow. In *CVPR*, pages 33–40. IEEE. (Cité pages 3 et 70.)

[270] Xiang, T. and Gong, S. (2008). Incremental and adaptive abnormal behaviour detection. *Computer Vision and Image Understanding*, 111(1) :59–73. (Cité page 14.)

[271] Xiao, J., Cheng, H., Han, F., and Sawhney, H. (2008). Geo-spatial aerial video processing for scene understanding and object tracking. In *CVPR*, pages 1–8. IEEE. (Cité page 14.)

[272] Xiao, J. and Shah, M. (2005). Motion layer extraction in the presence of occlusion using graph cuts. *PAMI*, 27(10) :1644–1659. (Cité pages 66 et 67.)

[273] Xiao, J., Yang, C., Han, F., and Cheng, H. (2009). Vehicle and person tracking in aerial videos. *Multimodal Technologies for Perception of Humans*, pages 203–214. (Cité page 68.)

[274] Xiong, Y. and Shafer, S. (1995). Dense structure from a dense optical flow sequence. In *ISCV*, page 1. Published by the IEEE Computer Society. (Cité page 153.)

[275] Yalcin, H., Hebert, M., Collins, R., and Black, M. (2005). A flow-based approach to vehicle detection and background mosaicking in airborne video. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 1202–vol. IEEE. (Cité pages 2, 16, 66 et 67.)

[276] Yilmaz, A., Javed, O., and Shah, M. (2006). Object tracking : A survey. *Acm Computing Surveys (CSUR)*, 38(4) :13. (Cité pages 17, 67, 81 et 169.)

[277] Yu, Q., Cohen, I., Medioni, G., and Wu, B. (2006). Boosted markov chain monte carlo data association for multiple target detection and tracking. *Pattern Recognition*, 2 :675–678. (Cité page 67.)

[278] Yu, Q. and Medioni, G. (2007). Map-enhanced detection and tracking from a moving platform with local and global data association. In *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*, pages 3–3. IEEE. (Cité page 2.)

[279] Yu, Q. and Medioni, G. (2009). Motion pattern interpretation and detection for tracking moving vehicles in airborne video. In *CVPR*, pages 2671–2678. (Cité pages 17, 67, 73, 74, 118, 136 et 139.)

[280] Yuan, C., Medioni, G., Kang, J., and Cohen, I. (2007a). Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *PAMI*, 29(9) :1627–1641. (Cité pages 74 et 136.)

[281] Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., and Zhang, B. (2007b). A formal study of shot boundary detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(2) :168–186. (Cité page 44.)

[282] Zhang, H., Zhang, L., Shen, H., and Li, P. (2009). A map approach for joint image registration, blur identification and super resolution. In *2009 Fifth International Conference on Image and Graphics*, pages 97–102. IEEE. (Cité page 39.)

[283] Zhou, H., Yuan, Y., and Shi, C. (2009). Object tracking using sift features and mean shift. *CVIU*, 113(3) :345–352. (Cité page 153.)

[284] Zimmer, H., Bruhn, A., Weickert, J., Valgaerts, L., Salgado, A., Rosenhahn, B., and Seidel, H. (2009). Complementary optic flow. In *Energy minimization methods in computer vision and pattern recognition*, pages 207–220. Springer. (Cité pages 158, 160, 161 et 162.)

[285] Zitova, B. and Flusser, J. (2003). Image registration methods : a survey. *Image and vision computing*, 21(11) :977–1000. (Cité page 16.)

[286] Zivkovic, Z. and van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7) :773–780. (Cité page 73.)

Titre Filtrage de segments informatifs dans des vidéos

Résumé Les travaux réalisés dans le cadre de cette thèse ont pour objectif d'extraire les différents segments informatifs au sein de séquences vidéo, plus particulièrement aériennes. L'interprétation manuelle de telles vidéos dans une optique de renseignement se heurte en effet au volume des données disponibles. Une assistance algorithmique fondée sur diverses modalités d'indexation est donc envisagée, dans l'objectif de repérer les "segments d'intérêt" et éviter un parcours intégral de la vidéo. Deux approches particulières ont été retenues et respectivement développées au sein de chaque partie. La partie 1 propose une utilisation des conditions de prise de vue (CPDV) comme modalités d'indexation. Une évaluation de la qualité image permet ainsi de filtrer les segments temporels de mauvaise qualité et donc inexploitables. La classification du mouvement image apparent directement lié au mouvement caméra, fournit une indexation de séquences vidéo en soulignant notamment les segments potentiels d'intérêt ou au contraire les segments difficiles présentant un mouvement très rapide ou oscillant. La partie 2 explore le contenu dynamique de la séquence vidéo, plus précisément la présence d'objets en mouvement. Une première approche locale en temps est présentée. Elle filtre les résultats d'une première classification par apprentissage supervisé en exploitant les informations de contexte, spatial puis sémantique. Différentes approches globales en temps sont par la suite explorées. De telles approches permettent de garantir la cohérence temporelle des résultats et réduire les fausses alarmes.

Mots-clés indexation de séquences vidéo, vidéo aérienne, détection d'activité, conditions de prise de vue, qualité image, apprentissage supervisé, information de contexte

Title Informative segment filtering in video sequences

Abstract The objective of this thesis is to extract the informative temporal segments from video sequences, more particularly in aerial video. Manual interpretation of such videos for information gathering faces an ever growing volume of available data. We have thus considered an algorithmic assistance based on different modalities of indexation in order to locate "segments of interest" and avoid a complete visualization of the video. We have chosen two methods in particular and have respectively developed them in each part of this thesis. Part 1 describes how viewing conditions can be used as a method of indexation. The assessment of image quality enables to filter out the temporal segments for which the quality is low and which can thus not be exploited. The classification of global image motion, which is directly linked to camera motion, leads to a method of indexation for video sequences. Indeed, it emphasizes possible segments of interest or, conversely, difficult segments for which motion is very fast or oscillating. Part 2 focuses on the dynamic content of video sequences, especially the presence of moving objects. We first present a local (in time) approach. This approach refines the results obtained after a first classification by supervised learning by using contextual information, spatial then semantic information. We have then investigated several methods for moving object detection which are global in time. Such approaches aim to enforce the temporal consistency of the detected objects and to reduce false detections.

Keywords video indexing, aerial video, activity detection, viewing conditions, image quality, supervised learning, contextual information