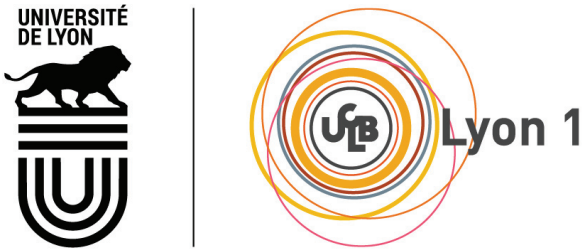
Détection d'objets en mouvement à l'aide d'une caméra mobile

Marie-Neige Chapel



N° d'ordre NNT : 2017LYSE1156

THESE DE DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

Ecole Doctorale N°512

Ecole Doctorale InfoMaths

Spécialité de doctorat : Informatique

Soutenue publiquement le 22/09/2017, par :

Marie-Neige Chapel

Détection d'objets en mouvement à l'aide d'une caméra mobile

Devant le jury composé de :

Bertolino Pascal, Maître de Conférences HDR, GIPSA-Lab,	Rapporteur
Cordier Frédéric, Maître de Conférences HDR, LMIA,	Rapporteur
Dipanda Albert, Professeur des Universités, Le2i,	Examinateur
Calabretto Sylvie, Professeure des Universités, LIRIS,	Examinatrice
Zeitouni Karine, Professeure des Universités, Laboratoire DAVID	Examinatrice
Bouakaz Saida, Professeure des Universités, LIRIS,	Directrice de thèse

Guillou Erwan, Maître de Conférences, LIRIS, Co-directeur de thèse

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Abstract

Moving objects detection in video streams is a commonly used technique in many computer vision algorithms. The detection becomes more complex when the camera is moving. The environment observed by this type of camera appeared moving and it is more difficult to distinguish the objects which are in movement from the others that composed the static part of the scene.

In this thesis we propose contributions for the detection of moving objects in the video stream of a moving camera. The main idea to differenciate between moving and static objects based on 3D distances. 3D positions of feature points extracted from images are estimated by triangulation and then their 3D motions are analyzed in order to provide a sparse static/moving labeling. To provide a more robust detection, the analysis of the 3D motions is compared to those of feature points previously estimated static. A confidance value updated over time is used to decide on labels to attribute to each point.

We make experiments on virtual (from the Previz project [1]) and real datasets (known by the community [Och+14]) and we compare the results with the state of the art. The results show that our 3D constraint coupled with a statistical and temporal analysis of motions allow to detect moving elements in the video stream of a moving camera even in complex cases where apparent motions of the scene are not similars.

Keywords : computer vision, moving objects detection, moving camera, feature points, 3D geometric constraint.

1. http ://previz.eu

Résumé

La détection d'objets mobiles dans des flux vidéo est une étape essentielle pour de nombreux algorithmes de vision par ordinateur. Cette tâche se complexifie lorsque la caméra utilisée est en mouvement. En effet, l'environnement capté par ce type de caméra apparaît en mouvement et il devient plus difficile de distinguer les objets qui effectuent réellement un mouvement de ceux qui constituent la partie statique de la scène.

Dans cette thèse, nous apportons des contributions au problème de détection d'objets mobiles dans le flux vidéo d'une caméra mobile. L'idée principale qui nous permet de distinguer les éléments mobiles de ceux qui sont statiques repose sur un calcul de distance dans l'espace 3D. Les positions 3D de caractéristiques extraites des images sont estimées par triangulation puis leurs mouvements 3D sont analysés pour réaliser un étiquetage éparse statique/mobile de ces points. Afin de rendre la détection robuste au bruit, l'analyse des mouvements 3D des caractéristiques est comparée à d'autres points précédemment estimés statiques. Une mesure de confiance, mise à jour au cours du temps, est utilisée pour déterminer l'étiquette à attribuer à chacun des points.

Nos contributions ont été appliquées à des jeux de données virtuelles (issus du projet Previz [2]) et réelles (reconnus dans la communauté [Och+14]) et les comparaisons ont été réalisées avec l'état de l'art. Les résultats obtenus montrent que la contrainte 3D proposée dans cette thèse, couplée à une analyse statistique et temporelle des mouvements, permet de détecter des éléments mobiles dans le flux vidéo d'une caméra en mouvement et ce même dans des cas complexes où les mouvements apparents de la scène ne sont pas uniformes.

Mots clés : Vision par ordinateur, détection d'objets mobiles, caméra mobile, points caractéristiques, contrainte géométrique 3D.

2. http ://previz.eu

Remerciements

Je remercie les membres du jury pour leurs questions et leurs remarques et tout particulièrement Pascal Bertolino et Frédéric Cordier pour leur travail de rapporteur sur le manuscrit de thèse.

J'adresse un grand merci à tous les doctorants et doctorantes qui ont partagé mon bureau. Mention spéciale au Girls' Office (trois filles dans un bureau de quatre places, c'est assez rare pour le mentionner !), sans qui je n'aurais peut être (certainement !) pas tenu le coup : Hélène, Karolina, Simon, Elsa, Jonathan, Hoang, Raafat. Une petite pensée également aux anciens membres du LIRIS 1 (R.I.P.) pour vos batailles de Nerf et votre bonne humeur : JD, Lérémy, Gérémy, Abdou, François, Matthieu.

Je remercie également tous les habitués du coin café du Nautibus pour les échanges scientifiques ou non (le débat git vs svn reste ouvert !), mais aussi pour les échanges de gâteaux (la pause goûter à 16h30 c'est sacré). Elodie, Lionel, Stéphanie, David, Guillaume, Julie, Nicolas, Thierry, Eliane, Amélie.

Je remercie également ma famille, mes parents qui m'ont donné les moyens de faire mes propres choix et qui m'ont toujours soutenue. J'adresse une mention particulière à Joseph Garnier qui a dû me supporter durant ces quatre années de thèse... mais puisqu'il est aussi (encore ?) doctorant on est quitte !

Pour finir, je remercie sincéremment Mohand-Hacid Saïd, sans qui cette thèse n'aurait certainement jamais aboutie.

Table des matières

1 Introduction

La détection d'objets mobiles est une étape clé de nombreux algorithmes de vision par ordinateur tels que la vidéo surveillance ou encore l'analyse du mouvement humain. Elle consiste à identifier dans le flux vidéo d'une caméra un mouvement physique réalisé dans un environnement 3D. Cette tâche est d'autant plus difficile dans des cas complexes tels que des changements d'illuminations brusques, des environnements dynamiques ou encore avec une caméra mobile.

Contexte de Previz

Le travail de cette thèse s'inscrit dans cadre du projet Previz [1] dont l'objectif est de fournir au réalisateur une prévisualisation en temps réel des effets spéciaux. La création d'un film passe par trois grandes étapes : la pré-production, la production et la post-production. Le travail effectué en pré-production concerne toute la préparation nécessaire au tournage : écriture du scénario, repérage des lieux de tournage, casting, etc. La production représente la phase de tournage du film à proprement parlé. Le montage des différentes prises de vues réalisées durant la phase de production est réalisé en post-production. Cela comprend également l'ajout de contenu tel que de la musique ou encore des effets spéciaux.

L'utilisation d'effets spéciaux est une pratique courante dans le monde cinématographique. Ils créent une illusion afin de modifier la réalité en utilisant divers procédés tels que le maquillage, les effets mécaniques ou encore les effets numériques. Bien que limités par la technologie, les outils informatiques permettent de créer du contenu de nature diverse et de l'intégrer dans une séquence vidéo manuellement ou semi-automatiquement. Un exemple très connu d'intégration semi-automatique est la *capture de mouvement*. La *capture de mouvement* (*motion capture* ou *mocap* en anglais), est un procédé très utilisé dans le monde cinématographique qui permet de capter le mouvement d'un sujet (un être humain ou un objet), pour l'appliquer à un sujet virtuel dans le but d'obtenir une animation réaliste. Les origines de ce procédé remontent au XIXe siècle avec Eadweard Muybridge et Etienne-Jule Marey pour leurs travaux sur la décomposition du mouvement avec la *chronophotographie*. Les techniques de reconstitution du mouvement à partir de dessins ou de photographies

1. http ://previz.eu

développées au XIXe siècle sont regroupées sous le terme de *précinéma*. L'étude du mouvement du corps humain a permis des avancées scientifiques dans le domaine de la santé et de l'anatomie puis elle a été utilisée au cinéma à partir des années 1990 sous le nom de *capture de mouvement*. Au cours des années, la capture de mouvement s'est développée et aujourd'hui divers procédés permettent de capter des mouvements : optique, mécanique, gyroscopique, magnétique. Couplées à un ordinateur, les informations fournies par ce système d'acquisition permettent d'automatiser les étapes d'analyse des images dans le but de rendre le processus autonome.

Durant la phase de production, le réalisateur doit être capable de se représenter mentalement le rendu final qu'il souhaite obtenir afin de guider ses équipes. Ce n'est qu'en post-production qu'il peut alors voir si les résultats obtenus correspondent à ses attentes. En proposant une prévisualisation des effets spéciaux durant la phase de production, le réalisateur peut réaliser des réajustements directement pendant le tournage favorisant sa créativité et permettant également d'améliorer le processus de production des films à effets spéciaux.

L'intégration des effets spéciaux nécessite de coupler spatialement et temporellement des éléments virtuels qui constituent les effets spéciaux avec les éléments réels constitués du décor et des acteurs pour obtenir un rendu final cohérent. Parmi les différents effets spéciaux, nous nous concentrons ici sur ceux qui doivent interagir avec le jeu des acteurs. Nous distinguons deux types d'acteurs : ceux qui sont voués à être totalement ou partiellement remplacés par du contenu virtuel, comme un avatar et ceux qui interagissent avec le virtuel en apparaissant dans leurs costumes à l'écran (cf. figure 1.2). Dans le premier cas, l'acteur porte des capteurs qui permettent de connaître la position de ses membres dans les images. La pose de l'avatar virtuel peut être automatiquement copiée sur celle de l'acteur afin d'obtenir un premier rendu avec effets spéciaux. Dans le second cas, aucune information sur la pose de l'acteur n'est fournie et nécessite donc une intégration manuelle des effets spéciaux effectuée en post-production. Dans ces circonstances, il est primordial de détecter l'acteur dans le flux vidéo pour pouvoir capturer son mouvement automatiquement. La figure 1.1 présente les différentes étapes du processus d'intégration de ce type d'éléments virtuels. La première étape, qui est l'*identification des éléments clés*, sera l'objet des contributions présentées dans ce document.

Les travaux réalisés au cours de ce travail de thèse se concentrent sur l'étape des éléments clés à identifier. Dans le cadre du projet Previz, les éléments clés sont les acteurs. Nous avons choisi de généraliser l'étape d'identification à l'ensembles des objets mobiles comme les animaux ou encore les voitures.

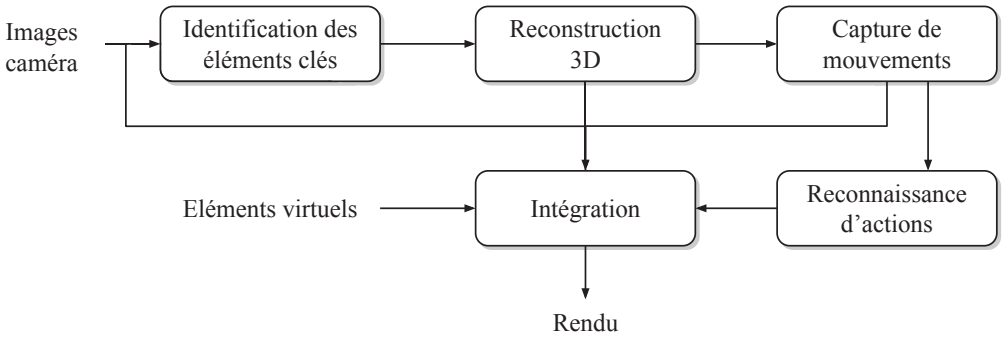


Fig. 1.1.: Schéma du processus d'intégration des effets spéciaux par rapport au jeu des acteurs.

Nous avons choisi de généraliser cette étape d'identification, qui se concentre dans le cadre du projet Previz la détection d'acteurs, à l'ensemble des objets mobiles qu'ils soient rigides ou non (par exemple des voitures).

Problématique

Le type de caméra utilisée pour réaliser la détection d'objets mobiles dépend du domaine d'application et sont soit statiques soit mobiles. Une séquence vidéo captée par une caméra statique présente des caractéristiques différentes de celle captée par une caméra mobile. Dans le flux vidéo d'une caméra stationnaire, les éléments statiques conservent leurs positions ainsi que leurs apparences sauf en cas d'événements externes tels que des changements d'illumination ou le déplacement d'un objet statique par un humain. L'apparence, la forme et la position des éléments mobiles varient quant à eux en fonction de leurs déplacements, des occultations et de leurs poses. A contrario dans le flux vidéo d'une caméra mobile tous les éléments, qu'ils soient statiques ou non, se comportent comme des éléments mobiles : ils changent de position et de forme dans les images, ils peuvent être occultés par d'autres éléments, ils peuvent apparaître et disparaître du champ de vision de la caméra et leurs apparences varient en fonction de leurs positionnements et leurs poses par rapport à la caméra.

Les êtres humains sont capables de détecter les objets/sujets en mouvement alors qu'ils réalisent eux-mêmes un déplacement dans l'espace. La tâche est beaucoup plus complexe à réaliser en traitant avec un processus informatique le flux vidéo d'une caméra mobile où tous les éléments apparaissent en mouvement. Les mouvements observés sont appelés *mouvements apparents* et ils dépendent du mouvement de la caméra ainsi que la distance 3D entre les éléments et la caméra. En fonction de ces deux éléments, les mouvements apparents des éléments statiques peuvent être

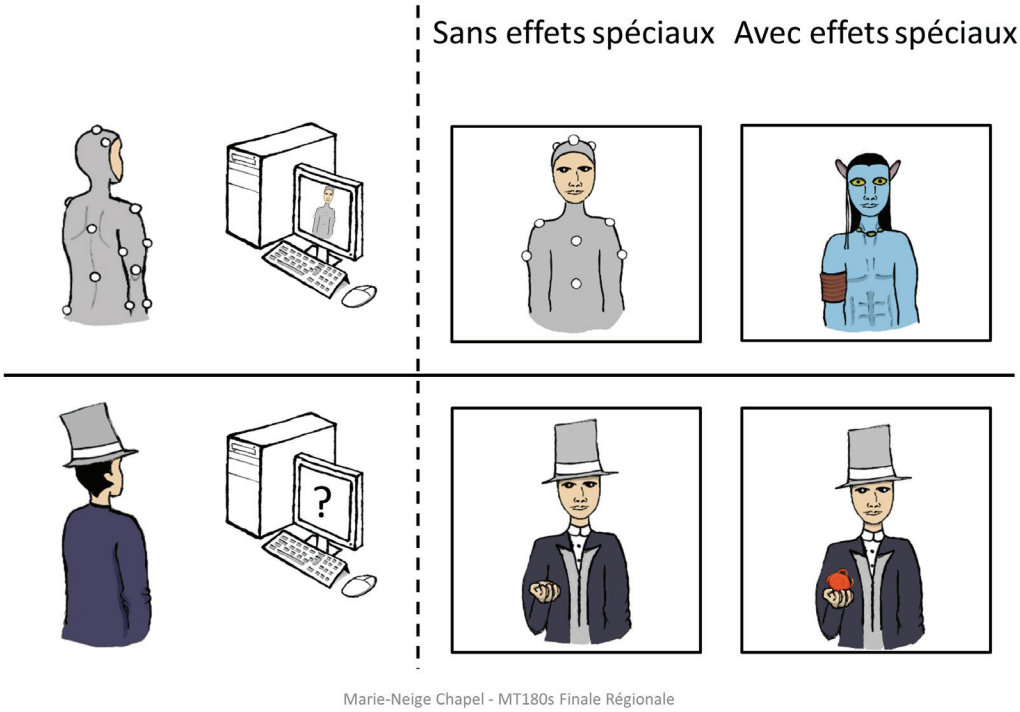


Fig. 1.2.: Deux types d'acteurs : avec et sans marqueur (diapositive du concours "Ma Thèse en 180 secondes", 2015).

uniformes ou totalement différents. Dans le premier cas, le mouvement apparent d'un objet mobile se détache généralement de celui de la partie statique de la scène. Dans le second cas, les mouvements apparents ne sont plus uniformes et il est même possible d'observer des mouvements contraires pour les éléments statiques (cf. figure 1.3). Les mouvements apparents des éléments mobiles ne sont plus discriminants et il est plus difficile de les détecter.

Les contributions qui ont été apportées au cours de ce travail de thèse apportent de nouvelles contributions à la détection d'objets mobiles dans le flux vidéo d'une caméra en mouvement. Nos premiers travaux portent sur la discrétisation de l'espace visible par la caméra à un instant t par un ensemble de plans. Au cours du temps, un point 3D peut se déplacer de deux manières : soit il se déplace dans un des plans de discrétisation de l'espace, soit il change de plan. En supposant que la caméra effectue un mouvement fluide et relativement lent, plusieurs seuils sont définis sur les deux types de déplacement afin de différencier les points qui appartiennent à des objets mobiles de ceux qui sont statiques. Nous avons par la suite proposé de nouvelles contributions qui s'appuient sur un ensemble constitué de points caractéristiques étiquetés statiques et qui évolue au cours du temps. Les déplacements 3D de tous les points caractéristiques sont comparés à ceux des points de l'ensemble stable statique, assimilé à un corps rigide, dans le but d'estimer leur mobilité. Une mesure statistique

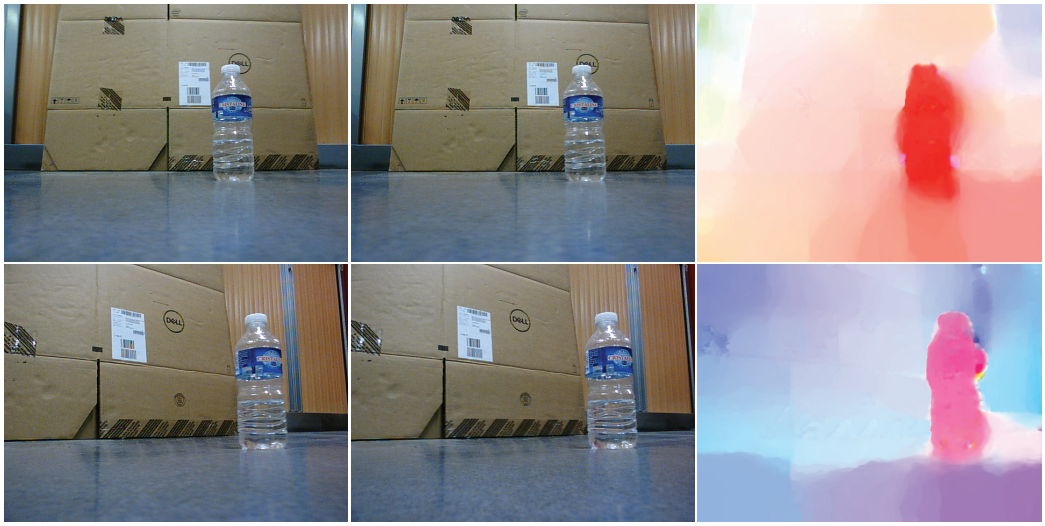


Fig. 1.3.: Différence de mouvements apparents d'une scène statique en fonction du mouvement de la caméra. Chaque ligne représente un mouvement de caméra différent et est composée de deux images prises à deux instants différents ainsi que d'une image de flot optique dense. La première ligne présente un mouvement de translation de la droite vers la gauche de la caméra. La seconde ligne présente le même mouvement caméra auquel un mouvement de rotation de la gauche vers la droite a été ajouté.

de confiance est estimée au cours du temps en fonction des déplacements relatifs des points caractéristiques par rapport à l'ensemble stable statique. Cette mesure est ensuite seuillée afin d'étiqueter ces points comme statiques ou mobiles. Pour obtenir un étiquetage plus robuste, les mouvements apparents 2D sont utilisés pour réduire les erreurs dues au bruit.

Organisation du document

Le présent document est organisé comme suit. Le chapitre 2 présente un état de l'art des méthodes de détection d'objets en mouvement. Les chapitres 3 et 4 présentent les travaux réalisés au cours de cette thèse. Le chapitre 5 présente une synthèse des travaux réalisés ainsi que des perspectives de recherche liées aux contributions proposées.

Notations

p	Point caractéristique 2D.
f	Flot optique 2D.

P	Point caractéristiques 3D.
$\widetilde{P}$	Point caractéristiques 3D estimé.
π	Plan 3D.

^{t}	Représente le temps courant et par abus de langage le numéro de l'image courante dans la séquence vidéo.
$_i$	Représente le numéro du point caractéristique.

I	Image caméra.
$\mathcal{C}$	Caméra.
$\mathcal{N}$	Ensemble stable de points caractéristiques statiques assimilé à un corps rigide.
$\mathcal{R}$	Repère.

2 État de l'art

Sommaire

2.1 Extraction de sujets mobiles dans le flux vidéo d'une caméra

Une scène filmée par une caméra est constituée de deux types d'éléments : des objets statiques et des objets/sujets mobiles. Dans les séquences vidéos sur lesquelles sont généralement appliquées les algorithmes de détection d'objets en mouvement, on remarque que les éléments mobiles évoluent devant la partie statique de la scène. Cette disposition en profondeur dans l'espace est définie dans le langage cinématographique par les termes *arrière plan* et *premier plan* (ou *fond*). Cette terminologie est également utilisée dans la littérature pour faire référence respectivement à la partie statique de la scène et aux sujets mobiles.

Les techniques présentées dans ce chapitre sont des techniques de détection d'éléments mobiles optiques qui s'appuient sur des informations extraites du flux vidéo de la caméra ainsi que sur des informations calculées a priori. Ces informations sont de natures diverses : apparence de la scène, mouvements apparents, paramètres de la caméra, etc. et permettent de distinguer le *premier plan* du *fond*. Dans la suite de ce chapitre, les méthodes ont été réparties en trois grandes catégories basées sur le type de caméra utilisée : caméra statique, caméra mobile avec mouvements contraints et caméra mobile sans mouvement contraint.

2.1.1 Les caméras fixes

Dans le but de détecter un objet en mouvement dans le flux vidéo d'une caméra statique, de nombreuses techniques s'appuient sur le fait que la partie statique de la scène reste inchangée durant la prise de vue et que les changements observés proviennent d'un objet en mouvement. Il est possible de mettre en évidence ces modifications et donc extraire le sujet en mouvement, en appliquant une soustraction entre l'image courante et une image qui ne contient pas d'objet en mouvement, c'est-à-dire une image de la scène. Les techniques de ce type sont regroupées sous le terme de *soustraction de fond* (*background subtraction*) dont le processus général est représenté par la figure 2.1. Le terme *fond* est employé pour la scène par opposition au terme *premier plan* qui lui désigne les sujets mobiles à détecter.

Le fond n'est pas décrit par une simple image, mais par un modèle statistique calculé sur plusieurs images provenant du flux vidéo de la caméra. Afin de palier aux divers changements que pourrait subir la scène durant la captation, ce modèle est mis à jour à chaque nouvelle image prise par la caméra. Le résultat de la soustraction entre l'image courante et le modèle de fond est un masque contenant les sujets en mouvement, appelé masque de premier plan.

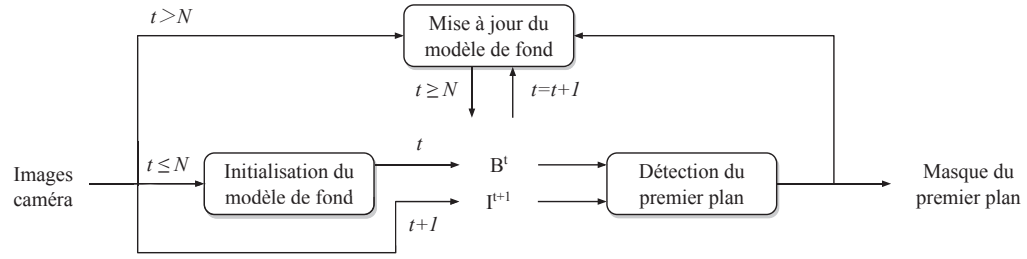


Fig. 2.1.: Processus de la soustraction de fond pour une caméra fixe (figure 1 de [Bou14]). N est le nombre d'images utilisées pour initialiser le modèle de fond B_t. I_t est l'image au temps t.

La méthode des mélanges de Gaussiennes (Mixtures of Gaussian ou MOG) proposée par Stauffer et Grimson [SG99] est la technique la plus utilisée pour modéliser le fond. Pour cela, chaque pixel de l'image est caractérisé par plusieurs Gaussiennes qui représentent sa probabilité d'observation :

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{2.1}$$

avec K le nombre de Gaussiennes, ω une pondération, Σ_i la matrice de covariance de la i^e Gaussienne μ_i est la valeur moyenne de la i^e Gaussienne et η est une fonction de densité de probabilité Gaussienne :

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-1/2(X_t-\mu_t)^\top \Sigma^{-1}(X_t-\mu_t)} \tag{2.2}$$

Une fois le modèle de fond initialisé, les pixels de l'image courante sont comparés avec ceux du modèle de fond. Dans le cas où le pixel correspond à l'une des Gaussiennes du modèle de fond, le pixel est considéré comme appartenant au fond et le modèle est mis à jour avec cette nouvelle donnée. Dans le cas où le pixel ne correspond à aucune Gaussienne, le pixel est classé comme appartenant au premier plan et la distribution la moins probable est remplacée par une nouvelle calculée sur le pixel correspondant (cf. figure 2.2). Ce modèle présente deux inconvénients majeurs : le premier est qu'un objet en mouvement qui s'immobilise pendant un certain temps sera intégré au modèle de fond, le second est l'aspect *pixel-wise* de la méthode qui traite les pixels indépendamment les uns des autres. Cet aspect ne permet pas de gérer les cas critiques tels que les changements d'illumination qui nécessitent une détection spatiale et/ou temporelle.

Harville et al. [Har+01] augmentent le *taux d'apprentissage* afin d'incorporer plus rapidement des objets statiques qui auraient été déplacés dans la scène, mais ce phénomène se produit également lorsque des sujets en mouvement s'arrêtent

(a) (b) (c)

Fig. 2.2.: La soustraction de fond par mélange de Gaussiennes. (a) L'image courante, (b) le modèle de fond, (c) le masque des pixels de premier plan. (Figure 1 de [SG99])

ponctuellement. Pour éviter cela, les auteurs proposent de mesurer l'activité de la scène pour chaque pixel en calculant les différences d'illumination au cours du temps. Les pixels qui présentent une forte activité appartiennent principalement à des éléments mobiles et voient de ce fait leur taux d'apprentissage réduit afin de limiter leur incorporation dans le modèle de fond. Wang et al. [WS05] utilisent des informations spatiales tandis que Porikli et al. [PT03] calculent un score de changement d'illumination pour adapter le taux d'apprentissage et donc gérer ces changements. Zhang et al. [Zha+05] distinguent deux cas pour lesquels un pixel a été étiqueté *premier plan* : soit il correspond à un objet mobile, sa valeur changera au cours du temps et il ne sera naturellement pas intégré au modèle de fond, soit il représente un objet mobile temporairement statique ou un nouvel objet statique introduit dans la scène et il sera intégré au modèle de fond seulement après un certain temps. Pour incorporer plus rapidement au modèle de fond la deuxième catégorie d'objets, sans avoir a manipuler le taux d'apprentissage, les auteurs utilisent un historique calculé sur le nombre de fois où un pixel a été étiqueté premier plan consécutivement. Dans le cas où cet historique d'étiquetage dépasse un certain seuil, la probabilité du modèle de fond actuel du pixel est abaissée au profit de la nouvelle probabilité qui représente le nouvel objet. D'autres méthodes présentées dans l'étude de Bouwmans [Bou+08] proposent également d'améliorer l'algorithme de Stauffer et Grimson [SG99] sur la rigueur statistique et en introduisant des contraintes spatiales et temporelles.

Une autre manière de représenter le fond est d'utiliser un sous-espace afin de réduire la dimension des données. L'une des méthodes les plus connues est celle de l'*Analyse en Composante Principale* (PCA) introduite par Oliver et al. [Oli+99]. Le modèle de fond est constitué d'un certain nombre de valeurs propres qui définissent un sous-espace vectoriel. La détection des objets en mouvement s'effectue par une projection de la nouvelle image dans le sous-espace puis par une différence entre l'image projetée et le modèle de fond. Pour détecter correctement les sujets en mouvement, il faut que ces derniers ne soient pas dominants dans l'image et qu'ils ne soient pas stationnaires pendant un long moment au risque d'être intégrés au modèle de fond. De plus, l'algorithme de mise à jour du fond a une complexité

élevée et est limité aux images en niveaux de gris. Par la suite, plusieurs méthodes ont été proposées pour réduire ces limitations telles que la méthode de Xu et al. [Xu+08] qui utilise une compensation d'erreur pour réduire l'influence du premier plan, Skocaj et al. [SL03] qui augmente la robustesse du modèle de fond ou encore Han et al. [HJ07] qui utilisent une PCA à deux dimensions qui est une adaptation de la PCA pour utiliser des images couleurs.

La méthode de la PCA est peu robuste face au bruit. Pour remédier à ce problème, la technique de la *Robust PCA* (RPCA) proposée par Candes et al. [Can+11] décompose les images de la séquence vidéo représentées par une matrice de données A en deux composants $A = L + S$ tels que L est une matrice de faible rang et S une matrice éparse de bruit. L représente la partie statique de la scène, i.e. le fond tandis que S contient les *outliers* qui sont les objets en mouvement et le bruit. Afin de séparer les deux matrices, Candes et al. [Can+11] proposent l'algorithme du *Principal Component Pursuit* (PCP). Mais cet algorithme présente plusieurs inconvénients tels que des temps de calcul élevés ou encore la perte des informations spatiales et temporelles. Bouwmans [BZ14] propose une étude comparative des algorithmes RPCA avec une résolution par l'algorithme du PCP.

Les algorithmes de détection d'objets en mouvement dans le flux vidéo d'une caméra statique sont largement étudiés depuis plusieurs dizaines d'années et plusieurs études proposent de catégoriser et de comparer les différentes approches existantes [Bou14; JS15]. Ces approches ne peuvent pas être directement appliquées au cas d'utilisation d'une caméra mobile puisqu'elles ne prennent pas en compte le changement d'apparence du fond dû à un changement de point de vue. Des adaptations de ces méthodes ont toutefois été proposées pour les caméras mobiles avec et sans contrainte de mouvement.

2.1.2 Les caméras à mouvement contraint

Parmi les caméras mobiles, certaines ont des contraintes physiques qui réduisent leur possibilité de mouvement. C'est le cas des caméras Pan Tilt Zoom (PTZ) qui peuvent, comme leur nom l'indique, effectuer des rotations selon l'axe des X (tilt) et des Y (pan) ainsi que des zooms. Leur centre optique est fixe ce qui les empêche de faire des translations. Ce type de caméra est généralement utilisé pour la vidéo-surveillance et l'utilisation d'algorithmes de détection d'objets en mouvement correspond parfaitement à leur utilisation. Puisque les mouvements de ces caméras sont entièrement automatisés, il est possible non seulement de détecter un sujet en mouvement mais également de le suivre, c'est-à-dire de le garder dans le champ de vision de la caméra en actionnant les moteurs. Everts et al. [Eve+07] se basent sur la couleur pour suivre la cible en utilisant plusieurs caméras. De la même manière

Varcheie et al. [VB11] utilisent la détection de mouvement pour obtenir des régions candidates au premier plan, de l'échantillonage sur les régions candidates, la couleur ainsi qu'une classification pour détecter la cible à chaque nouvelle frame. Mian [Mia08] propose de détecter et de suivre des visages en utilisant des caractéristiques pseudo-Haar et l'algorithme du *CAMSHIFT* (Continuously Adaptive Mean Shift). Chen et al. [Che+08] utilise deux caméras afin de calculer l'orientation des caméras pour suivre un sujet en mouvement. Par la suite, il proposera une méthode qui permet de réaliser le suivi sans connaître les paramètres intrinsèques des caméras a priori [Che+09]. Ce type de méthodes permet de suivre un sujet en mouvement en l'identifiant généralement par une *bounding box* mais elles ne permettent pas d'identifier les pixels qui appartiennent au sujet.

L'avantage de ce type de caméra est que leur champ visuel est restreint puisque leur centre optique est fixe, donc pas de changement de position mais seulement des changements de direction. Partant de ce constat, il est possible de construire une vue panoramique de la zone observable par la caméra qui représente le modèle de fond au même titre que pour les caméras statiques. Le modèle de fond panoramique est une mosaïque d'images prise par la caméra selon différents points de vues. Ces images sont ensuite alignées entre elles pour constituer une seule grande image. Le fond peut ainsi être modélisé de la même manière que pour des caméras statiques, généralement par des mélanges de gaussiennes. Une fois le modèle de fond construit, l'image prise par la caméra à l'instant t doit être comparée à celle du fond afin d'extraire la silhouette de l'élément mobile. Pour cela, il faut recaler l'image courante dans celle du panorama. La figure 2.3 présente le schéma général des méthodes de soustraction de fond avec des caméras PTZ.

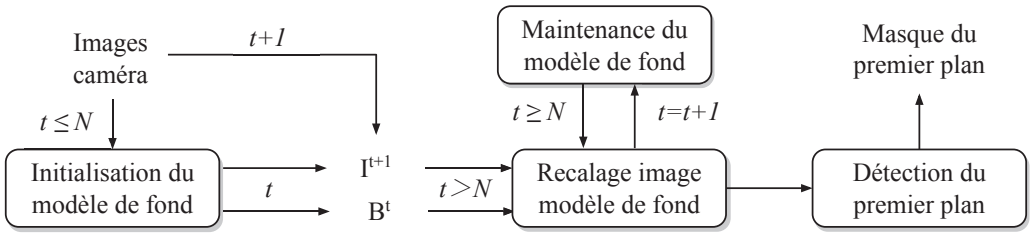


Fig. 2.3.: Processus de la soustraction de fond pour une caméra PTZ. N est le nombre d'images utilisées pour initialiser le modèle de fond B^t. I^t est l'image au temps t.

Mittal et al. [MH00] proposent de modéliser chacun des pixels du panorama par un mélange de gaussiennes. La transformation affine qui permet de recaler l'image sur le fond est calculée à partir de la mise en correspondance de points caractéristiques, puis affinée par la méthode de Levenberg-Marquardt. Chacun des pixels de l'image courante est ensuite comparé avec celui du modèle de fond dont le mélange de gaussienne est la plus proche pour faire une soustraction de

fond. Une recherche dans le voisinage proche permet de rectifier les petites erreurs d'alignement. Le modèle de fond est mis à jour avec les informations de l'image courante. Certaines techniques [Bha+00 ; Kan+03 ; HE03] utilisent les angles de la caméra pour recaler les images sur le panorama. Bevilacqua et al. [Bev+05] proposent une méthode qui permet de construire le panorama en temps réel au fur et à mesure de la prise de vue et non par le biais d'une étape d'initialisation off-line. Les images sont alignées en calculant les homographies sur des points caractéristiques extraits et suivis par l'algorithme KLT [ST94]. Cependant, les points caractéristiques qui appartiennent à un élément mobile ne doivent pas être pris en compte lors de l'alignement des images puisque le recalage doit s'effectuer sur la partie statique de la scène dans le but d'extraire les sujets en mouvement. Pour cela, les auteurs ont choisi de partitionner les points caractéristiques présents sur le bord de l'image en utilisant les similarités des trajectoires. Lorsque l'image est correctement alignée avec le modèle de fond, elle est intégrée dans ce dernier pour la mise à jour en utilisant la règle de mise à jour α :

$$B_x^t = (1 - \alpha)B_x^{t-1} + \alpha I_x^t \tag{2.3}$$

avec B_x^t le modèle de fond pour le pixel x au temps t et I_x^t le pixel x dans l'image I.

Au lieu de construire un panorama, Cui et al. [Cui+14] suggèrent d'utiliser deux caméras : une avec un grand angle, statique et à faible résolution et une autre mobile (PTZ) et à haute résolution. La caméra grand angle remplace la construction de la mosaïque et permet de maintenir l'intégralité du modèle de fond à jour à chaque nouvelle frame. Une difficulté survient lors du recalage des deux images du fait que les caméras n'ont pas les mêmes champs de vues ni les mêmes résolutions. Pour palier à cela, les auteurs proposent un recalage en trois étapes : une région grossière est définie dans l'image de fond en se basant sur la trajectoire de l'élément mobile calculée par *mean-shift* qui correspond à la portion de la scène observée par la caméra PTZ ; une transformation affine est calculée par *Least Square Approximation* (LSA) via une mise en correspondance de caractéristiques *Speeded Up Robust Features* (SURF) sur laquelle est appliquée l'algorithme du *RANdom SAmple Consensus* (RANSAC) pour la rendre plus robuste ; la transformation est ensuite affinée par la méthode de *sum squared differences* (SSD) (cf. figure 2.4).

Xue et al. [Xue+13] utilisent une seule caméra PTZ et construisent un panorama modélisé par un mélange de Gaussiennes. Pour recaler l'image courante sur le panorama, les auteurs utilisent des points caractéristiques SURF. Le descripteur SURF est invariant à l'échelle et plus la résolution de l'image est haute, meilleure est l'invariance. Xue et al. utilisent une caméra dont la résolution est relativement petite ce qui limite l'invariance du descripteur. Cet effet se répercute lors de la mise en

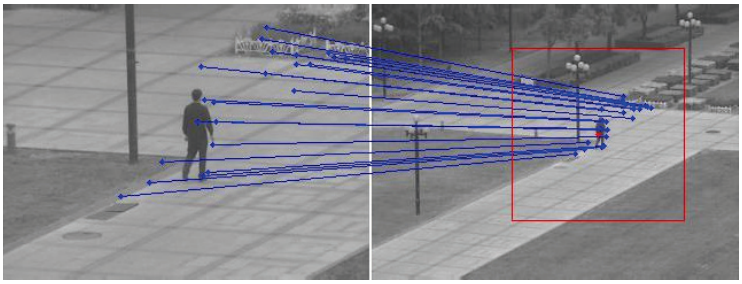


Fig. 2.4.: Mise en correspondance des points caractéristiques entre la caméra PTZ et la caméra statique en bleue. Le carré rouge délimite la région grossière dans la caméra statique. (figure 4 de [Cui+14])

correspondance des points caractéristiques entre deux images à différentes échelles. Plus la différence d'échelle sera élevée, moins il y aura de points caractéristiques mis en correspondance. Pour pallier cela, les auteurs proposent de créer une hiérarchie d'image prises à différents niveaux de zoom. Un niveau regroupe les images d'un même niveau de zoom et les niveaux sont liés entre eux par la mise en correspondance de points caractéristiques dans les images. Cette hiérarchie construite off-line permet de faire correspondre l'image courante avec une image de la hiérarchie pour remonter ensuite jusqu'au panorama.

Au lieu de maintenir un modèle de fond sous forme de panorama, d'autres méthodes se contentent d'un modèle de fond de même résolution et de même angle d'ouverture que l'image courante [MB94 ; Rob+09 ; Kad+13]. L'image courante est recalée sur l'image précédente et le modèle de fond est maintenu à jour à chaque nouvelle frame. Murray et al. [MB94] calculent la correspondance des pixels entre deux images en utilisant l'orientation de la caméra qui est connue. Robinault et al. [Rob+09] et Kadim et al. [Kad+13] calculent une homographie pour recaler les deux images. Ce type de méthodes a l'avantage de ne pas avoir à besoin d'une étape d'initialisation pour la construction du panorama contrairement à celles qui en utilisent un. De plus, le maintient à jour du modèle de fond est moins lourd et plus robuste puisqu'il ne couvre que la partie visible par la caméra et ne souffre donc pas de modifications qui pourraient survenir hors champ.

L'avantage du type de caméra étudié dans cette section est la contrainte de mouvement qui permet de simplifier le problème. En effet, puisqu'il n'y a pas de translation il n'y a pas de parallaxe à prendre en compte. Cependant, elles ne peuvent pas être appliquée aux séquences captées par des caméras sans contrainte de mouvement qui ont de la parallaxe.

2.1.3 Les caméras mobiles

Cette catégorie regroupe les méthodes qui s'appliquent à des caméras qui n'ont aucune - ou quasiment aucune - contrainte physique de mouvement. Dans cette section, les différentes méthodes utilisées ont été reparties dans cinq catégories. Chaque catégorie représente une classe de méthodes caractérisée par sa représentation de l'environnement de captation.

2.1.3.1 Compensation du mouvement de la caméra

De la même manière que pour les caméras statiques et pour les caméras PTZ, les techniques usuelles de soustractions de fond peuvent être utilisées avec des caméras mobiles. Puisque la caméra est en mouvement, il faut dans un premier temps recaler l'image avec celle du fond pour pouvoir extraire la silhouette du sujet en mouvement (cf. figure 2.5), comme pour les PTZ.

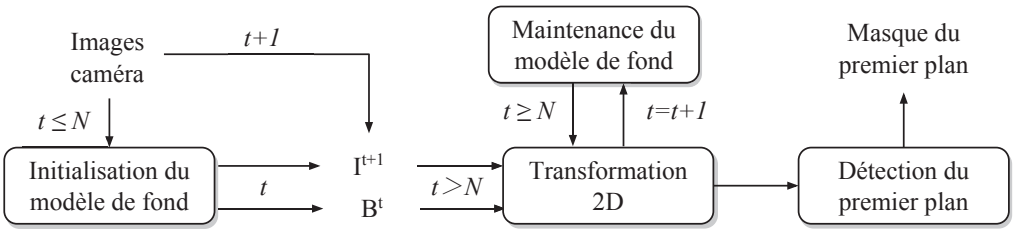
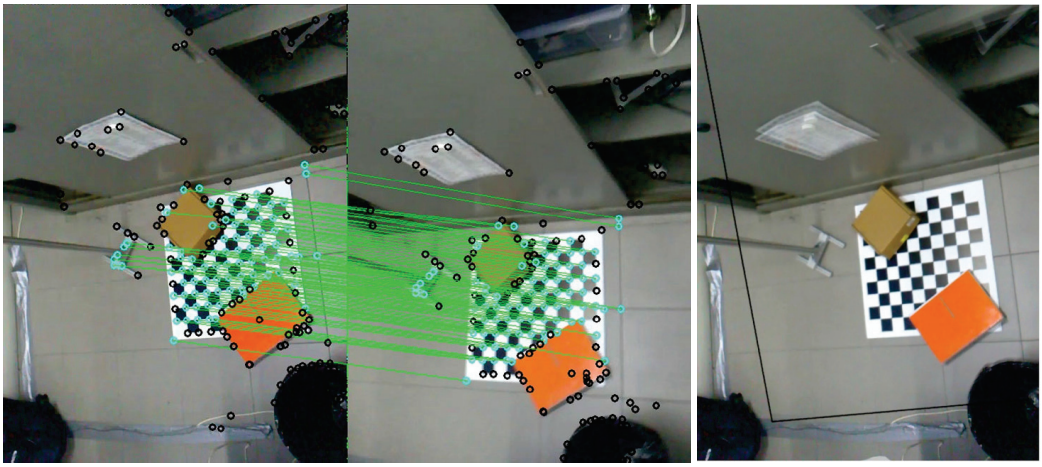


Fig. 2.5.: Processus de la soustraction de fond pour une caméra mobile. N est le nombre d'images utilisées pour initialiser le modèle de fond B_t. I_t est l'image au temps t.

Les techniques utilisées pour recaler les images captées par la caméra mobile au cours du temps sont généralement des homographies ou des transformations affines 2D. Ce type de transformation établit une relation linéaire entre deux images d'une même surface planaire dans l'espace 3D. Utiliser de telles transformations pour compenser le mouvement de la caméra revient à approximer la partie statique de la scène par un plan unique. Cette hypothèse est vérifiée lorsque la caméra effectue des rotations sans translation ou lorsque la scène est suffisamment éloignée de la caméra, c'est le cas par exemple avec des images aériennes. Dans le cas contraire, les parties de la scène statique qui résultent de la parallaxe ne sont correctement recalées comme le montre la figure 2.6 et des faux négatifs apparaissent après l'étape de soustraction de fond.

Viswanath et al. [Vis+15] utilisent les points caractéristiques étiquetés comme fond pour calculer une homographie qui permet de compenser le mouvement de la caméra entre l'image courante et le modèle de fond. Une fois recalée, l'image



(a) (b)

Fig. 2.6.: Exemple d'erreur de recalage entre l'image courante et le fond par homographie. L'homographie calculée sur la mise en correspondance représentée par l'image (a) représente le plan du sol qui est parfaitement recalé sur le modèle de fond tandis que l'armoire présente un décalage dans la superposition (b) (figures 1 et 2 de [Rom+14]).

courante est comparée au fond, modélisé par des mélanges de Gaussiennes, en comparant les intensités des pixels et leurs voisinages. Kim et al. [Kim+13] modélisent le fond en utilisant une gaussienne par pixel et utilisent un modèle spatio-temporel pour le mettre à jour. L'aspect temporel du modèle est représenté par un âge associé à chaque pixel qui détermine le *taux d'apprentissage* du fond. L'âge correspond au nombre de frames successives dans lequel un point de la scène est apparu (cf. figure 2.7). L'image courante est recalée sur la précédente en calculant une homographie sur des points caractéristiques extraits et suivis par la méthode *Lucas Kanade Tracking* (LKT).

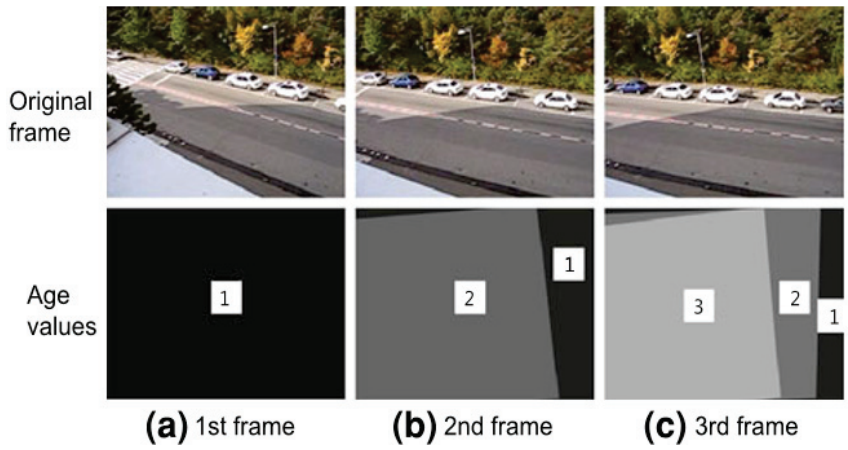


Fig. 2.7.: Valeur des âges au cours de la séquence vidéo. (figure 8 de [Kim+13])

Pour retirer un maximum de faux négatifs dans l'étiquetage final, le voisinage des pixels est pris en considération lors de la soustraction de fond. La prise en compte du voisinage peut cependant déformer et réduire la silhouette du sujet mobile lorsque ce dernier a une couleur similaire avec celle du fond. Pour pallier cela, les auteurs ajoutent une dernière étape qui suit et raffine la silhouette du sujet. Le suivi est réalisé par un correcteur *proportionel, intégrateur, dérivé* (PID) qui met en correspondance les silhouettes extraites à l'instant $t-1$ et t. Si la région suivie maintient son étiquette sur plusieurs frames, les auteurs considèrent qu'il s'agit d'un objet rigide et non de faux négatifs générés par un problème de recalage des images. De ce fait, pour extraire un maximum de points appartenant au sujet mobile, les voisins ne sont plus pris en compte dans l'étape de soustraction de fond. En plus de cela, un algorithme morphologique probabiliste permet de retrouver des parties du sujet dont la couleur est proche de celle du fond en modifiant l'étiquette des pixels dont la couleur du voisinage est plus proche des pixels détectés mobiles que de ceux du fond. De manière similaire à Kim et al. [Kim+13], Yi et al. [Yi+13] utilisent la notion d'âge ainsi qu'une seule gaussienne par opposition aux mélanges de gaussiennes. Au lieu de modéliser chacun des pixels, l'image courante est divisée en une grille et chacune des cases est décrite par un *Single Gaussian Model*. Le mouvement de caméra est alors compensé via des homographies calculées pour chacune des cases de la grille. Le modèle de fond est ensuite mis à jour en fonction des portions de la grille de l'image précédente recouvertes. Afin de prévenir une contamination du modèle de fond par des pixels appartenant au sujet en mouvement, les auteurs proposent de maintenir un deuxième modèle de fond candidat. Les modèles sont échangés lorsque le modèle candidat est plus âgé que le modèle courant. Romanoni et al. [Rom+14] utilisent une approche temporelle et spatio-temporelle pour corriger les écarts de recalage du fond et ils utilisent pour cela une représentation par histogramme pour décrire les distributions spatiales et temporelles des pixels. L'approche temporelle compare les intensités des pixels entre l'image courante et l'image recalée. L'approche spatio-temporelle compare quant à elle les différences d'intensité entre le pixel avec son voisinage et le modèle de fond en utilisant la distance de *Bhattacharyya*. Les histogrammes utilisés pour l'approche spatio-temporelle sont créés en additionnant les histogrammes du voisinage du pixel. La combinaison des deux approches permet de supprimer les faux négatifs de la classification. Wan et al. [Wan+14] utilisent dans un premier temps une méthode itérative pour classifier les points caractéristiques extraits des images par la méthode *KLT*. Deux images consécutives sont recalées par l'estimation des paramètres d'une transformation affine en utilisant la technique de *RANSAC* sur les points caractéristiques. Après recalage, les deux images sont soustraites pour extraire les parties de l'image qui ne sont pas recalées. La région extraite ne correspond pas totalement à un élément mobile puisqu'elle inclue les parties du fond qui étaient occultées par l'élément. Pour supprimer ces faux négatifs, deux modèles de mélanges de Gaussiennes sont calculés à partir des points précédemment étiquetés et sont utilisés pour mettre à jour la probabilité que chaque point caractéristique

appartienne à l'élément mobile. Une fois les étiquettes raffinées, les paramètres de la transformation affine entre les deux image sont de nouveau estimés en se basant uniquement sur les points étiquetés comme appartenant au fond. Ce processus se répète jusqu'à obtenir une convergence. Lorsque le procesus d'étiquetage des points caractéristiques se termine, l'image courante est segmentée en appliquant l'algorithme du *graph-cut* géodésique qui combine la segmentation géodésique et l'algorithme du *graph-cut* pour obtenir une segmentation du sujet plus précise. Jung et al. [JS04] utilisent le *Frame Differencing* couplé à un filtre à particules adaptatif pour détecter les sujets en mouvement. Grâce à cela, la direction et la vélocité du sujet dans l'image sont estimées. Zhou et al. [Zho+13] utilisent une matrice de faible rang pour représenter le modèle de fond. Les objets en mouvements sont perçus comme un changement d'intensité qui ne peut correspondre à la représentation par la matrice de rang faible du modèle de fond.

Certaines méthodes s'appuient sur l'analyse des mouvements apparents pour séparer le sujet en mouvement de la scène statique. Liu et al. [LG09] considèrent que le mouvement apparent dominant observé dans les images caméra est celui du fond et qu'il est différent de celui du sujet mobile. Le calcul de la discrépance de flot optique dans chaque paire d'image permet de mettre en évidence les zones mobiles. Sur un sujet humain les zones mobiles ne représentent que certaines parties du corps, comme par exemple les mains d'un pianiste qui se baladent sur le clavier tandis que le reste de son corps reste globalement immobile. Une dernière étape est appliquée à la fin de la séquence vidéo pour extraire l'intégralité de la silhouette du sujet en segmentant les images sur la base de toutes les zones précedemment extraites. Schubert et al. [SM14] proposent de supprimer les faux négatifs qui apparaissent après le recalage de deux images consécutives en réalisant un filtrage en deux étapes. La première étape consiste à renforcer la contrainte de continuité de mouvement en effectuant un vote. Pour cela, une fenêtre temporelle est définie sur les masques obtenus après soustraction de fond et est centrée sur le masque de l'image courante. Chaque pixel qui figure comme appartenant à un sujet mobile dans le masque courant subit un processus de vote qui décide de modifier ou non l'étiquette attribuée au pixel. Pour réaliser ce vote, chacun des pixels appartenant à un élément mobile dans les masques de la fenêtre temporelle insère une valeur dans l'espace de vote qui dépend de la position de son masque dans la fenêtre. Plus le masque est éloigné du masque courant moins la valeur sera élevée (cf. figure 2.8). Un seuil est ensuite appliqué sur la plus grande valeur pour valider ou non l'étiquette du pixel testé.

La seconde étape du filtrage des faux négatifs vérifie la consistance des trajectoires obtenues par la calcul du flot optique sur les images antérieurs et postérieures à l'image courante dans le but de supprimer les points erronés. Une dernière étape d'opérations morphologiques est appliquée aux masques pour raffiner la silhouette

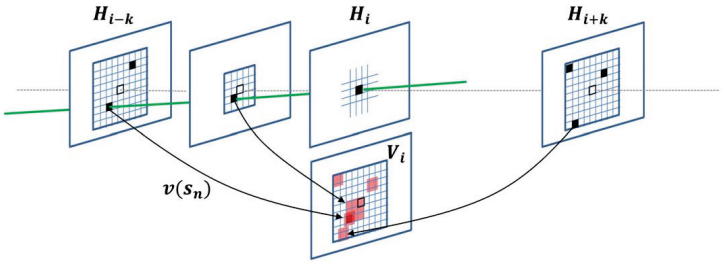


Fig. 2.8.: Système de vote avec H_i un masque obtenu par soustraction de fond, V_i l'espace de vote et $v(s_n)$ l'emplacement dans l'espace de vote (figure 5 de [SM14]).

des éléments mobiles en appliquant d'abord une érosion puis une dilatation. DeGol et al. [DN14] relient les points caractéristiques extraits d'une même image par une triangulation de Delaunay afin d'obtenir un graphe. Les arêtes du graphe sont pondérées en fonction de la dissimilarités des mouvements apparents. Les arêtes du graphe qui lient des points statiques à des points mobiles sont supprimées sur la base d'un seuillage calculé sur les poids des arêtes et les clusters résultants qui contiennent peu de points sont ceux qui correspondent à des sujets en mouvement. Hu et al. [Hu+15] utilisent la contrainte épipolaire pour étiqueter les points caractéristiques comme statiques ou mobiles. Les points étiquetés mobiles dans la précédente image sont également étiquetés mobiles dans l'image courante. Une homographie est ensuite calculée avec les points statiques pour recaler deux images consécutives, puis les pixels appartenant aux parties de l'image non recalées sont extraits par soustraction des deux images après recalage. Les auteurs calculent également le flot optique global des points caractéristiques statiques après recalage pour l'appliquer à l'image recalée afin de supprimer au maximum les erreurs de recalage. Les pixels extraits par *Frame Differencing* sont agglomérés en région puis une vérification est réalisée sur chacune des régions pour déterminer si elle appartient à un sujet en mouvement. Si une région contient des points caractéristiques étiquetés mobiles, alors cette région représente un sujet en mouvement, dans le cas contraire elle est supprimée. Le problème est que cette étape supprime également des régions correctes. Pour augmenter le nombre de régions à l'instant t, les régions qui ont été calculée sur les trois dernières images sont fusionnées et ajoutées à l'image courante. Afin d'améliorer la silhouette des sujets en mouvement, les régions mobiles sont sous-échantillonnées, les opérateurs de dilatation et d'érosion sont ensuite appliqués, puis un sur-échantillonnage est réalisé pour retrouver la taille originale de l'image. Kim et al. [Kim+16] utilisent un algorithme de *clustering* sur des trajectoires représentées dans un espace 3D dont les deux premières dimensions représentent le mouvement apparent et la troisième représente la distance à la caméra.

Les méthodes par compensation proposent une solution satisfaisante au problème de détection d'éléments mobiles dans le cas où la caméra se déplace dans une

zone suffisamment éloignée de la scène pour limiter au maximum les effets de parallaxe. En revanche, si de fortes parallaxes apparaissent dans les images, ce type de technique ne sera pas en mesure de différencier correctement les sujets mobiles des erreurs de recalages.

2.1.3.2 Plane+Parallax

Les méthodes connues sous le nom de *Plane+Parallax* sont très proches de la philosophie des méthodes dites par compensation du mouvement de la caméra dans le sens où elles estiment elles aussi une transformation géométrique 2D entre deux images basée sur le plan dominant de la scène. La différence est que les parties de l'image qui ressortent après soustraction de fond sont soit dues à la parallaxe soit dues à un sujet en mouvement. Pour différencier les deux, les méthodes s'appuient sur les distances entre les points 3D et le plan dominant estimés au cours du temps.

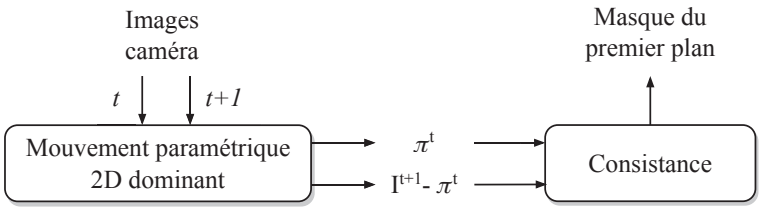


Fig. 2.9.: Processus général des méthodes *Plane+Parallax*. π^t est le plan dominant au temps t et I^{t+1} est l'image caméra au temps $t+1$.

La méthode proposée par Irani et al. [IA98] considère deux cas : soit la scène peut être approchée par un plan 2D, soit la scène contient des variations de profondeurs significatives. Dans le premier cas, on revient à la catégorie des méthodes par compensation où le mouvement de la caméra est compensé en calculant une transformation 2D entre deux images. Après recalage, les portions de l'image qui correspondent aux éléments statiques sont correctement alignées tandis que les objets en mouvement ne le sont pas. Lorsque la scène ne peut plus être approchée par un seul plan, une série de plans est estimée : la transformation 2D entre deux images est estimée puis les images sont recalées. Les parties des images qui ne sont pas correctement recalées sont récupérer pour réestimer une nouvelle transformation 2D. Ainsi la scène est représentée non plus par un mais par plusieurs plans. Les incohérences qui restent après recalage de tous les plans entre deux images proviennent des objets en mouvements. Dans le cas où la scène ne peut pas être approchée par plusieurs plans, la décomposition en *Plane+Parallax* est utilisée. La transformation 2D calculée pour recaler les images permet de compenser les mouvements de rotation et de zoom. Ainsi, les mouvements résiduels proviennent soient

de la parallaxe soit du sujet mobile. Ceux qui proviennent de la parallaxe sont dus au mouvement de translation de la caméra et forment un champ de trajectoires radiales centrées sur le focus d'expansion (*Focus Of Expansion*, FOE.) Grâce à cela, il est possible de différencier les mouvements générés par le mouvement de la caméra de deux des sujets mobiles. Toutefois, l'utilisation des mouvements apparents qui forment un champ radial peuvent biaiser l'estimation du FOE. En effet, si le nombre de mouvements apparents d'un sujet mobile est plus important que celui du fond et qu'ils forment un champ radial, alors le FOE estimé sera faux, comme le montre l'image 2.10.

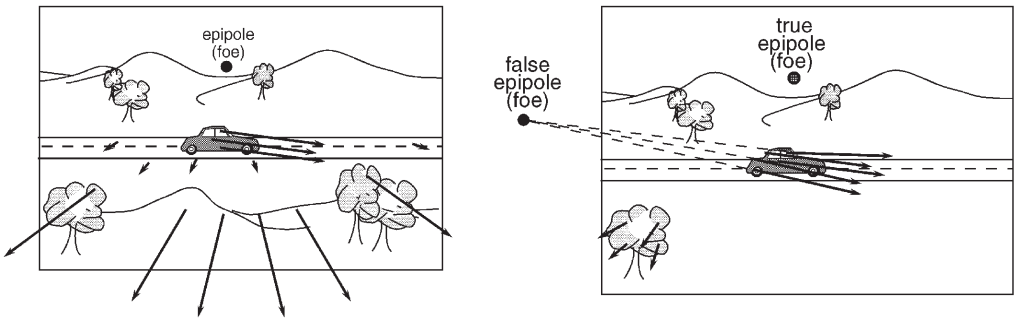


Fig. 2.10.: Estimation du FOE via les mouvements apparents. (a) Estimation correcte du FOE avec les mouvements apparents du fond. (b) Estimation incorrecte du FOE avec les mouvements apparents de l'objet en mouvement. (Figure 4 de [IA98])

Pour palier à cela, Irani et al. proposent une contrainte de consistance 3D qui s'appuie sur trois images consécutives. Les points sont testés par couple et nécessite que l'un des deux soit connu comme étant statique, il faut donc l'intervention de l'utilisateur pour sélectionner un point appartenant à la scène. Cette méthode nécessite une mise en correspondance dense pour étiqueter l'ensemble des pixels de l'image via la contrainte de consistance 3D. Sawhney et al. [Saw+99] proposent une méthode qui s'appuie uniquement sur quelques points mis en correspondance entre trois images consécutives. A partir de ces correspondances, les homographies entre les deux paires d'images sont calculées et ces dernières sont utilisées à leur tour pour estimer les épipoles. La contrainte qui permet de distinguer un problème de recalage dû à la parallaxe de celui dû à un élément mobile se déroule en deux étapes. Tout d'abord, deux images sont recalées et les parties de l'image qui ne s'alignent pas sont supposées provenir de la parallaxe. Puis les deux autres images sont recalées en conservant les précédentes régions supposées parallaxes. Si ces régions ne sont toujours pas correctement alignées, alors elles proviennent d'un élément mobile et non de la parallaxe.

Les techniques précédentes nécessitent que le plan de référence soit constant au cours du temps. Or, le plan dominant n'est pas toujours le même en fonction de la partie de scène captée par la caméra à deux instants différents. Kang et al.

[Kan+05] et Yuan et al. [Yua+07] ont étendu la contrainte de structure pour qu'elle soit valable dans le cas où le plan dominant change. Kang exprime les disparités en termes de probabilités utilisées pour filtrer les zones résiduelles mais aussi pour mieux suivre l'objet en mouvement. Yuan estime une contrainte bilinéaire qui relie deux plans dominants différents en utilisant les homographies calculées entre trois images différentes.

Il y a un cas particulier pour lequel les méthodes *Plane+Parallax* précédentes sont incapables de distinguer un sujet mobile de la scène : la caméra et l'objet se déplacent dans la même direction et leurs vélocités satisfont une relation de proportionalité constante. La figure 2.11 représente ce phénomène. Le point 3D reconstruit par triangulation se situe au même endroit alors que l'objet s'est déplacé. La contrainte de structure est alors vérifiée et le point est considéré comme statique.

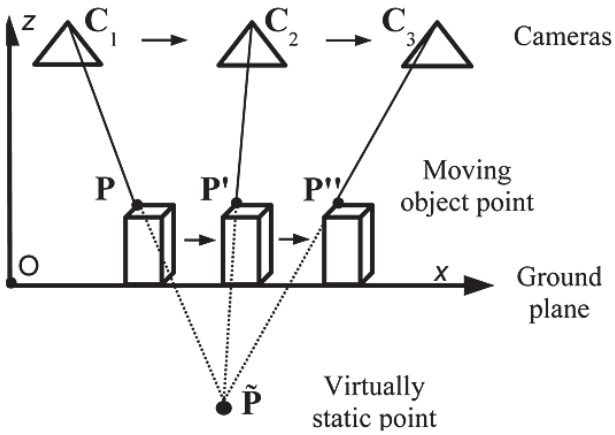


Fig. 2.11.: Cas dégénéré de mouvement de caméra et de mouvement d'un objet pour la détection (Figure 4 de [Yua+07]).

Les méthodes *Plane+Parallax* permettent de prendre en compte les effets de parallaxe dans la détection de sujets mobiles. Cependant, elles nécessitent que la scène puisse être approchée par un seul plan ce qui est généralement le cas pour des scènes prises de loin comme c'est le cas avec des images aériennes. Dans le cas d'une scène complexe où l'approximation par un seul plan n'est pas suffisante, des méthodes proposent d'en utiliser plusieurs.

2.1.3.3 Multi plans

L'environnement dans lequel nous évoluons est généralement constitué d'une multitude d'objets qui vont créer des discontinuités dans les profondeurs et donc dans le flot optique d'une caméra qui évolue près de la scène. La représentation de la scène par un plan unique n'est pas viable et la nécessité d'en utiliser plusieurs devient alors une solution évidente.

Aldeson et Wang ont mis au point à travers plusieurs travaux [Ade91 ; WA93 ; WA94] une représentation de la scène sous forme de plans superposés et ordonnés en profondeur. Chaque plan contient trois cartes : une carte d'intensité c'est la carte d'apparence, une carte alpha qui définie l'opacité ou la transparance de la carte en chaque point et la carte de vélocité qui décrit la manière dont la carte devra être recalée au cours du temps. Même dans le cas d'une scène complexe, un plan étendu est créé par accumulation d'information au cours du temps, comme un panorama, pour représenter le fond de la scène. Les mouvements sont segmentés avec un modèle affine pour extraire les différents plans de la scène.

Les discontinuités du flot optique mettent en évidence les variations de profondeurs de la scène. Les mouvements apparents similaires représentent alors un objet ou un ensemble d'objets situés à la même distance de la caméra, donc sur un même plan. De nombreuses techniques pour segmenter les mouvements et en extraire des plans ont été élaborées. Darrell et al. [DP91] utilisent un M-estimateur pour estimer les paramètres de mouvement tandis qu'Ayer et al. [AS95] utilisent un ML-estimateur. Tous deux estiment automatiquement le nombre de plans en utilisant le concept du *Minimum Description Length*. Ke et al. [KK02] segmentent les mouvements en les partitionnant dans un sous-espace définit par les homographies qui représentent les plans de la scène. Zeinik-Manor et al. [ZMI02] contraignent un sous-espace linéaire sur des homographies. Smith et al. [Smi+04] proposent quant à eux d'extraire les plans en estimant le mouvement des contours.

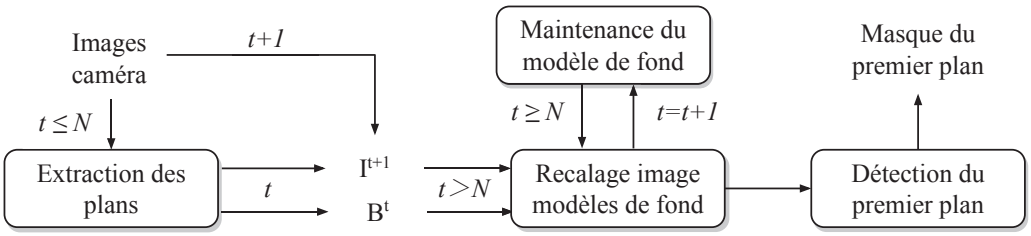


Fig. 2.12.: Processus général des méthodes multi-plans.

Afin de détecter et d'extraire des sujets en mouvement, la représentation multi-plans a été combinée avec la modélisation de fond afin d'extraire la silhouette des objets en mouvement (cf. figure **??**). Jin et al. [Jin+08] extraient les plans de la scène en utilisant une cascade de RANSAC. Pour ce faire, des points caractéristiques sont extraits et mis en correspondance entre deux images consécutives, puis un premier plan est estimé en calculant une homographie avec RANSAC. Les points caractéristiques qui ne sont pas recalés avec cette homographie sont réutilisés pour estimer une nouvelle homographie qui correspond à un autre plan dans la scène. L'opération est réitérée m fois ou jusqu'à ce qu'il n'y ait plus suffisamment de points caractéristiques pour calculer un autre plan. Les pixels de l'image courante sont

ensuite assignés à un plan en minimisant la différence d'intensité entre le pixel de l'image courante et l'image recalée d'après les homographies précédemment calculées (cf. figure 2.13). Les pixels qui ne font pas partie de la zone de recouvrement des deux images sont assignés à un plan en utilisant un *Minimal Span Tree* (MST). Un panorama est construit à partir des différents plans extraits des images au cours du temps. Les pixels du panorama sont décrits par des mélanges de Gaussiennes et la détection des éléments mobiles est réalisée par soustraction de fond.

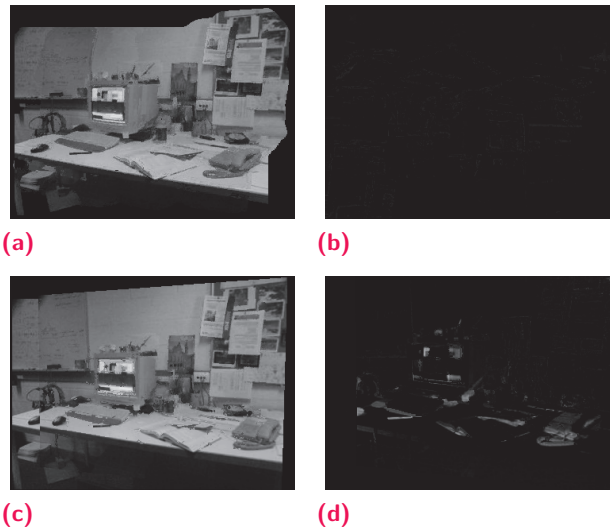


Fig. 2.13.: (a) Compensation en utilisant quatre homographies, (b) disparité entre (a) et l'image précédente, (c) compensation en utilisant une homographie, (d) disparité entre (c) et la frame précédente. (Figure 3 de [Jin+08])

Zamalieva et al. [Zam+14] proposent de représenter la scène par un ou plusieurs plans. Pour cela, les auteurs utilisent le *Geometric Robust Information Criterion* (GRIC) qui est un modèle de sélection Bayésien afin de décider d'utiliser une transformation homographique ou la matrice fondamentale. Dans les deux cas, les objets en mouvement sont extraits en minimisant une fonction d'énergie basée sur le mouvement et sur l'apparence de manière spatiale et temporelle en utilisant les modèles de fond et de premier plan précédents. A l'inverse de la méthode proposée par Jin, Zamalieva et al. [ZY14] construisent plusieurs plans équidistants et parallèles au plan dominant de la scène (cf. figure 2.14). Chaque plan possède un modèle de fond de la partie de la scène qu'il décrit. Les pixels de l'image courante sont ensuite assignés aux plans les plus proches en utilisant les points de fuite. Pour détecter les objets en mouvements, une soustraction de fond est réalisée pour chacun des plans entre le modèle de fond d'un plan et les pixels de l'image courante qui lui ont été assignés.

Les méthodes qui représentent la scène avec plusieurs plans ont l'avantage d'avoir une représentation plus juste de la structure de la scène. En revanche, elles nécessitent de trouver la bonne granularité de représentation : trop peu de plans peut introduire des problèmes de recalage lorsqu'il y a de la parallaxe et trop de

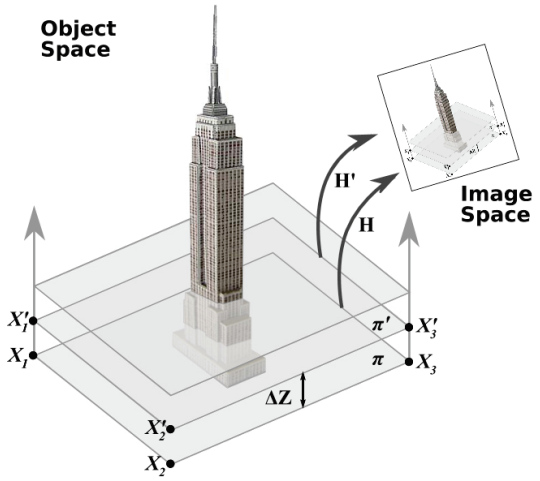


Fig. 2.14.: Génération des plans parallèles équidistants. (Figure 4 de [ZY14])

plans peut représenter un élément mobile par un plan et pourrait être considéré comme appartenant à la scène.

2.1.3.4 Segmentation de trajectoires

Cette classe de techniques segmente les trajectoires du flot optique pour ensuite les étiqueter statique/mobile. L'étiquetage éparse obtenu est ensuite étendu à toute l'image via des approches probabilistes.

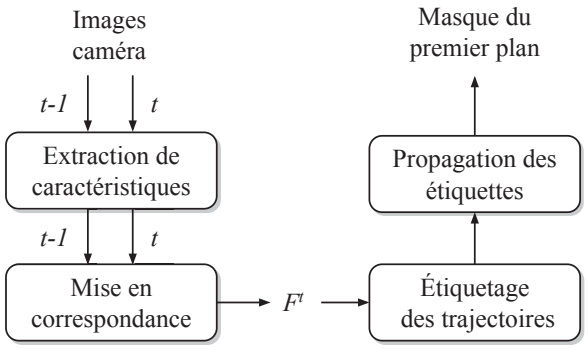


Fig. 2.15.: Processus général des méthodes par segmentation de trajectoires. F^t est le flot optique calculé entre les instants t et $t-1$.

Thompson et Pong [TP90] examinent deux classes de méthodes pour la détection d'objets en mouvement : la contrainte épipolaire et l'utilisation du flot optique avec l'information de profondeur. Dans le cas de la contrainte épipolaire, des discontinuités dans le flot optique translationnel sont recherchées, via le FOE et une connaissance du mouvement de la caméra. Dans la seconde classe de méthodes, des

connaissances a priori sur la scène telles que les surfaces, les discontinuités ou la position 3D de certains points permettent d'estimer le mouvement de la caméra ou de détecter les discontinuités créées par l'objet en mouvement.

Ogale et al. [Oga+05] proposent trois classes de méthodes en fonction du mouvement apparent (cf. figure 2.16). La première classe le fond à un mouvement apparent uniforme et l'objet en mouvement a un mouvement apparent totalement distinct qui peut être séparé en appliquant un algorithme de *clustering*. Dans le cas où l'objet et le fond se déplacent dans la même direction dans le flux vidéo de la caméra, les auteurs utilisent les différences d'estimations de profondeur pour discriminer l'objet mobile : entre les profondeurs estimées par les occultations et celles estimées par le calcul de structure à partir de mouvement (*Structure From Motion*, SFM)) ou entre les profondeurs déduites des mouvements apparents et celles calculées par la stéréoscopie.

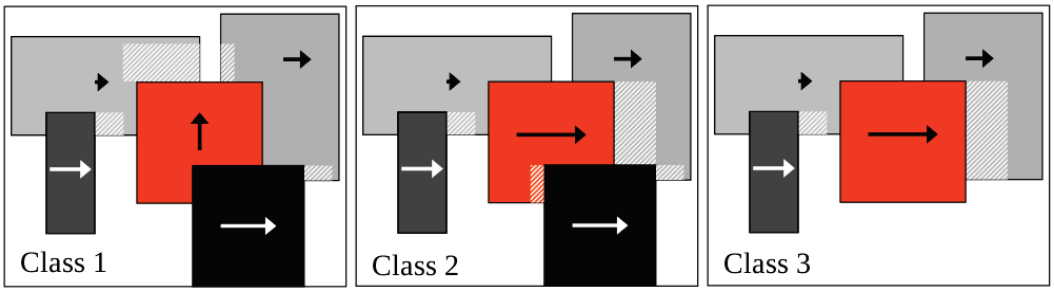


Fig. 2.16.: Exemples des mouvements représentant les trois classes de méthodes. L'objet orange est l'objet mobile a détecter. (Figure 2 de [Oga+05])

Sheikh et al. [She+09] créent un modèle de fond sur les trajectoires des mouvements apparents observés dans le flux vidéo d'une caméra mobile. Des points caractéristiques sont extraits puis traqués sur toute la séquence vidéo. Le modèle de fond est un sous-espace 3D décrit par les trajectoires qui appartiennent au fond. Pour construire le sous-espace, trois trajectoires sont choisies aléatoirement puis un consensus est calculé sur l'erreur de projection sur le sous-espace 3D formé par les trajectoires sélectionnées. Si un consensus est trouvé, le sous-espace formé représente le fond et toutes les trajectoires qui appartiennent à ce sous-espace sont étiquetées statiques. Dans le cas contraire, le consensus est recalculé avec trois autres trajectoires. Berger et al. [BS14] maintiennent le mouvement des points caractéristiques dans un sous-espace et les étiquette en utilisant un *Markov Random Field* (MRF). Cui et al. [Cui+12] distinguent les trajectoires appartenant à la scène de celles appartenant à un objet en mouvement en modélisant le fond par une matrice de faible rang avec les trajectoires du flot optique et en utilisant une contrainte sur le faible nombre de trajectoires devant appartenir à un objet mobile.

Afin de regrouper les points caractéristiques, Brox et al. [BM10] et Elqursh et al. [EE12] calculent des affinités sur des trajectoires long terme provenant de points caractéristiques suivis sur plusieurs images consécutives. Les affinités de deux points caractéristiques sont définies par la différence de leurs mouvements et sur leurs positions 2D. Les points caractéristiques sont ensuite regroupés en utilisant des algorithmes de *clustering* pour obtenir une segmentation à un niveau objet. Elqursh va plus loin en utilisant des heuristiques telles que la compacité ou la proximité spatiale pour étiqueter les groupes comme statique/mobile. Par la suite, Brox a consolidé sa méthode notamment en optimisant le regroupement des points caractéristiques [Och+14]. Elqursh et Elgammal ont quant à eux proposé une nouvelle méthode [EE13] qui divise un groupe de points caractéristiques lorsque qu'il y a de grandes variations d'affinités à l'intérieur du groupe. La méthode proposée par Nonaka et al. [Non+13] inclut la distance cosinus entre deux trajectoires pour calculer leur similarité. Les points caractéristiques sont calculés sur deux images consécutives et sont extraits sous forme de bloc pour les répartir sur l'image. Afin de limiter la sursegmentation, Keuper et al. [Keu+15] proposent une fonction non linéaire qui permet de regrouper les pixels qui ont une couleur et un mouvement similaires même s'ils sont éloignés spatiallement dans l'image. De plus, dans le cas où un objet se déplace rapidement et que ses trajectoires sont courtes, un segment qui relie une trajectoire perdue avec une nouvelle trajectoire créée est inséré pour ne former qu'une seule trajectoire. Narayana et al. [Nar+13] utilisent uniquement les orientations des trajectoires pour les regrouper et les étiqueter. Cette technique n'est appliquable que lorsque la caméra suit un mouvement de translation et échoue lorsqu'il y a des rotations.

Contrairement aux méthode présentées précédemment, seule une partie des pixels de l'image est utilisée pour la segmentation et l'étiquetage. Dans le but d'extraire la silhouette des sujets en mouvement, il est nécessaire d'ajouter une dernière étape qui permet un étiquetage dense de l'image à partir de l'étiquetage éparse obtenu (cf. figure 2.17).

(a) (b) (c)

Fig. 2.17.: (a) L'image originale, (b) les points caractéristiques blancs appartiennent à un élément mobile et les noirs au fond, (c) l'étiquetage dense de l'image construit à partir des étiquettes des points caractéristiques. (Figure 1 de [She+09])

L'étiquetage dense passe généralement par la minimisation d'une fonction d'énergie sur les étiquettes en estimant les probabilités des pixels d'appartenir à chacunes

des étiquettes. Certaines méthodes [She+09 ; Cui+12 ; EE12] proposent d'utiliser une estimation par noyau pour calculer les probabilités qu'un pixel x appartienne à l'étiquette k en se basant sur les pixels appartenant aux trajectoires déjà étiquetées. Plusieurs caractéristiques peuvent être utilisées pour calculer les probabilités telles que la couleur ou le flot optique. Puisque l'espace des solutions est trop élevé pour tester toutes les solutions de manière exhaustive, un algorithme d'optimisation qui permet de trouver une solution globale est utilisé pour minimiser l'énergie, tel que l'algorithme du *graph-cut* [BFL06].

Ochs et Brox proposent d'utiliser leur étiquetage éparse basé sur des trajectoires à long terme [BM10] pour réaliser un étiquetage dense [OB11]. Une approche variationnelle est appliquée sur une hiérarchie de régions obtenues par segmentation des images. La formulation du modèle continu permet de limiter les artefacts qui apparaissent avec une approche basée sur un graphe (cf. figure 2.18). Les mêmes auteurs ont par la suite utilisé un algorithme primal-dual pour résoudre le problème de segmentation [Och+14].

Fig. 2.18.: Différence entre un modèle discret (haut), et un modèle continue (bas) (figure 3 de [OB11]).

Les méthodes de détection d'éléments mobiles par segmentation des trajectoires issues des mouvements apparents procèdent à un étiquetage en deux étapes : un étiquetage des points caractéristiques extraits des images qui sont ensuite propagées pour obtenir un étiquetage dense. L'avantage de ce double étiquetage par rapport aux méthodes présentées précédemment est qu'il n'y a pas de recalage entre images qui peut conduire à des erreurs dues à la parallaxe. Cependant, il est généralement admis que le fond a un mouvement apparent uniforme ce qui n'est pas toujours le cas.

2.1.3.5 Reconstruction

La question de la reconstruction 3D de l'environnement a été beaucoup étudiée notamment pour les systèmes robotisés qui peuvent se déplacer de manière autonome dans l'espace. La technique du SLAM (Simultaneous Localization And Mapping) introduite par Leonard et al. [LDW91] permet, comme son nom l'indique, de cartographier l'espace et de localiser le robot à l'intérieur de celui-ci. Cet algorithme est assimilé au *paradoxe de l'oeuf ou de la poule* : pour pouvoir se localiser il est nécessaire d'avoir la carte de l'espace dans lequel on évolue ; à l'inverse, pour pouvoir cartographier l'espace il faut savoir où on se situe. Actuellement, de nombreux algorithmes [FP+15] proposent de réaliser cette tâche en utilisant généralement un filtre de Kalman [Mou+06 ; ED07 ; KM07]. Les techniques de SLAM admettent que l'entité autonome se déplace dans un environnement entièrement statique. Pour lever cette contrainte, en plus de résoudre le problème de localisation et de cartographie, certaines méthodes s'attaquent en plus à la détection et au suivi d'objets en mouvement afin de proposer une solution qui permette d'utiliser des environnements dynamiques (cf. figure 2.19).

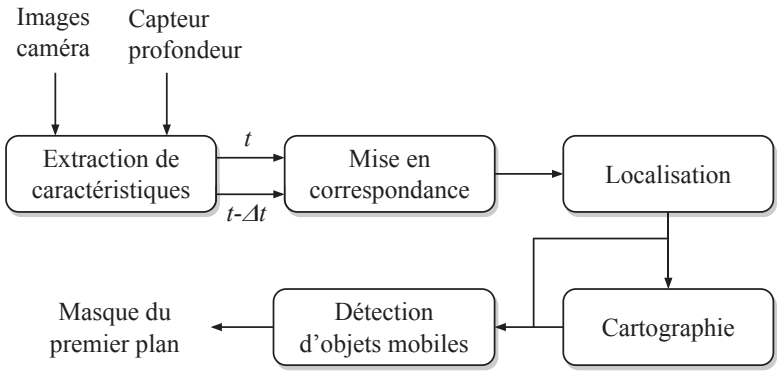


Fig. 2.19.: Processus général des méthodes par reconstruction (SLAM).

Wang et al. [WT02 ; Wan+03] proposent une méthode de détection d'objets en mouvement qui s'adapte avec une multitude d'algorithmes *SLAM*. Les auteurs utilisent les reconstructions précédentes pour savoir si un objet est ou non mobile. Par exemple, si un objet se situe dans un endroit où il n'y avait rien auparavant, alors cet objet est mobile. Les objets mobiles sont ensuite suivis en utilisant un modèle de mouvement de l'élément à détecter ainsi qu'une zone de recherche pour améliorer les performances. Par la suite, Wang et al. [Wan+07] ont proposé deux nouvelles approches pour détecter un objet en mouvement. La première est une détection basée consistance qui se déroule en deux étapes. Tout d'abord, la carte de l'environnement est utilisée pour comparer les coordonnées polaires des points extraits avec le scanner avec celles du scan récupéré à l'instant courant. Les segments

3D, qui sont issus d'une agglomération des points 3D sur un critère de distance, sont étiquetées mobiles si plus de la moitié des points qui le constituent ont été étiquetés mobiles. La deuxième méthode de détection permet de détecter des objets qui se déplacent lentement, comme un piéton, et qui sont décrits par peu de points 3D. Pour cela, la technique suppose qu'un ensemble de points 3D situés dans une zone où se situait précédemment un objet mobile doit être lui aussi mobile.

Pour garantir le bon fonction de leur algorithme de *SLAM* dans un environnement qui contient des objets en mouvement, Migliore et al [Mig+] proposent de détecter ces objets par une contrainte géométrique 3D proche de celle de la contrainte épipolaire. Cette contrainte consiste à vérifier l'intersection de trois droites de projections d'un même point caractéristique à trois positions de caméra différentes. Cependant, du fait du manque de précision des capteurs utilisés pour estimer la position de la caméra, l'application de la contrainte n'est pas aisée. Pour surmonter cette difficulté, les auteurs utilisent la *Géométrie Projective Incertaine* [Heu04] qui leur permet de vérifier les relations entre les droites de projections dans un monde incertain. Kundu et al. [Kun+11] utilisent la contrainte épipolaire combinée à la contrainte *Flow Vector Bound* (FVB) [Kun+10] pour détecter les points caractéristiques qui appartiennent à un élément mobile. La contrainte FVB permet d'aider à détecter les points appartenant à un objet en mouvement qui se déplacent le long du plan épipolaire, ce qui n'est pas réalisable par la contrainte épipolaire seule. Pour cela, des bornes minimale et maximale sont estimées pour un point caractéristique par rapport à la structure de la scène puis le déplacement minimal et maximal dans l'image le long de la droite de projection du point sont également estimés. Si le vecteur de déplacement du point caractéristique n'est pas compris entre les bornes minimales et maximales de déplacement précédemment calculées, alors le point appartient à un objet en mouvement.

Les méthodes de détection d'objets en mouvement par reconstruction ont l'avantage de pouvoir s'appuyer sur des informations 3D en plus des informations 2D qui proviennent des images, mais elles nécessitent d'avoir des capteurs en plus de la caméra couleur.

2.2 Synthèse et positionnement

La détection d'objets en mouvement dans le flux vidéo d'une caméra mobile est une tâche complexe du fait que la partie statique de la scène apparaît en mouvement dans le flux vidéo et que ce mouvement se mélange avec celui du ou des objets en mouvement. Les mouvements apparents de la partie statique de la scène peuvent être plus ou moins différents selon le mouvement de la caméra et la structure de la scène.

Dans le cas où la scène est suffisamment éloignée de la caméra ou si la scène ne contient que peu d'objets, son mouvement apparent sera relativement uniforme (cf. figure 2.20). Il est alors plus aisé de distinguer un objet mobile de la scène si celui-ci a un mouvement apparent discriminant. Dans le cas d'une scène complexe, constituée de plusieurs objets situés à différentes profondeurs, les mouvements apparents qui seront observés dans le flux vidéo dépendront fortement du mouvement de la caméra. En effet, ils peuvent être comme dans le cas précédent plutôt uniformes si la caméra effectue principalement des translations ou au contraire produire des mouvements très différents lorsque qu'il y a une combinaison rotation/translation dans le mouvement de la caméra (cf. figure 2.20).

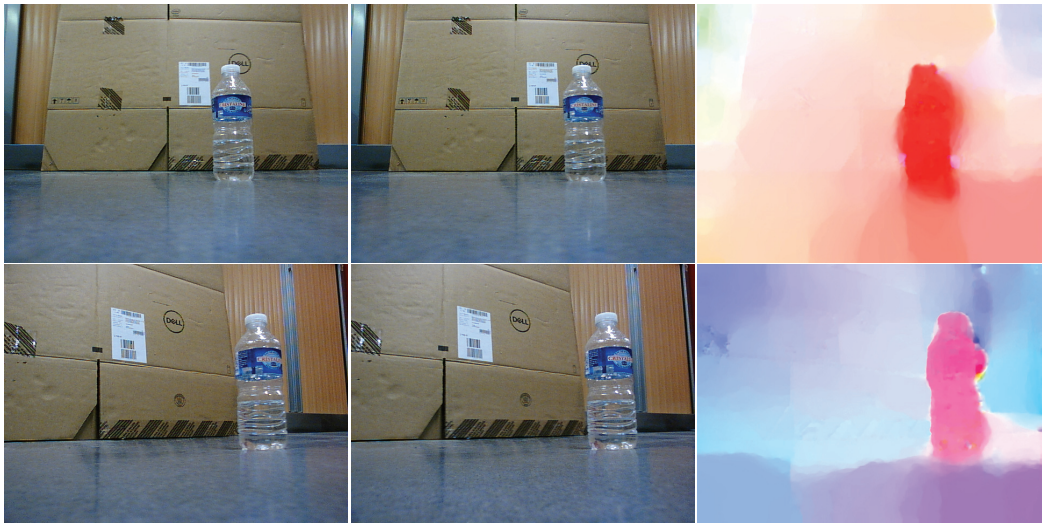


Fig. 2.20.: Différence de mouvements apparents d'une scène statique en fonction du mouvement de la caméra. Chaque ligne représente un mouvement de caméra différent et est composée de deux images prises à deux instants différents et d'une image de flot optique dense. La première ligne présente un mouvement de translation de la droite vers la gauche de la caméra. La seconde ligne présente le même mouvement auquel un mouvement de translation et de rotation de la caméra.

La section précédente montre que la littérature sur le thème de la détection d'objets en mouvement dans le flux vidéo d'une caméra est très riche et très variée dans les solutions apportées. Cette section présente les avantages et les inconvénients ainsi que les limites des différentes méthodes en reprenant la catégorisation faite dans la précédente section.

D'une manière générale, les méthodes utilisées pour détecter des sujets en mouvement dans le flux vidéo d'une caméra mobile peuvent adapter les méthodes proposées pour des caméras fixes comme c'est le cas des méthodes par *compensation de mouvement* ou *multi plans* qui utilisent des modèles de fond. Les méthodes de *segmentation de trajectoires* travaillent sur les mouvements apparents qui ne concernent qu'un certains nombre de pixels contrairement aux autres méthodes qui utilisent l'intégralité de l'image.

Les méthodes par *compensation de mouvement* et *Plane+Parallax* nécessitent que la scène puisse être approchée par un plan, ce qui est le cas des scènes qui ne contiennent pas ou très peu d'objets, donc peu complexes, et des scènes qui sont filmées par une caméra très éloignée comme c'est le cas des images aériennes. En plus de cela, les méthodes *Plane+Parallax* nécessitent qu'il y ait un seul plan dominant dans la scène. Ce problème de plan unique est adressé par les méthodes de *représentation multi-planaires* mais nécessitent que l'objet à détecter ne soit pas trop grand au risque de l'approximer par un plan.

Actuellement, aucune méthode ne permet, à notre connaissance, de détecter un sujet qui évolue dans une scène complexe, i.e. composée de divers objets qui créent de grandes variations de profondeurs du point de vue de la caméra, dans le flux vidéo d'une caméra totalement libre de mouvement.

Comme l'expliquent Thompson et Pong [TP90], il est très difficile de détecter un objet en mouvement en se basant uniquement sur des caractéristiques 2D. Dans les approches qui seront présentées dans les chapitres suivants, nous avons choisi d'introduire des informations 3D estimées à partir d'informations 2D provenant des images captées par la caméra pour réaliser la détection d'objets en mouvement avec une caméra mobile. Les approches proposées reposent principalement sur une contrainte 3D dont l'énoncé général est le suivant : dans l'espace 3D, un point qui appartient à la partie statique de la scène doit conserver sa position au cours du temps, contrairement à un point appartenant à un élément mobile. Pour pouvoir appliquer cette contrainte, des points caractéristiques sont extraits et suivis dans les images de la caméra. Puis leurs positions 3D est estimée à partir de leur position 2D en utilisant deux images. L'intérêt de travailler dans l'espace 3D plutôt que dans le 2D est que l'information 3D obtenue ne dépend pas des mouvements apparents qui ne sont pas suffisants pour distinguer la partie statique de la partie mobile.

3

Représentation en plans cohérents d'une scène pour la détection d'objets mobiles dans des séquences vidéos.

Sommaire

3.1 Introduction

Lorsqu'on observe le monde à travers les images captées par une caméra mobile, tous les éléments qui le constituent apparaissent en mouvement. Il est difficile de distinguer les mouvements apparents qui relèvent d'objets statiques de ceux qui dépendent d'objets mobiles. Les informations 2D observées dans les images ne sont pas suffisantes pour distinguer tous les cas de figures. Pour palier à cela, il est possible d'enrichir les informations 2D avec des informations 3D de la scène filmée.

L'approche présentée dans ce chapitre s'apparente à celle proposée par Zamalieva et al. [ZY14]. Après avoir identifié le plan dominant de la scène, les auteurs représentent celle-ci par des plans équidistants et parallèles au plan dominant (cf. figure 3.1). De cette manière, la scène a une représentation 3D.

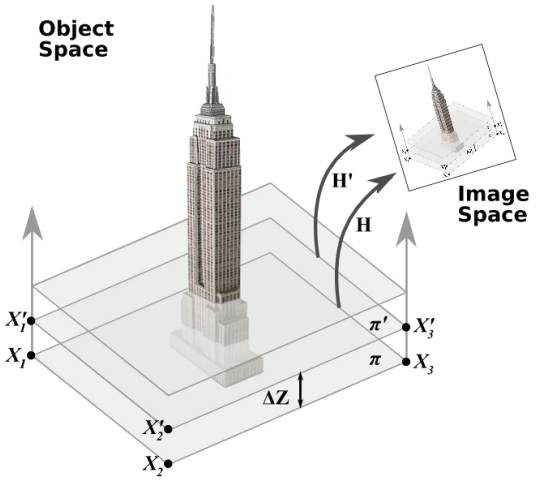


Fig. 3.1.: Ensemble de plans hypothétiques pour représenter la scène (figure 1 de [ZY14])

A chaque plan est associé un modèle de fond qui modélise l'ensemble des pixels en supposant qu'ils appartiennent à ces plans. Les points de fuite sont utilisés pour déterminer le plan d'appartenance d'un pixel de l'image caméra. Un algorithme de soustraction de fond est appliqué entre l'image courante et chacun des plans dans le but d'extraire les silhouettes des objets en mouvements. Toutefois, puisque la caméra est mobile il est possible que le plan dominant change en fonction de la portion de scène visible depuis la caméra. Dans ce cas-là, les modèles de fond associés aux plans ne sont plus cohérents avec la portion de scène visible depuis la caméra et la soustraction n'est plus possible. Il est donc impératif de conserver durant toute la séquence le même plan dominant. Pour palier cela, la méthode présentée dans ce chapitre propose d'utiliser le plan image de la caméra comme base de construction des autres plans afin d'étiqueter les points caractéristiques extraits des images comme statiques ou mobiles.

3.2 Approche

Dans l'approche présentée ici, la scène est représentée en 3D sous forme de plans construits par rapport au plan image de la caméra. Les plans sont parallèles les uns aux autres et équidistants au cours du temps (cf. figure 3.2). Les plans construits changent dans l'espace 3D puisqu'ils sont basés sur le plan image de la caméra qui est en mouvement. Comparée à la méthode de Zamalieva [ZY14], la technique de construction de plans présentée ici a l'avantage de ne pas contraindre le mouvement de la caméra puisque le plan de base nécessaire à la construction est toujours disponible et ce quelle que soit la position et l'orientation de la caméra dans l'environnement.

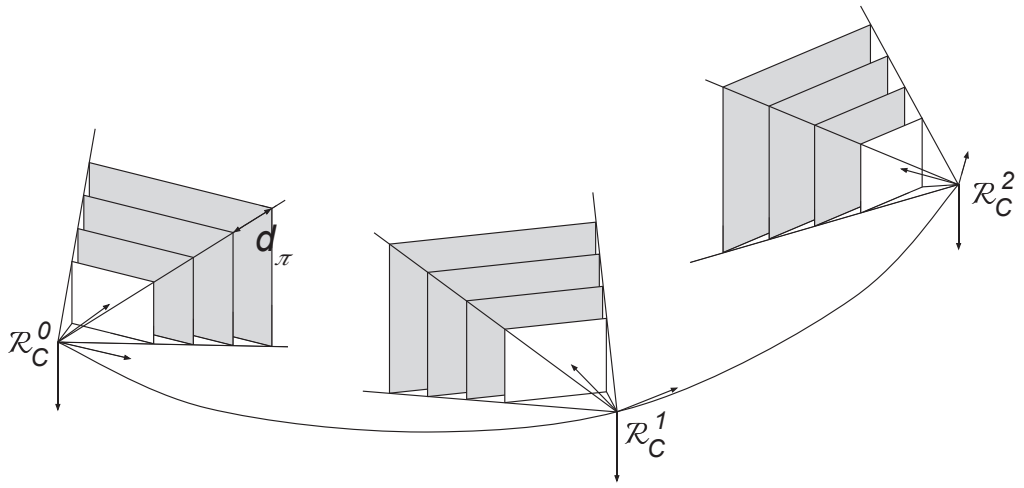


Fig. 3.2.: Schéma de la représentation de la scène par des plans 2D construits par rapport au plan image de la caméra pour chaque nouvelle image.

La méthode présentée dans ce chapitre travaille sur des points caractéristiques extraits et suivis dans les images d'une séquence vidéo. Ces points sont associés à l'un des plans 3D, puis étiquetés en utilisant une contrainte géométrique 3D. Pour pouvoir créer les plans et assigner les points caractéristiques à l'un d'eux, les paramètres intrinsèques de la caméra sont calculés une fois a priori et les paramètres extrinsèques sont supposés connus (ces paramètres doivent être fournis avec le jeu de données). Le schéma global de l'approche est représenté par la figure 3.3 et comporte trois phases : la création des plans, l'estimation des profondeurs relatives et la classification des points caractéristiques.

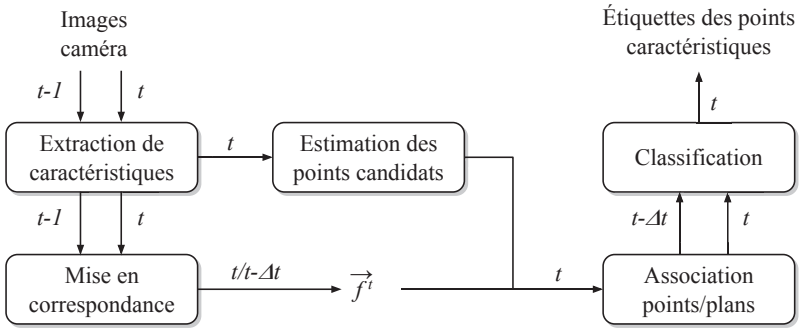


Fig. 3.3.: Schéma global de l'approche.

3.2.1 Estimation des points candidats

Pour pouvoir estimer les mouvements apparents des objets qui constituent l'environnement, il faut dans un premier temps extraire des éléments caractéristiques et les mettre en correspondance au cours du temps. Pour cela, nous avons choisi de travailler sur des points caractéristiques et nous avons utilisé l'algorithme du LDOF (*Large Displacement Optical Flow*) de Brox et Malik [BM11] qui est une méthode robuste à l'extraction et au suivi de points caractéristiques dans les images. Les positions 3D de ces points sont ensuite estimées dans le but de déterminer leurs étiquettes *statique* ou *mouvement*. Pour pouvoir réaliser cette approximation, un ensemble de K plans parallèles au plan image de la caméra $\mathcal{C}^t$ et équidistants sont construits (cf. figure 3.4). Les points caractéristiques sont projetés sur chacun des plans pour obtenir des candidats 3D :

$$P^t_{i,k} = (O^t, p^t_i) \cap \pi_k \tag{3.1}$$

avec p^t_i le i$^{\text{ème}}$ point caractéristique fourni par l'algorithme du LDOF, $\pi_k = d_\pi k$ le k$^{\text{ème}}$ plan et O^t le centre optique de la caméra $\mathcal{C}^t$. Cette projection est possible parce que la matrice intrinsèque de la caméra est connue. Ainsi, il est possible de calculer la droite de projection définie par (O^t, p^t_i) :

$$(O^t, p^t_i) = A^{-1} p^t_i \tag{3.2}$$

avec A la matrice intrinsèque. La distance d_π entre les plans qui détermine la granularité de représentation de la scène est choisie manuellement en fonction de la scène filmée.

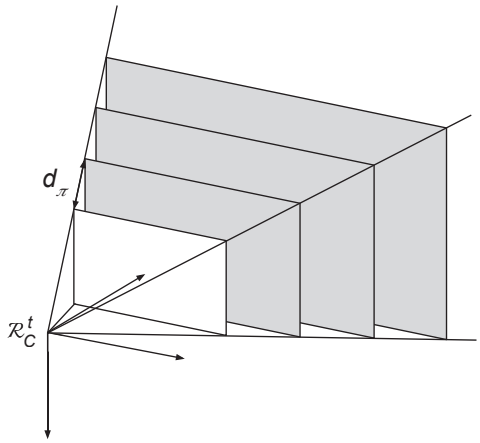


Fig. 3.4.: Schéma de construction des plans équidistants et parallèles au plan image de la caméra.

3.2.2 Association points/plans

Une seule image n'est pas suffisante pour estimer la position 3D d'un point caractéristique. A l'aide de la matrice intrinsèque, il est possible de connaitre la droite de projection mais elle ne permet pas d'obtenir la distance entre le point et la caméra. Pour cela, il est nécessaire d'utiliser une image provenant d'un autre point de vue, avec une composante translationnelle non nulle entre les positions des deux caméras.

Pour cette approche, le mouvement de la caméra est considéré connu. Ainsi, les points candidats 3D $P_{i,k}^t$ peuvent être à leur tour projetés sur le plan image de la caméra à l'instant $t + \Delta t$. La figure 3.5 représente ce processus de double projection des points caractéristiques.

Avec la mise en correspondance des points caractéristiques réalisée par la méthode du LDOF, la position 2D dans l'image caméra des points caractéristiques est connue à l'instant t et $t + \Delta t$. Le flot optique d'un point p_i est alors déduit de ces deux positions : $\overrightarrow{f}_i = \overrightarrow{p_i^t p_i^{t+\Delta t}}$. Le flot optique obtenu entre le point extrait via la méthode du LDOF à l'instant et le point projeté depuis le plan k dans l'image à $t+\Delta t$ est le flot optique du point candidat : $\overrightarrow{f}_{i,k} = \overrightarrow{p^t p_{i,k}^{t+\Delta t}}$. Pour choisir le plan auquel sera associé le point p_i, un score minimal est calculé sur l'amplitude et la direction des deux flots optiques :

$$\widetilde{P}_i^t = \underset{P_{i,k}^t}{\arg\min}(\alpha D_a(\overrightarrow{f}_{i,k}, \overrightarrow{f}_i) + (1-\alpha) D_p(\overrightarrow{f}_{i,k}, \overrightarrow{f}_i)) \tag{3.3}$$

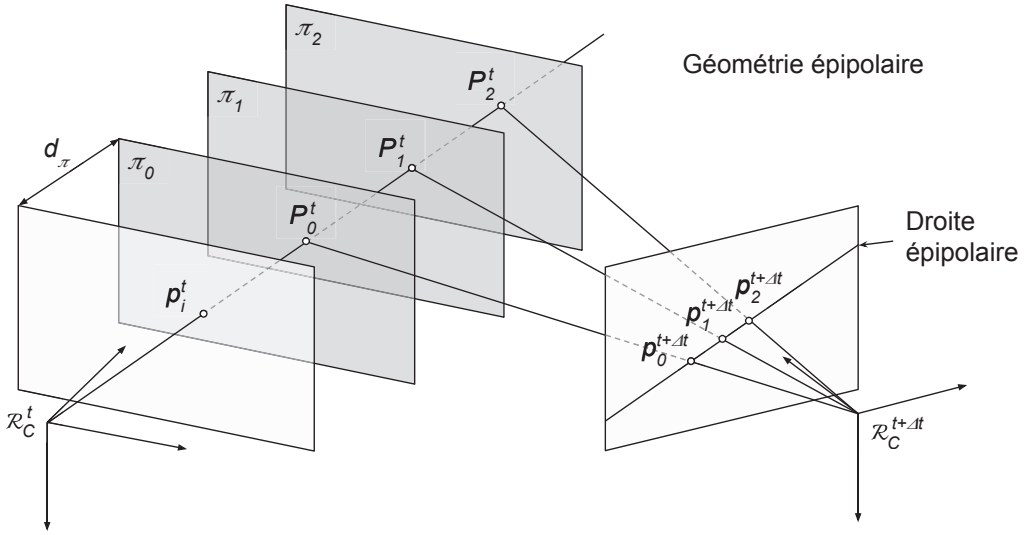


Fig. 3.5.: Schéma des projections du point caractéristique p_i^t sur les plans et reprojection sur le plan image de la caméra à $t + \Delta t$.

où $\overrightarrow{f}_i$ et $\overrightarrow{f}_{i,k}$ représentent le flot optique du point réel et du point candidat respectivement et $\alpha \in [0, 1]$ un poids. D_a et D_p définissent respectivement une différence d'amplitude et une différence de pente :

$$D_a(\overrightarrow{f}_{i,k}, \overrightarrow{f}_i) = 1 - e^{-\left(\left|\|\overrightarrow{f}_{i,k}\| - \|\overrightarrow{f}_i\|\right|\right)} \tag{3.4}$$
$$D_p(\overrightarrow{f}_{i,k}, \overrightarrow{f}_i) = 1 - \left(1 + \left(\frac{\overrightarrow{f}_{i,k}.\overrightarrow{f}_i}{\|\overrightarrow{f}_{i,k}\|.\|\overrightarrow{f}_i\|}\right)\right)/2$$

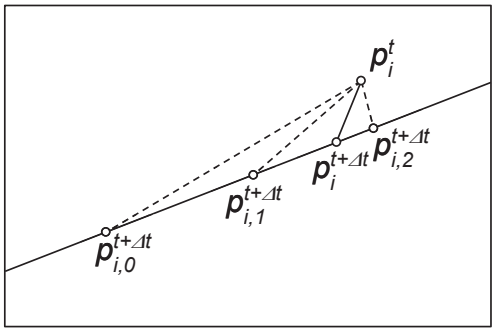


Fig. 3.6.: Schéma du flot optique d'un point caractéristique p_i entre les images t et $t + \Delta t$ et des flots optiques candidats en fonction des plans.

A l'issue de cette étape, les points caractéristiques sont assignés à un des plans π_k d'après l'équation 3.3. Seuls les points caractéristiques qui ont été suivis pendant au moins Δt images sont utilisés à cette étape puisqu'il est nécessaire d'avoir deux ensembles de points caractéristiques mis en correspondance entre t et $t - \Delta t$.

3.2.3 Classification

Nous souhaitons étiqueter les points P_i^t comme statiques ou mobiles. Pour cela nous analysons le mouvement 3D de ces points dans la structure en plan de la reconstruction. La classification des points caractéristiques comme *statique* ou *mouvement* est réalisée par une fonction de seuillage sur le déplacement de ces points dans le repère 3D. Pour pouvoir appliquer ce seuillage, il est nécessaire que les deux positions 3D du point à tester soient dans le même repère. Soit p_i^t le point à tester à l'instant t, $\widetilde{P}_{i,k}^t$ et $\widetilde{P}_{i,k}^{t-\Delta t}$ ses positions 3D au temps t et $t-\Delta t$ définies dans les repères caméra $\mathcal{C}^t$ et $\mathcal{C}^{t-\Delta t}$ respectivement. Un changement de repère est appliqué à $\widetilde{P}_{i,k}^{t-\Delta t}$ pour avoir sa position dans le repère camera $\mathcal{C}^t$:

$$\widetilde{P}_{i,k'}^{\prime t} = R \times \widetilde{P}_{i,k}^{t-\Delta t} + T \tag{3.5}$$

avec R et T les matrices de rotation et de translation entre les deux caméras.

Il est désormais possible d'appliquer la fonction de seuillage sur les positions 3D du point à tester :

$$\begin{cases} p_i^t \in statique & \text{si } |\widetilde{P}_{i,k'}^{\prime t} - \widetilde{P}_{i,k}^t| < \varepsilon, \\ p_i^t \in mouvement & \text{sinon.} \end{cases} \tag{3.6}$$

La plus grande difficulté dans l'application de la fonction de seuillage 3.6 est le choix du seuil ε. Cela vient du fait que les positions 3D des points caractéristiques ne sont pas exactes mais dépendantes de la granularité de la représentation de la scène par les plans. Ainsi, avec ce type de représentation définie dans un repère mobile, les points statiques et mobiles vont changer de positions relatives dans la représentation de la scène. On distingue deux types de mouvement qu'un point 3D $P_{i,k}$ peut réaliser entre deux pas de temps :

— $\widetilde{P}_{i,k}$ se déplace dans le plan π_k auquel il a été assigné.

— $\widetilde{P}_{i,k}$ passe du plan π_k au plan $\pi_{k+/-\Delta}$.

Un point qui change de plan se déplace au moins d'une distance d_π alors qu'un point qui se déplace dans le plan peut avoir un déplacement plus ample ou plus petit que d_π dépendante du mouvement de la caméra.

Afin de détecter les éléments mobiles, deux seuils ont été définis par rapport aux deux types de mouvement que nous avons observés :

— Un seuil ε_d pour les points qui sont restés dans le même plan entre deux pas de temps.

— Un seuil ε_π pour les points qui ont changé de plans entre deux pas de temps.

De plus, un autre élément permet de déterminer l'étiquette d'un point : la distance entre le point et la caméra. Les points caractéristiques sont reconstruits dans le cône de vision 3D de la caméra dans une zone délimitée par l'objet statique ayant la plus petite profondeur et celui ayant la plus grande profondeur au cours de la séquence vidéo. Avec un mouvement apparent incohérent au mouvement de la caméra, les profondeurs des objets mobiles sont erronées et peuvent sortir de l'espace de reconstruction défini par les éléments statiques de la scène. Pour détecter ces erreurs et les étiqueter en conséquence, nous définissions deux bornes d_{min} et d_{max} qui délimitent la zone de reconstruction des points caractéristiques statiques dans le cône de vision de la caméra.

D'après les contraintes que nous avons énoncées ci-dessus, les étiquettes des points caractéristiques sont obtenues par les équations de seuillages suivantes :

$$\begin{cases} p_i^t \in mouvement & \text{si } |\widetilde{P}_{i,k}^t - \widetilde{P}_{i,k}^{\prime t-\Delta t}| > \varepsilon_d, \\ p_i^t \in mouvement & \text{si } |\widetilde{P}_{i,k}^t - \widetilde{P}_{i,k'}^{\prime t-\Delta t}| > \varepsilon_\pi, \\ p_i^t \in mouvement & \text{si } d_{min} > |\widetilde{P}_{i,k}^t| > d_{max}, \\ p_i^t \in statique & \text{sinon.} \end{cases} \tag{3.7}$$

Les étiquettes sont estimées uniquement pour les points qui ont été suivis pendant au moins $2\Delta t$ images consécutives puisqu'il est nécessaire d'avoir deux reconstructions espacées de Δt images.

3.3 Expérimentations

3.3.1 Jeux de données et méthode d'évaluation

Les expérimentations menées pour la méthode présentée dans ce chapitre ont été réalisées sur un jeu de données virtuelles. Le jeu de données virtuelles est un ensemble de séquences vidéo qui provient des scènes réalisées pour le projet Previz et ont été élaborées avec les logiciels Unity et Blender. La position de la caméra nécessaire à notre méthode est calculée de manière exacte par Unity et Blender. Le tableau 3.1 présente un descriptif des différentes séquences vidéos utilisées.

Séquence	#images	Mouvement caméra
tigre1	40	Translation
tigre2	40	Translation
plafonnier	50	Translation/Rotation
bouteille	19	Translation

Tab. 3.1.: Description du jeu de données.

Afin d'évaluer la méthode présentée dans ce chapitre, trois valeurs statistiques ont été calculées sur les points caractéristiques : la précision P, le rappel R et la F-mesure F.

$$P^t = \frac{\{l_i | l_i = moving\} \cap \{vt_i | vt_i = moving\}}{\{l_i | l_i = moving\}} \tag{3.8}$$

$$R^t = \frac{\{l_i | l_i = moving\} \cap \{vt_i | vt_i = moving\}}{\{vt_i | vt_i = moving\}} \tag{3.9}$$

$$F^t = \frac{2P^t R^t}{P^t + R^t} \tag{3.10}$$

avec l et vt les étiquettes des points caractéristiques attribuées par notre méthode et par la vérité terrain respectivement.

La précision évalue le bruit dans l'étiquetage qui est caractérisé par des points étiquetés mobiles par la méthode alors qu'ils sont statiques (faux positifs). Une valeur de précision élevée indiquera un faible bruit. Le rappel définit le pourcentage de points caractéristiques correctement étiquetés mobiles parmi tous ceux qui sont réellement mobiles. La F-mesure est une mesure qui combine les valeurs de précision et de rappel par une moyenne harmonique. Chacune de ces trois valeurs statiques est calculée sur l'ensemble des points caractéristiques étiquetés à un instant t. Pour obtenir une évaluation statistique qui représente non pas une seule image mais la séquence vidéo, P, R et F sont calculés sur toutes les images de la séquence puis la moyenne est calculée pour obtenir $\widetilde{P}$, $\widetilde{R}$ et $\widetilde{F}$.

Les masques de vérité terrain des séquences virtuelles ont été créés automatiquement avec Unity et Blender de manière exacte pour l'intégralité de chacune des séquences.

3.3.2 Evaluation

Le tableau 3.2 et les figures 3.7 et 3.8 présentent les résultats que nous avons obtenus sur les différentes séquences présentées dans la sous section précédente.

Séquence	Précision	Rappel	F-mesure
tigre1	0.586066	0.754617	0.659746
tigre2	0.827586	0.468500	0.598300
plafonnier	0.870552	0.965466	0.915556
bouteille	0.0875912	0.0085197	0.015529

Tab. 3.2.: Résultats des valeurs moyennes de la précision, du rappel et de la F-mesure.

Dans la séquence *tigre1* (cf. figure 3.7), l'élément mobile qui est le tigron a une plus grande amplitude de mouvement apparent que la scène statique. Avec une caméra qui effectue uniquement une translation, l'élément est reconstruit comme un élément statique qui serait disposé très près de la caméra. Avec le seuillage sur la distance minimale, le sujet mobile est correctement détecté. D'après le tableau 3.2, la séquence *tigre1* présente une précision moyenne peu élevée. Cela provient des faux positifs générés par le bruit de la méthode de suivi des points caractéristiques. Le sujet mobile de l'exemple *tigre2* (cf. figure 3.7) est non rigide et a un mouvement apparent opposé à celui de la scène statique. Certains points mobiles ont tendance à changer de plans et sont détectés par le seuil ε_π. Cependant, avec des mouvements apparents relativement parallèles à ceux de la scène statique, les points de l'élément mobile correspondent aux critères des éléments statiques ce qui explique une valeur moyenne de rappel basse.

La séquence *plafonnier* (cf. figure 3.8) présente un cas particulier de mouvement apparent. En effet, le plafonnier situé au milieu des images de la séquence a un mouvement apparent différent du reste de la scène statique alors qu'il est lui-même statique. Dans cette séquence, seule la bouteille disposée sur le bar est mobile et son un mouvement apparent va dans la même direction que celui de la partie arrière de la scène (le bar et les étagères). Grâce à la contrainte de la zone de reconstruction définie sur le cône de vision de la caméra, notre méthode permet d'étiqueter correctement le plafonnier comme statique et de détecter la bouteille comme élément mobile du fait que sa reconstruction est en dehors de la zone délimitée par les seuils d_{min} et d_{max}.

Les résultats des valeurs moyennes de précision, de rappel et de F-mesure pour la séquence *bouteille* (cf. figure 3.8) montrent que l'objet mobile n'a pas été détecté par notre méthode. Cette séquence présente un exemple de cas dégénéré de mouvement de la caméra et de l'objet à détecter qui ne permet pas de distinguer un objet mobile du reste de la scène statique. Ce cas particulier est proche du cas dégénéré des

méthodes *Plane+Parallaxe* : lorsque l'objet mobile effectue un mouvement cohérent avec celui de la caméra (cf. figure 3.9), l'objet mobile se comporte comme un élément statique et devient indétectable. La séquence *bouteille* a été réalisée avec une caméra qui effectue un mouvement de translation de la gauche vers la droite et l'objet mobile effectue le mouvement inverse.

La figure 3.9 présente un exemple du cas dégénéré. D'après cette figure, le point reconstruit entre $t - \Delta t$ et t est le même que celui entre t et $t + \Delta t$ alors que le point

(a) Séquence tigre1

(b) Séquence tigre2

Fig. 3.7.: Résultat sur le jeux de données virtuelles. Chaque groupe d'image contient : image originale à $t - \Delta t$ et à t, étiquetage des points caractéristiques (en vert les points étiquetés statiques et en rouge ceux étiquetés mobiles), vérité terrain, points caractéristiques, reconstruction vue de dessus. Les zones rouges sur les deux dernières images mettent en évidence l'objet mobile à détecter.

(a) Séquence plafonnier

(b) Séquence bouteille

Fig. 3.8.: Résultat sur le jeux de données virtuelles. Chaque groupe d'image contient : image originale à $t - \Delta t$ et à t, étiquetage des points caractéristiques (en vert les points étiquetés statiques et en rouge ceux étiquetés mobiles), vérité terrain, points caractéristiques, reconstruction vue de dessus. Les zones rouges sur les deux dernières images mettent en évidence l'objet mobile à détecter.

est mobile. Dans ce cas-là, le point 3D reconstruit est cohérent avec le mouvement de la caméra et est estimé statique. Ce type d'élément mobile est détectable uniquement dans le cas où l'objet se déplace suffisamment rapidement pour que sa reconstruction soit en dehors de la zone délimitée par les seuils d_{min} et d_{max}.

3.4 Synthèse

L'approche présentée dans ce chapitre permet la détection d'éléments mobiles dans le flux vidéo d'une caméra en mouvement en utilisant une contrainte basée sur l'analyse des mouvements 3D relatifs de points caractéristiques. La détection est réalisée via un étiquetage statique/mobile de points caractéristiques extraits et suivis au cours de la séquence vidéo. Pour y parvenir, les positions 3D de points caractéristiques sont estimées sur des plans 3D liés à la caméra. La détection est alors réalisée par différents seuillages sur les déplacements 3D et les profondeurs des points.

L'un des avantages de cette approche est de travailler dans le repère 3D, là où les déplacements sont plus facilement différentiables contrairement au repère 2D

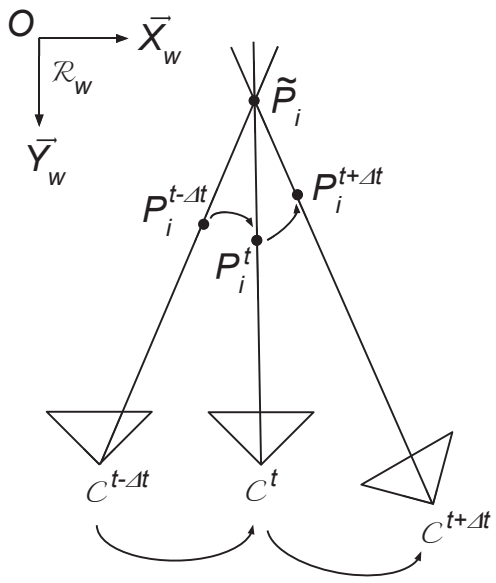


Fig. 3.9.: Estimation de la position 3D d'un point par triangulation à l'aide d'une caméra mobile, avec $\widetilde{P}_i$ la position estimée du point.

du plan image de la caméra où les mouvements apparents des objets statiques et mobiles se mélangent plus aisément. Notre méthode permet également de différencier des éléments mobiles des éléments statiques dans des cas plus extrêmes où les mouvements apparents de la partie statique de la scène divergent fortement.

Avec notre représentation 3D, la détection des objets mobiles est fortement dépendante de la granularité de la représentation par les plans. En effet, si peu de plans sont utilisés, les objets mobiles auront moins tendance à changer de plan et donc le seuil ε_π serait moins efficace. De plus, cela augmenterait les erreurs de reconstruction des points statiques du fait de la projection sur les plans, au risque que les différentes fonctions de seuillages estiment des étiquettes erronées. En outre, comme nous l'avons observé dans les expérimentations, certains mouvements relatifs entre la caméra et les objets mobiles à détecter génèrent des reconstructions 3D cohérentes empêchant la détection des éléments mobiles.

La méthode présentée dans le chapitre suivant propose une amélioration de l'estimation des étiquettes des points caractéristiques en utilisant plusieurs points dans l'estimation de l'étiquette de chacun des points et en tenant compte de la temporalité.

4 Couplage d'informations 2D et 3D pour un étiquetage éparse

Sommaire

4.1 Introduction

L'approche présentée dans ce chapitre étiquette les points caractéristiques comme *statique/mouvement* en s'appuyant sur les positions 3D des points estimées. L'identification d'un mouvement dans le flux vidéo d'une caméra mobile est une tâche complexe étant donné que les objets statiques apparaissent en mouvement et que ces mouvements se confondent avec ceux du ou des sujets mobiles. Nous avons vu dans le chapitre précédent que l'utilisation des positions 3D des points caractéristiques permettaient de distinguer les éléments mobiles de ceux statiques. En revanche, le bruit généré par l'estimation des positions 3D des points rend la méthode instable. L'approche présentée ici reprend l'idée principale de l'approche décrite dans le chapitre précédent à savoir qu'un point est considéré en mouvement s'il se déplace significativement dans l'environnement 3D au cours du temps, autrement dit si son déplacement 3D est supérieur à un certain seuil. Afin de distinguer la partie mobile de la partie statique de la scène, une contrainte géométrique 3D est appliquée sur les points caractéristiques en se basant sur un élément connu statique pour rendre la méthode plus robuste au bruit. Un aspect temporel est ajouté à l'étiquetage pour rendre la méthode plus stable.

On ne considère que les scènes composées de deux types d'éléments : les objets statique et les objets/sujets mobiles. La partie statique de la scène peut être assimilée à un corps rigide. Un corps rigide est un corps solide non articulé qui ne subit aucune déformation. De cette caractéristique découle la propriété suivante : les distances 3D entre chacun des points qui constituent le corps rigide restent invariantes au cours du temps. Cette propriété définit une contrainte géométrique 3D que doivent respecter les points caractéristiques appartenant à la partie statique de la scène et qui discrimine ceux appartenant à un objet en mouvement.

Contrairement à l'approche présentée dans le chapitre précédent, la méthode que nous proposons dans ce chapitre estime les positions 3D des points caractéristiques par triangulation sans utiliser de structure intermédiaire à la représentation de la scène. Cette technique nécessite de connaître les paramètres intrinsèque et extrinsèque de la caméra. De la même manière que pour la méthode présentée dans le chapitre précédent, les paramètres intrinsèques sont calculés une fois a priori. En revanche, les paramètres extrinsèques ne sont pas connus mais estimés à la volée à chaque pas de temps en se servant de la mise en correspondance des points caractéristiques. L'étiquetage proposé est également éparse mais prend en compte la temporalité pour éviter des changements d'étiquettes trop fréquents et s'appuie sur plusieurs points caractéristique pour un résultat plus robuste.

4.2 Approche

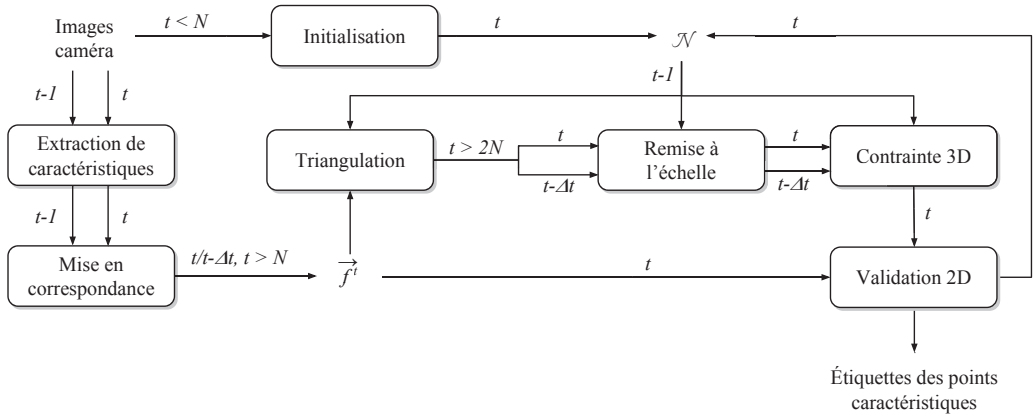


Fig. 4.1.: Schéma général de l'approche.

L'approche comporte cinq étapes exécutées séquentiellement ainsi qu'une étape d'initialisation (cf. figure 4.1). Des points caractéristiques sont *extraits* et *mis en correspondance* par l'algorithme du LDOF (Large Displacement Optical Flow) de Brox et Malik [BM11]. Les positions 3D des points sont ensuite *triangulées* et *mises à l'échelle* en s'appuyant sur les mises en correspondance. La contrainte du corps rigide est utilisée pour associer une étiquette à chaque point caractéristique. Ces étiquettes sont ensuite validées en utilisant le flot optique 2D estimé sur deux images consécutives. Une étape préliminaire d'*initialisation* est réalisée durant les premières secondes de la séquences afin d'identifier un ensemble de points caractéristiques statiques nécessaire à l'application de la contrainte géométrique 3D.

Dans un premier temps nous introduisons notre contrainte géométrique 3D afin de comprendre l'articulation des différentes étapes du processus autour de cette contrainte puis les étapes seront décrites plus en détails dans un second temps.

4.2.1 Contrainte géométrique 3D

Dans la scène, tous les objets statiques restent physiquement au même endroit, il est ainsi possible de les considérer comme une seule entité représentant un corps rigide. Tous les éléments statiques qui constituent le corps rigide conservent leurs distances 3D au cours du temps ce qui n'est pas le cas entre un objet statique et un objet mobile.

La contrainte géométrique 3D permet de différencier les points caractéristiques statiques de ceux qui sont mobiles au cours du temps en utilisant $\mathcal{N}$ comme référence statique. Cette contrainte est utilisée durant l'étape d'*étiquetage des points caractéris-*

tiques (cf. section 4.2.3) qui produit un premier étiquetage *statique/mouvement* des points caractéristiques pour chaque nouvelle image caméra.

$$Err(P_i, P_j)^t = |dist(P_i, P_j)^t - dist(P_i, P_j)^{t-\Delta t}| \tag{4.1}$$

La contrainte géométrique 3D qui vérifie la stabilité d'un point P_i est définie par :

$$\begin{cases} statique & \text{si } Err(P_i, P_j)^t < \varepsilon_{statique}, \text{ avec } P_j \in \mathcal{N} \\ mouvement & \text{sinon.} \end{cases} \tag{4.2}$$

Il faut noter ici que l'utilisation d'un ensemble stable de points étiquetés statiques comme base de comparaison est très important. En effet, si le point à tester n'est pas comparé avec un point de l'ensemble stable statique mais un point quelconque, des incertitudes surviennent sur la nature du point :

— si $Err(P_i, P_j)^t > \varepsilon_{statique}$, l'un des deux points voire les deux points sont mobiles mais il est impossible d'identifier lequel,

— si $Err(P_i, P_j)^t < \varepsilon_{statique}$, les deux points sont statiques ou les deux points appartiennent à un même objet mobile rigide.

La comparaison avec l'ensemble stable statique permet de lever l'ambiguïté sur le point qui est à l'origine de la différence puisque l'un des deux est connu comme étant statique.

Notons ici que l'erreur est calculée sur des points provenant de deux reconstructions à t et à $t - \Delta t$. L'utilité d'attendre Δt images permet de laisser le temps à l'objet de se déplacer dans le but de le détecter avec la contrainte 3D. Cette attente entre les deux reconstructions laisse également le temps à la caméra de se déplacer pour réduire l'incertitude sur la reconstruction (cf. figure 4.2). Avec un déplacement de caméra important, l'incertitude sur l'estimation des positions 3D des points est réduite. En revanche, Δt doit être suffisamment petit pour que les deux reconstructions aient des points caractéristiques en commun afin de réaliser la contrainte géométrique 3D. Dans nos expérimentations, nous avons choisi $\Delta t = 4$.

4.2.2 Reconstruction et remise à l'échelle

L'étiquetage *statique/mouvement* des points caractéristiques est réalisé via une contrainte 3D qui s'applique sur deux ensembles de points 3D. Il est donc nécessaire de retrouver l'information 3D à partir des informations 2D des images. Nous nous plaçons dans le cas d'un système stéréoscopique où la caméra mobile peut être perçue comme deux caméras en considérant deux pas de temps successifs. Grâce à

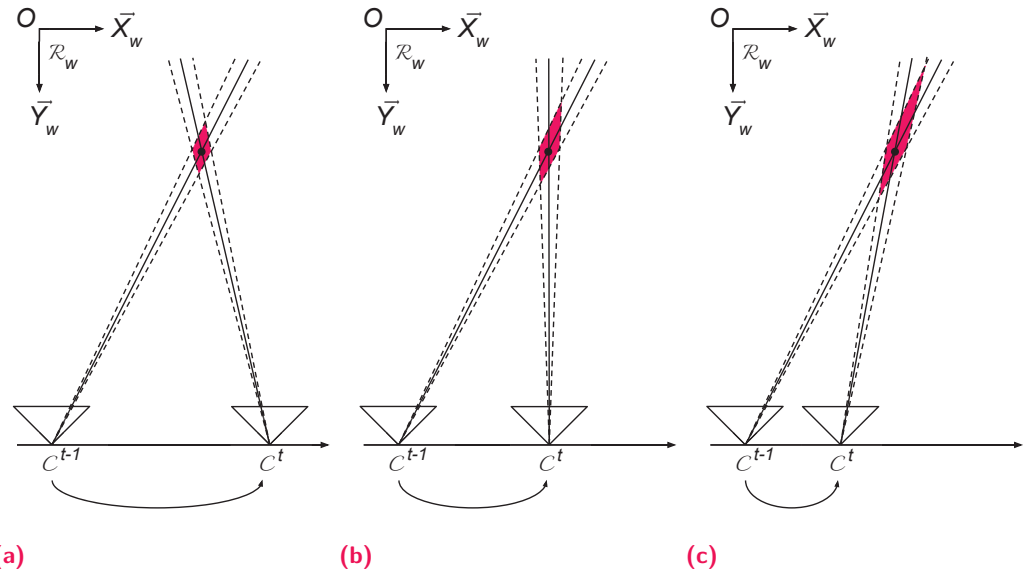


Fig. 4.2.: Schémas représentant la zone d'incertitude lors de la triangulation.

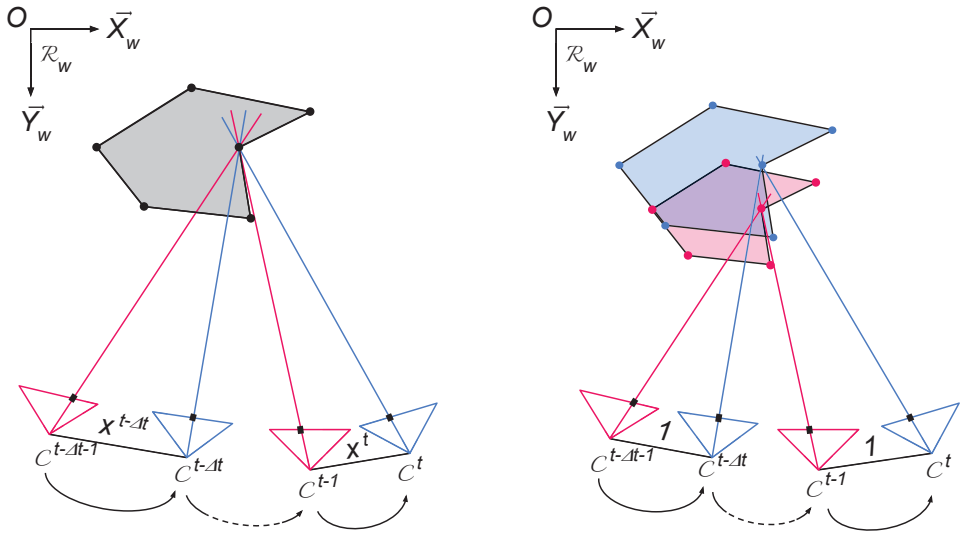
la mise en correspondance des points caractéristiques entre les images prises par la caméra mobile au cours du temps, il est possible de remonter à l'information 3D en utilisant un algorithme de triangulation [HZ03] en se servant de la position de la caméra à l'instant t comme référentiel.

Pour réaliser une triangulation, il est nécessaire de connaître les paramètres intrinsèques (passage 2D à 3D) et extrinsèques (position et orientation 3D) de la caméra. Les paramètres intrinsèques sont calculés une fois a priori et sont donc connus pour toute la séquence vidéo. Les paramètres extrinsèques sont quant à eux estimés via les points caractéristiques après leur mise en correspondance durant la prise en utilisant la méthode proposée par Nistér [Nis04]. Cette technique nécessite un minimum de cinq points pour estimer la position et l'orientation de la seconde caméra par rapport au repère de la première. Cette estimation est toutefois soumise à un facteur d'échelle puisque l'estimation de la translation n'est pas exacte mais donne uniquement un vecteur de direction dont l'amplitude vaut toujours 1.

Les paramètres extrinsèques estimés par la méthode de Nistér sont obtenus à un facteur d'échelle près et les reconstructions 3D ne sont pas comparables (cf. figure 4.3). Cette différence d'échelle provient du vecteur de direction estimé par la méthode de Nistér qui ne retranscrit pas l'amplitude du mouvement de la caméra. Cependant la contrainte 3D utilisée pour étiqueter les points caractéristiques comme *statique/mouvement* compare les distances 3D des points provenant de deux reconstructions à deux pas de temps différents. De ce fait, il est nécessaire que les deux reconstructions utilisées soient à la même échelle, mais elles ne nécessitent pas d'avoir le même facteur que la scène réelle (cf. figure 4.4).

Afin d'obtenir une estimation stable du facteur d'échelle à appliquer a une reconstruction, celui-ci est calculé uniquement à partir des points qui font partie de l'ensemble stable statique. Soit $s_{i,j}$ le facteur d'échelle calculé à partir de deux points P_i et P_j, et pour chaque couple de points statiques, provenant de deux reconstructions à l'instant t et $t-1$:

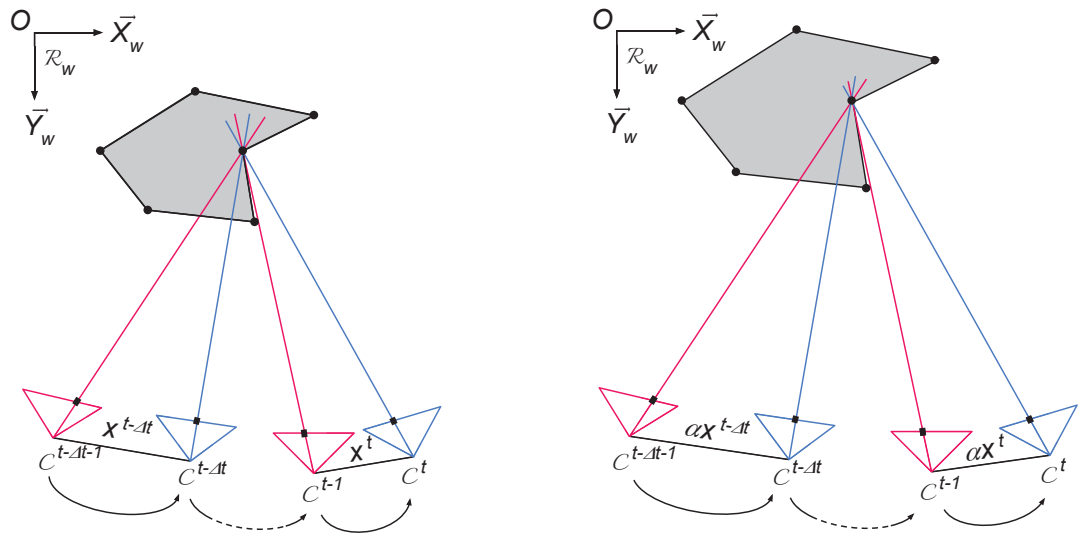
$$s_{i,j} = \frac{|\overrightarrow{P_iP_j}|^{t-1}}{|\overrightarrow{P_iP_j}|^{t}}, P_i, P_j \in \mathcal{N}^{t-1} \cap \mathcal{N}^{t} \tag{4.3}$$



(a) (b)

Fig. 4.3.: Schémas en vue de dessus de deux reconstructions des points caractéristiques d'un objet rigide statique via une caméra qui effectue des mouvements de translation et de rotation. Une première reconstruction est obtenue avec la caméra à $t - \Delta t - 1$ et $t-1$ affichée en rose et la seconde avec la caméra à $t - \Delta t$ et t en bleue. (a) Positionnement réel de la caméra dans le repère monde. La reconstruction obtenue par les caméras roses est la même que celle obtenue par les caméras bleues. Les points 3D de l'objet des deux reconstructions se recouvrent et ils sont à l'échelle de taille réelle de l'objet. (b) Positionnement de la caméra après estimation des paramètres extrinsèques via la méthode de Nister. Les caméras sont espacées de manière régulières et perdent la notion de la distance qui les séparent. Les points 3D de l'objet des deux reconstruction ne se recouvrent pas et ils ne sont plus à l'échelle.

Dans le cas où les positions 3D des points caractéristiques et le mouvement de la caméra sont calculés de manière exacte, le facteur d'échelle calculé pour une paire de points est le même pour toutes les paires de points appartenant à l'ensemble stable statique. Cependant l'extraction, la traque des points et l'estimation des paramètres extrinsèques de la caméra sont bruités et par conséquent les facteurs d'échelles $s_{i,j}$ de chaque paire seront différents. N'ayant aucune connaissance a priori sur la structure de la scène, l'estimation du facteur d'échelle repose uniquement sur les facteurs d'échelles bruités de chaque paire de points caractéristiques (cf. équation 4.3). Pour



(a) sont (b)

Fig. 4.4.: Schémas en vue de dessus de deux reconstructions des points caractéristiques d'un objet rigide statique via une caméra qui effectue des mouvements de translation et de rotation. Une première reconstruction est obtenue avec la caméra à $t - \Delta t - 1$ et $t - 1$ affichée en rose et la seconde avec la caméra à $t - \Delta t$ et t en bleue. (a) Positionnement réel de la caméra dans le repère monde. La reconstruction obtenue par les caméras roses est la même que celle obtenue par les caméras bleues. Les points 3D de l'objet des deux reconstructions se recouvrent et ils sont à l'échelle de taille réelle de l'objet. (b) Positionnement de la caméra après remise à l'échelle. Les caméras retrouvent la notion de distance spatiale qui les séparent. Les points 3D de l'objet des deux reconstructions se recouvrent et ils sont soumis à un facteur d'échelle α par rapport à leur position réelle.

estimer le facteur d'échelle optimal s à appliquer à la seconde reconstruction, nous commençons par estimer un facteur d'échelle optimal s_i pour chaque point P_i :

$$s_i = f_j(s_{i,j}) \tag{4.4}$$

puis nous utilisons l'ensemble des facteurs s_i pour estimer le facteur s à appliquer à l'ensemble des points :

$$s = f_i(s_i) \tag{4.5}$$

Après expérimentations avec la moyenne et la médiane, les fonctions f_j et f_i ont été choisies comme étant la fonction médiane qui offre de meilleurs résultats. Le facteur d'échelle s résulte donc d'une médiane de médianes de tous les facteurs d'échelles calculés sur les points caractéristiques de l'ensemble stable statique. Ce facteur d'échelle est alors appliqué aux points de la reconstruction obtenue à l'instant t :

$$\widetilde{P}_i^t = sP_i^t \tag{4.6}$$

De la même manière que pour l'estimation des positions 3D expliquée dans le chapitre précédent, seuls les points caractéristiques suivis pendant au moins Δt images sont utilisés pour cette étape, les autres étant trop jeunes pour faire partie des deux ensembles de points caractéristiques à t et $t - \Delta t$.

4.2.3 Etiquetage des points caractéristiques

A l'issue de l'étape de reconstruction et de remise à l'échelle des positions 3D des points caractéristiques, nous avons deux ensembles de points comparables. Il est donc possible d'appliquer l'équation 4.2 pour obtenir un étiquetage, cependant :

— chaque point $\widetilde{P}_i^t$ n'est comparé qu'avec un seul point de l'ensemble stable statique et le bruit peut fausser le résultat,

— un point peut changer fréquemment d'étiquette au cours du temps à cause du bruit et génère donc des instabilités dans l'étiquetage qui peuvent affecter la stabilité de $\mathcal{N}$.

Pour réaliser l'étiquetage des points caractéristiques, ces derniers seront comparés à l'ensemble de l'ensemble stable statique. Nous subdivisons l'ensemble des points caractéristiques en deux sous-ensembles :

$$\begin{aligned} \mathcal{V}_i^t &= \{\widetilde{P}_j^t | \widetilde{P}_j^t \in \mathcal{N}^t, Err(\widetilde{P}_i, \widetilde{P}_j)^t < \epsilon_{err3D}\} \\ \mathcal{I}_i^t &= \mathcal{N}^t \backslash \mathcal{V}^t \end{aligned} \tag{4.7}$$

avec $\mathcal{V}_i^t$ l'ensemble des points caractéristiques de l'ensemble stable statique avec lequel le point $\widetilde{P}_i$ est défini statique d'après la contrainte géométrique et $\mathcal{I}_i^t$ son complément par rapport à $\mathcal{N}^t$.

Dans le but d'éviter des changements d'étiquette trop fréquent au cours du temps, l'étiquette l_i d'un point caractéristique est définie par une *valeur de confiance* $Conf_i$:

$$l_i = \begin{cases} statique & \text{si } Conf_i < \epsilon_{mouvement}, \\ mouvement & \text{si } Conf_i > \epsilon_{statique}, \\ incertain & \text{sinon.} \end{cases} \tag{4.8}$$

avec $\epsilon_{mouvement} \in [-1, 0]$, $\epsilon_{statique} \in [0, 1]$ et $Conf_i^t \in [-1, 1]$. Plus la valeur de confiance sera proche de 1 ou de -1, plus la certitude de l'étiquette attribuée au point caractéristique sera haute et inversement si elle est proche de zéro. Les points qui ont une valeur de confiance qui s'approche de zéro ou qui oscille autour de cette valeur posent problème parce qu'ils peuvent introduire des faux-négatifs dans $\mathcal{N}$. Pour éviter cela, une troisième étiquette a été introduite : *incertain*. Elle permet

de mettre de côté les points qui ont une valeur de confiance peu sûre pour ne pas introduire d'erreur dans l'étiquetage final qui ne comporte que les étiquettes *statique* et *mouvement*. La valeur de confiance d'un nouveau point caractéristique est initialisée à 0.

La valeur de confiance est mise à jour à chaque nouvelle image captée par la caméra :

$$Conf_i^t = Conf_i^{t-1} + U_i^t \tag{4.9}$$

avec U_i^t la valeur de mise à jour au temps t du ième point. La valeur de confiance est calculée par une accumulation de valeurs positives et/ou négatives qui vont influencer le choix de l'étiquette à attribuer au point caractéristique à l'instant t. Le calcul de la valeur de mise à jour pour chaque nouveau pas de temps U_i^t se fait sur la base d'un vote à la majorité pour décider si la valeur sera positive (bonus) ou négative (malus).

Calcul de la valeur bonus U_i^t est une valeur positive dans le cas où $\widetilde{P}_i$ est estimé statique au temps t, c'est-à-dire si $|\mathcal{V}^t| > \alpha|\mathcal{N}|^t$. La valeur de mise à jour dépend des points de l'ensemble stable statique avec lesquels $\widetilde{P}_i$ vérifie la contrainte géométrique 3D :

$$U_i^t = re^{-\left(\frac{\overline{Err_i^t}}{\epsilon_{err3D}}\right)^2} \tag{4.10}$$

avec r un coefficient qui détermine la borne supérieure de la valeur bonus, ϵ_{err3D} le seuil de différence de distances et $\overline{Err_i^t}$ l'erreur de distance moyenne calculée sur l'ensemble $\mathcal{V}^t$:

$$\overline{Err_i^t} = \frac{1}{|\mathcal{V}_i|} \sum_{\widetilde{P}_j \in \mathcal{V}_i^t} Err(\widetilde{P}_i, \widetilde{P}_j)^t \tag{4.11}$$

Calcul de la valeur malus U_i^t est une valeur négative dans le cas où P_i est estimé mobile au temps t, c'est-à-dire si $|\mathcal{I}^t| \leq \alpha|\mathcal{N}^t|$. Le malus est calculé de la même manière que la valeur de bonus :

$$U_i^t = -re^{-\left(\frac{\overline{Err_i^t} - 1.}{max_{err} - \epsilon_{err3D}}\right)^2} \tag{4.12}$$

avec max_{err} l'erreur maximale de distance entre $\widetilde{P}_i^t$ et $\widetilde{P}_i^{t-\Delta t}$ et $\overline{Err_i^t}$ l'erreur de distance moyenne calculée sur l'ensemble $\mathcal{I}$:

$$\overline{Err_i^t} = \frac{1}{|\mathcal{I}_i|} \sum_{\widetilde{P}_j \in \mathcal{I}_i^t} Err(\widetilde{P}_i, \widetilde{P}_j)^t \tag{4.13}$$

max_{err} est l'erreur de distance maximale qui permet de plafonner les erreurs de distances des points de l'ensemble $\mathcal{I}$: $\overline{Err_i^t} = min(\overline{Err_i^t}, max_{err})$. Cette nécessité

d'erreur plafond vient du fait que l'ensemble $\mathcal{I}$ regroupe les points dont l'erreur de distance est issue d'un objet mobile ou du bruit. L'erreur de distance ici n'est donc pas bornée or la valeur malus qui dépend de l'erreur de distance moyenne nécessite une borne maximale pour ne pas impacter un point trop fortement à cause d'une erreur due au bruit. Nous bornons les différences de distances à la valeur $\max_{err}$. La figure 4.5 représente les courbes de bonus et de malus avec $\epsilon_{err3D} = 0.10$ et $max_{err} = 0.5$. Une large discontinuité située à ϵ_{err3D} apparait entre la courbe des valeurs bonus et celle des malus. La contrainte géomérique 3D que nous avons définie est très restrictive quant à l'attribution d'un bonus ce qui explique l'attribution d'une valeur bonus élevée à proximité de la limite ϵ_{err3D}. Il est cependant plus aisé d'obtenir une valeur malus, notamment à cause du bruit. De ce fait, la pente de la courbe des valeurs malus est plus douce et les valeurs moins élevées pour éviter les erreurs d'étiquetage.

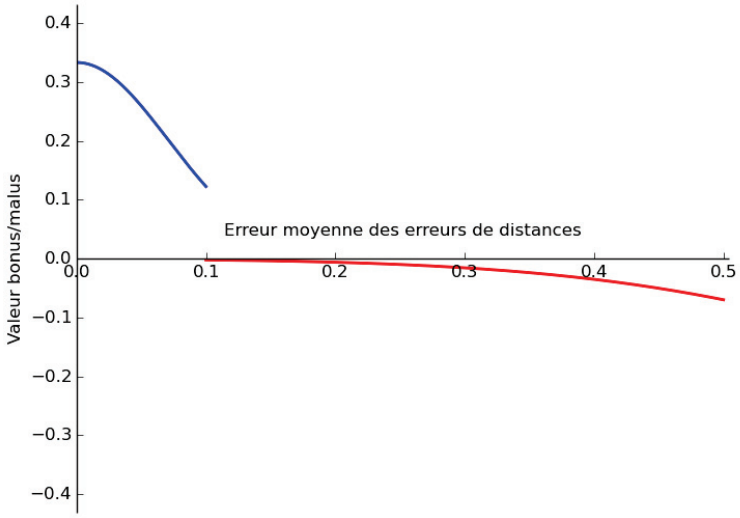


Fig. 4.5.: Courbes des différentes valeurs possible pour la valeur de mise à jour U_i^t en fonction de l'erreur de distance moyenne $\overline{Err_i^t}$. La courbe bleue est la courbe pour les valeurs bonus et la rouge pour les valeurs malus.

Une fois la valeur de confiance mise à jour, les étiquettes des points caractéristiques sont également mises à jour en utilisant l'équation 4.8. Le premier étiquetage n'apparait qu'après $2\Delta t$ images puisque notre contrainte géométrique s'appuie sur deux ensembles de points caractéristiques reconstruits et espacés de Δt images. L'image 4.6 présente un exemple d'étiquetage obtenu avec notre méthode. Sur cette image, le sujet en mouvement est correctement détecté mais on remarque que des faux positifs apparaissent sur la partie statique de la scène dus au bruit observé durant l'extraction, la mise en correspondance des points caractéristiques et au calcul des paramètres extrinsèques qui se répercute sur l'estimation de la position 3D des

points. L'étape de validation 2D qui suit va permettre de supprimer un maximum de ces valeurs aberrantes.

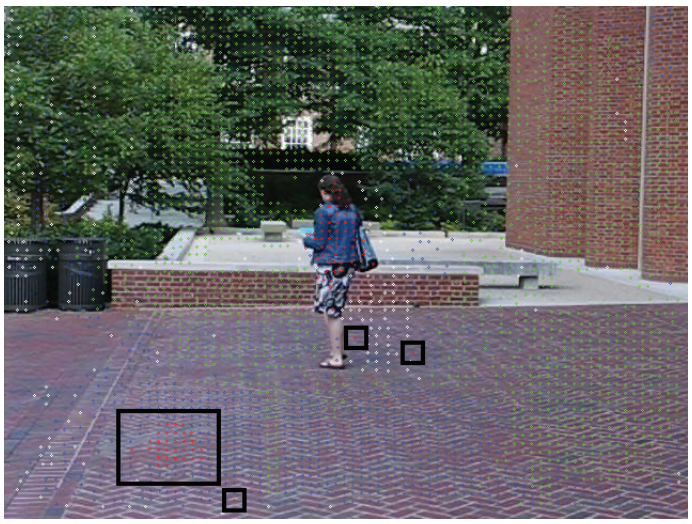


Fig. 4.6.: Résultat de l'étiquetage après l'étape de la contrainte géométrique 3D. Les points ont été colorés en fonction de leur étiquette : vert = *statique*, rouge = *mouvement*, bleu = *incertain* et blanc = point pour le moment trop jeune pour pouvoir appliquer les calculs. Les carrés noirs mettent en évidence les faux négatifs qui apparaissent dans l'étiquetage des points caractéristiques.

4.2.4 Validation 2D

L'étiquetage obtenu après application de la contrainte géométrique 3D laisse apparaître des faux négatifs mais également des faux positifs à cause des étapes d'extraction, de mise en correspondance des points caractéristiques et d'estimation des paramètres extrinsèques de la caméra qui sont bruitées.

L'étape de la validation 2D utilise les mouvements apparents pour supprimer un maximum de faux positifs de l'étiquetage. Le mouvement apparent d'un point caractéristique dépend du mouvement de la caméra et de la distance 3D entre la caméra et le point caractéristique. L'idée ici est que les mouvements apparents similaires et proches spatialement appartiennent à un même objet ou à plusieurs objets qui peuvent être assimilés à un seul objet plus grand (cf. annexe A). Le fait est qu'un objet ne possède qu'un seul état : statique ou mobile. Il est donc incohérent qu'un objet ait des points caractéristiques étiquetés *statique* et d'autres étiquetés *mobile*.

Les mouvements apparents sont représentés par des trajectoires calculées en fonction de la position des points caractéristiques entre deux images à t et $t - \Delta t$: c'est le flot optique. Les trajectoires sont regroupées dans des groupes $\mathcal{G}_j^t = (p_i^t, f_i^t)$ en

fonction de leur amplitude et de leur orientation en minimisant la fonction d'énergie suivante :

$$E(i,j) = E_{Dir}(\overrightarrow{f_i}, \overrightarrow{f_j}) + E_{Mag}(\overrightarrow{f_i}, \overrightarrow{f_j}) + E_{Dist}(p_i, p_j) \tag{4.14}$$

avec i et j les indices de deux points caractéristiques de $\mathcal{P}^t$. Les deux premiers termes de E modélisent la similarité de deux trajectoires tandis que le dernier terme représente la proximité spatiale et sont décrits par les équations suivantes :

$$\begin{cases} E_{Dir}(\overrightarrow{f_i}, \overrightarrow{f_j}) &= \frac{\overrightarrow{f_i}.\overrightarrow{f_j}}{||\overrightarrow{f_i}||.||\overrightarrow{f_j}||} \\ E_{Mag}(\overrightarrow{f_i}, \overrightarrow{f_j}) &= (u_i - u_j)^2 + (v_i - v_j)^2 \\ E_{Dist}(p_i, p_j) &= |p_i - p_j| \end{cases} \tag{4.15}$$

Il n'y a aucune contrainte sur le nombre de groupe, l'algorithme crée les groupes par propagation jusqu'à ce que tous les points caractéristiques appartiennent à un groupe. L'image 4.7 montre les groupes de points caractéristiques créés sur l'image après minimisation de l'énergie E. On remarque ici que la majorité des points appartenant à la partie statique de la scène sont regroupés ensemble tandis que l'objet en mouvement est découpés en plusieurs groupes.

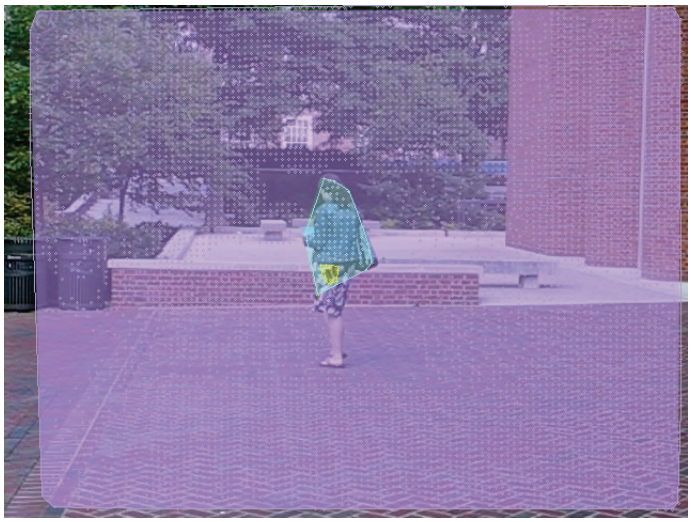
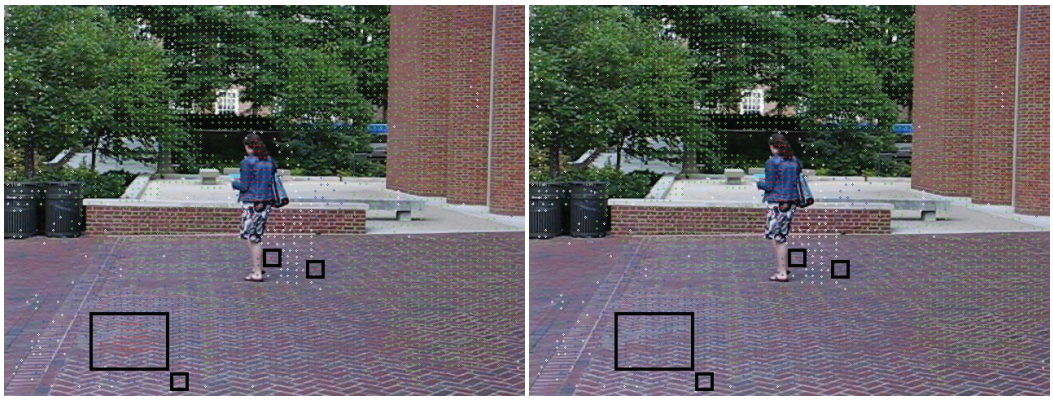


Fig. 4.7.: Groupes de points caractéristiques créés sur le flot optique et la proximité spatiale. Chaque couleur représente un groupe différent.

D'une manière générale, le bruit génère des étiquettes faux positifs plutôt que des faux négatifs qui sont des points caractéristiques étiquetés *statique* sur un sujet mobile. Ainsi, les groupes formés par la fonction d'énergie qui contiennent des points étiquetés différemment sont très souvent des groupes qui ne contiennent que des points appartenant à la partie statique de la scène. La valeur de confiance des points étiquetés *mobile* est réinitialisée à zéro pour supprimer les faux positifs sans les inclure dans l'étiquetage final en leur attribuant l'étiquette *incertain*.

Le résultat de l'étiquetage après l'étape de la validation 2D est présentée par la figure 4.8. On constate que la majorité des étiquettes faux positifs ont été supprimées tout en conservant un bon étiquetage sur le sujet mobile.



(a) (b)

Fig. 4.8.: Résultat de l'étiquetage (a) avant l'étape de la validation 2D et (b) après.

L'étiquetage obtenu à l'issue de cette étape permet de mettre automatiquement à jour $\mathcal{N}$ qui est uniquement constitué des points caractéristiques étiquetés *statique* (cf. équation 4.16). Ainsi les points qui sortent du champ visuel de la caméra ou qui ne sont plus étiquetés *statique* ne font plus partie de $\mathcal{N}$ tandis que les points caractéristiques nouvellement extraits qui viennent d'être étiquetés *statique* enrichissent $\mathcal{N}$.

$$\mathcal{N}^t = \{p_i^t | l_i^t = statique\} \tag{4.16}$$

4.2.5 Initialisation

La contrainte géométrique 3D s'appuie sur un ensemble de points caractéristiques déjà étiquetés *statique* pour mettre à jour les valeurs de confiance afin d'obtenir un étiquetage de ces points à l'instant t. L'ensemble stable statique utilisé à l'instant t est celui qui a été formé à l'instant $t-1$. A l'instant $t=0$ il n'est pas possible d'appliquer le processus de détection d'un objet en mouvement pour trois raisons :

— l'ensemble stable statique ne contient aucun point caractéristique,

— les positions 3D des points caractéristiques ne peuvent pas être calculées à partir d'une seule image,

— le flot optique ne peut pas être calculé sur une seule image.

Les étapes de la contrainte géométrique 3D et de la validation 2D ne peuvent donc pas être appliquées. L'étape d'initialisation permet de résoudre ces trois problèmes.

Durant l'initialisation il est imposé qu'aucun objet en mouvement n'apparaisse dans le champ de vision de la caméra. Ainsi, tous les points caractéristiques extraits

dans les images sont obligatoirement statiques et peuvent être directement étiquetés *statique* en initialisant leur valeur de confiance à 1. Cela permet d'initialiser l'ensemble stable de points caractéristiques étiquetés statique nécessaire à l'étape de la contrainte géométrique 3D ce qui résout le premier problème. Les deux autres problèmes se résolvent de la même manière. Il faut que la caméra capture plusieurs images tout en se déplaçant durant deux ou trois secondes.

4.3 Expérimentations

La méthode présentée dans ce chapitre a été testée sur des données réelles et virtuelles, avec et sans l'étape de validation 2D et comparée à celle d'Elqursh [EE12].

4.3.1 Jeux de données et méthode d'évaluation

La méthode a été appliquée à des données captées dans des environnements réels et virtuels. Nous avons utilisé huit séquences vidéos qui proviennent du jeu de données de Freiburg-Berkeley Motion Segmentation (FBMS-59) introduit par Ochs et al. [Och+14], ainsi que sur deux séquences vidéos créées avec l'environnement virtuel produit pour le projet Previz. Le tableau 4.1 et le tableau 4.2 présentent un descriptif des séquences issues respectivement de FBMS-59 et de Previz.

Séquence	**#images**	**Mouvement caméra**
cars1	19 (1)	Translation/Rotation
cats05	88 (1)	Translation/Rotation
dogs01	200 (1)	Translation/Rotation
horses03	45 (1)	Translation/Rotation
marple2	56 (195)	Translation/Rotation
people1	37 (1)	Translation/Rotation
people2	30 (1)	Translation/Rotation

Tab. 4.1.: Description des jeux de données réelles provenant du jeu de données de FBMS-59. La deuxième colonne contient le nombre d'images utilisées pour la séquences ainsi que le numéro de la première image.

Séquence	**#images**	**Mouvement caméra**
plafonnier1	50	Translation/Rotation
plafonnier2	50	Translation/Rotation

Tab. 4.2.: Description des jeux de données virtuelles provenant des données du projet Previz.

Pour pouvoir appliquer notre méthode, il est essentiel que les séquences vidéos respectent deux contraintes : les premières images ne doivent pas contenir d'objet en mouvement pour réaliser l'étape d'initialisation (cf. section 4.2.5). De plus, il est nécessaire que la caméra soit toujours en mouvement et doit obligatoirement effectuer au moins un mouvement de translation pour pouvoir estimer les positions 3D des points caractéristiques par triangulation nécessaires à l'estimation de la contrainte 3D. Les séquences vidéos du jeu de données FBMS-59 ne respectent pas toujours ces deux contraintes. La première contrainte peut être aisément résolue en fournissant les masques des premières images durant la phase d'initialisation dans le but de supprimer automatiquement les points caractéristiques situés sur les objets mobiles. La phase d'initialisation est réalisée uniquement sur les cinq premières images et les masques ont été créés manuellement. En ce qui concerne la seconde contrainte, nous avons choisi parmi les séquences vidéos de FBMS-59 celles dont la caméra effectue une translation et pour certaines séquences où la caméra alterne entre déplacement et arrêt, nous avons appliqué notre algorithme que sur une portion de la séquence qui respecte la seconde contrainte.

Afin d'évaluer et de comparer la méthode présentée dans ce chapitre, trois valeurs statistiques moyennes ont été calculées sur les points caractéristiques : la précision $\widetilde{P}$, le rappel $\widetilde{R}$ et la F-mesure $\widetilde{F}$ (voir section 3.3.1).

Des masques de vérité terrain sur la segmentation de mouvements sont fournis avec les jeux de données de FBMS-59 qui associent une couleur à un mouvement. Ces masques ont été adaptés pour que les éléments mobiles soient représentés en blanc et la partie statique de la scène en noir. Nous avons complété les masques de vérité terrain fournis avec les jeux de données par d'autres que nous avons créés manuellement afin de calculer les résultats quantitatifs représentatifs des méthodes testées. Concernant les séquences virtuelles, les masques ont été générés automatiquement pour l'intégralité de la séquence.

4.3.2 Évaluation qualitative et quantitative avec et sans l'étape de suppression des faux positifs

La méthode a été testée sur les jeux de données avec et sans l'étape de validation 2D qui permet de supprimer les étiquettes fausses positives. Le tableau 4.3 présente les moyennes des calculs de précision, de rappel et de f-mesure.

Pour toutes les séquences sur lesquelles a été appliquée la méthode, la précision moyenne est nettement meilleure lorsque l'étape de validation 2D est appliquée puisque la précision dépend des faux positifs supprimés par l'étape de validation 2D. En revanche, la valeur de rappel moyenne est sensiblement identique et s'explique

Séquence	Précision	Rappel	F-mesure
cars1_avec	0.943935	1.000000	0.971159
cars1_sans	0.746951	0.935115	0.830508
cats05_avec	0.940217	0.729958	0.821853
cats05_sans	0.343137	0.801527	0.480549
dogs01_avec	0.984071	0.970332	0.977153
dogs01_sans	0.257169	0.97079	0.406621
horses03_avec	0.947165	0.92686	0.936902
horses03_sans	0.497375	0.931204	0.648417
marple2_avec	0.929692	0.998651	0.962939
marple2_sans	0.801435	0.983558	0.883206
people1_avec	0.942529	0.987952	0.964706
people1_sans	0.405941	0.982036	0.574431
people2_avec	0.938073	0.992718	0.964623
people2_sans	0.641956	0.990268	0.778947

Tab. 4.3.: Valeurs moyennes de précision, rappel et F-mesure de la méthode avec (ex. cars1_avec) et sans (ex. cars_sans) l'étape de validation 2D.

par le fait que les faux négatifs dont elle dépend ne sont pas supprimés par l'étape de validation 2D. La figure 4.9 présente des résultats qualitatifs de la méthode. Sans l'étape de validation, les faux positifs s'accumulent de plus en plus au fur et à mesure de la séquence réduisant le nombre de points caractéristiques étiquetés statiques qui sont nécessaires au calcul des paramètres extrinsèques de la caméra. Contrairement à la contrainte 3D qui travaille sur des données qui ont accumulées du bruit sur les étapes d'extraction, de mise en correspondance, de reconstruction et de remise à l'échelle des positions 3D des points caractéristiques, la validation 2D n'utilise que les étapes d'extraction et de mise en correspondance.

4.3.3 Comparaison avec l'état de l'art

Une comparaison des résultats obtenus a été réalisée avec la méthode de Elqursh [EE12] qui segmente les trajectoires du flot optique. Cette méthode calcule un étiquetage complet des images contrairement à notre méthode qui calcule uniquement un étiquetage éparse. Afin de réaliser cette comparaison, nous calculons les valeurs moyennes de précision, de rappel et de F-mesure uniquement sur les pixels qui ont été étiquetés *statique* ou *mobile* par notre méthode (cf. figure 4.10).

Séquences vidéos en environnement réel Les séquences vidéos de FBMS-59 se déroulent généralement dans des scènes "simples", où les objets sont assez éloignés de la caméra et où les éléments mobiles évoluent devant. Les tableaux 4.4, 4.5, 4.6 présentent les résultats quantitatifs obtenus et la figure 4.16 présente les résultats qualitatifs.

Fig. 4.9.: Résultats de l'étiquetage des points caractéristiques sur des séquences réelles avec et sans l'étape de validation 2D. Les points caractéristiques étiquetés statiques sont affichés en vert, ceux étiquetés mobiles en rouge et ceux étiquetés incertains en bleu.

D'après le tableau 4.4, les valeurs de précisions moyennes sont généralement plus élevées pour notre méthode comparée à celles d'Elqursh, ce qui signifie qu'il y a globalement moins de bruit généré sur l'étiquetage des points caractéristiques mobiles. Cela vient du fait que la méthode d'Elqursh réalise un étiquetage d'après une analyse du flot optique en considérant que le mouvement apparent d'un élément mobile est différent de celui la partie statique de la scène, ce qui n'est pas toujours le cas. Notre méthode s'appuie quant à elle sur une analyse du mouvement 3D

Fig. 4.10.: Exemple des données utilisées pour la comparaison entre notre méthode et celle d'Elqursh. A gauche l'image originale, au centre le masque binaire généré par la méthode d'Elqursh et à droite les points caractéristiques extraits par notre méthode et étiquetés d'après le masque au centre.

Séquence	Chapel	Elqursh
cars1	0.943935	0.875576
cats05	0.940217	1.000000
dogs01	0.984071	0.356742
horses03	0.947165	0.989305
marple2	0.929692	0.997067
people1	0.942529	0.981308
people2	0.938073	0.303453

Tab. 4.4.: Valeurs de précision moyennes.

Séquence	Chapel	Elqursh
cars1	1.000000	0.8189660
cats05	0.729958	0.3544300
dogs01	0.970332	0.2216400
horses03	0.926860	0.2292440
marple2	0.998651	0.457604
people1	0.987952	0.9545450
people2	0.992718	0.9450800

Tab. 4.5.: Valeurs de rappel moyennes.

Séquence	Chapel	Elqursh
cars1	0.971159	0.8463250
cats05	0.821853	0.5233640
dogs01	0.977153	0.2734120
horses03	0.936902	0.3722330
marple2	0.962939	0.6273060
people1	0.964706	0.9677420
people2	0.964623	0.4593990

Tab. 4.6.: Valeurs de F-mesure moyennes

des points qui est plus discriminante pour les objets en mouvement qu'une analyse des mouvements apparents. En ce qui concerne les valeurs de rappels moyennes,

Fig. 4.11.: Résultats de l'étiquetage des points caractéristiques sur séquences réelles avec les méthodes de Chapel [Cha+17] et de Elqursh [EE12].

notre méthode surpasse celle d'Elqursh sur les séquences vidéos testées ici. Les valeurs moyennes de précision et de rappel se répercutent sur celles de F-mesures et indiquent que le rapport moyen entre les valeurs de précisions et de rappels sont meilleurs sur l'étiquetage des points caractéristiques utilisés par notre méthode.

Le tableau des valeurs de précisions moyennes 4.4 présente le bruit de l'étiquetage mobile des points caractéristiques. D'après les résultats obtenus par les deux méthodes, notre méthode présente des résultats plus stables mais elle est surpassée sur certaines séquences par la méthode d'Elqursh. Notre méthode est sensible à

deux éléments qui sont sources de bruit dans l'étiquetage : le premier provient de la mise en correspondance des points caractéristiques (cf. figure 4.12). En effet, certains points caractéristiques ont le même mouvement apparent que l'objet mobile alors qu'ils sont situés sur la partie statique de la scène. Ces points caractéristiques sont étiquetés "mobile" a juste titre et provoquent un bruit dans l'étiquetage. Le second élément qui provoque des erreurs d'étiquetage provient de la reconstruction réalisée avec la méthode le mouvement de caméra estimé Nister. Cette méthode utilise exactement cinq points caractéristiques pour réaliser son estimation. Ces cinq points choisis aléatoirement parmi l'ensemble stable de points caractéristiques ne sont pas toujours représentatifs du mouvement apparent et entraînent une erreur d'estimation du mouvement de la caméra. La reconstruction ainsi obtenue est incorrecte et l'application de la contrainte 3D sur les positions 3D obtenues génère des erreurs d'étiquettes. Des erreurs ponctuelles de reconstruction sont généralement maîtrisées par la nature temporelle de l'étiquetage mais des erreurs trop fortes et trop fréquentes généreront des erreurs à plus long terme.

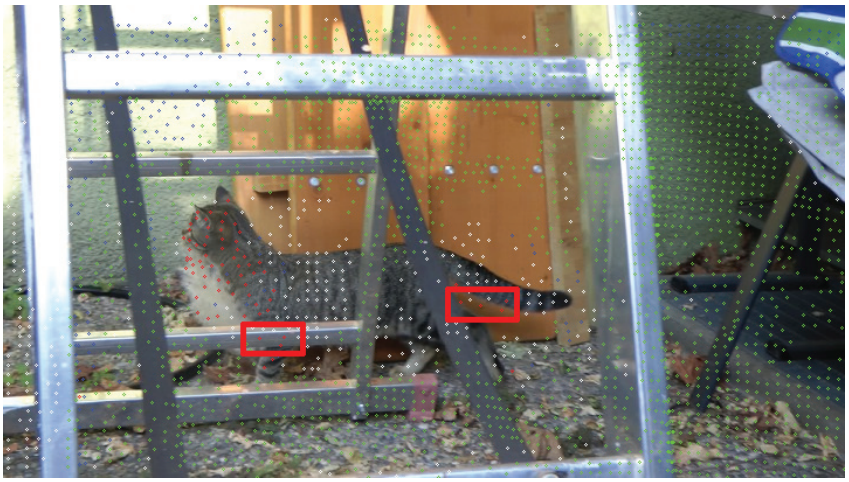


Fig. 4.12.: Erreur de mise en correspondance de points caractéristiques qui génère des erreurs d'étiquetage.

La méthode d'Elqursh est quant à elle sensible à deux éléments : le premier est le mouvement apparent qui est utilisé pour réaliser une première segmentation de points caractéristiques et un premier étiquetage (cf. figure 4.13). Dans certains cas, l'objet mobile a un mouvement apparent très proche voire identique à celui de la partie statique de la scène. Des instabilités sur la segmentation des points et leur étiquetage apparaissent. Le second élément provient de la propagation de l'étiquetage éparse à l'ensemble de l'image (cf. figure 4.14). L'étiquetage éparse obtenu à l'instant t est utilisé pour mettre à jour l'étiquetage dense. Dans le cas où l'étiquetage éparse est erroné, l'étiquetage dense va accumuler des erreurs et générer des erreurs. De plus, dans certains cas, les objets ont une apparence proche de celle de la partie statique de la scène provoquant un effet de camouflage qui induit en erreur l'étiquetage final de l'image.

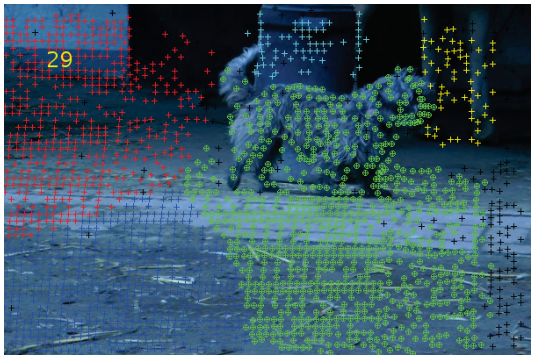


Fig. 4.13.: Erreur de segmentation de trajectoires avec la méthodes d'Elqursh. Chaque couleur représente un groupe et le groupe dont les points caractéristiques sont affichés avec des cercles est celui estimé mobile.

Fig. 4.14.: Erreur d'étiquetage dense avec la méthode d'Elqursh. A gauche, segmentation des trajectoires. Au centre, l'étiquetage dense estimé. A droite la vérité terrain.

Séquences vidéos en environnement virtuel Les séquences vidéos que nous avons réalisées dans l'environnement virtuel de Previz présentent la particularité d'avoir des mouvements apparents non uniformes pour les éléments statiques et même parfois totalement opposés (cf. figure 4.15). Pour les séquences virtuelles que nous présentons ici, le plafonnier situé au milieu des images est un objet statique mais qui a un mouvement apparent différent du reste de la partie statique de la scène. Le mouvement de caméra de la séquence *plafonnier2* est l'inverse de celui de *plafonnier1*. Afin de vérifier la stabilité de la méthode dans ce genre de situation, un objet mobile a été ajouté dans l'environnement. Les tableaux 4.7, 4.8, 4.9 présentent les résultats quantitatifs obtenus et la figure 4.16 présente les résultats qualitatifs.

Fig. 4.15.: Différence de mouvements apparents d'une scène statique en fonction du mouvement de la caméra (séquence plafonnier1). Deux images d'origines à deux pas de temps différents et l'image du flot optique dense.

Dans la séquence *plafonnier1*, la bouteille qui est en mouvement sur le bar a une direction de mouvement apparent similaire à la majeure partie de la scène statique contrairement au plafonnier qui a quant à lui un mouvement apparent opposé bien qu'il soit statique. Nous observons que les résultats obtenus avec la méthode

Séquence	Chapel	Elqursh
plafonnier1	0.9529820	0.265712
plafonnier2	0.0107527	0.149141

Tab. 4.7.: Valeurs de précision moyennes.

Séquence	Chapel	Elqursh
plafonnier1	0.70965000	0.285226
plafonnier2	0.00295858	0.872781

Tab. 4.8.: Valeurs de rappel moyennes.

Séquence	Chapel	Elqursh
plafonnier1	0.81351000	0.275124
plafonnier2	0.00464037	0.25475

Tab. 4.9.: Valeurs de F-mesure moyennes.

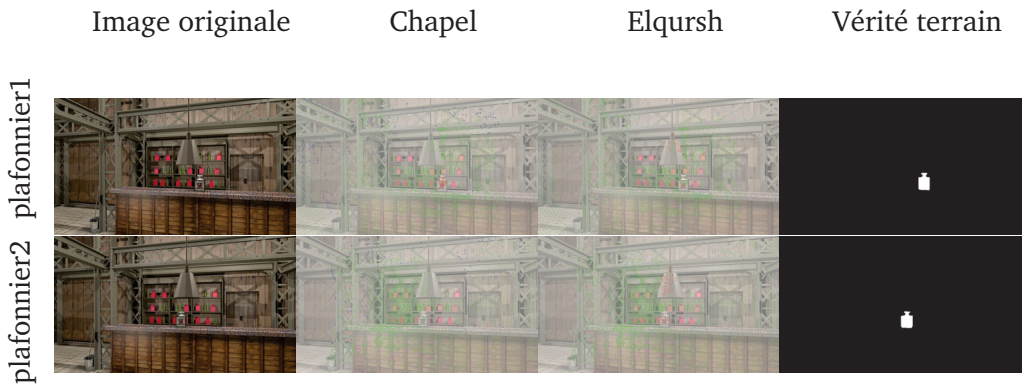


Fig. 4.16.: Résultats de l'étiquetage des points caractéristiques sur séquences virtuelles avec les méthodes de Chapel [Cha+17] et de Elqursh [EE12].

d'Elqursh détectent principalement le plafonnier comme objet mobile contrairement à notre méthode qui ne détecte que la bouteille comme objet mobile (cf. figure 4.17). En se basant uniquement sur les mouvements apparents 2D, la méthode d'Elqursh n'est pas en mesure de détecter des objets dans ce genre de situation.

Le plafonnier et la bouteille mobile de la séquence *plafonnier2* ont le même mouvement apparent. La méthode d'Elqursh détecte les deux objets comme mobile du fait de leurs mouvements apparents similaires (cf. figure 4.18). Notre méthode détecte correctement le plafonnier comme un objet statique. En revanche, la bouteille est également étiquetée statique. Cette erreur de détection provient du cas dégénéré que nous avons déjà présenté dans le chapitre précédent à la section 3.3.2. Ici, la position 3D de la bouteille est erronnée, mais elle est cohérente avec le mouvement de la caméra. De ce fait, la bouteille a un comportement statique dans l'espace de

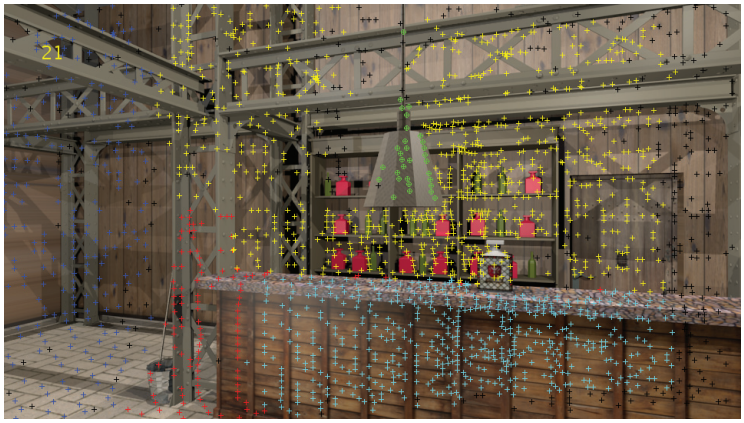


Fig. 4.17.: Segmentation des trajectoires par la méthode d'Elqursh [EE12]. Chaque groupe correspond à une couleur et le groupe dont les points sont affichés par des cercles est le groupe estimé mobile.

reconstruction 3D et ne peut pas être détectée comme mobile par notre contrainte géométrique 3D. De rares points sont parfois détectés sur la bouteille expliquant des valeurs non nulles pour notre méthode pour les différentes mesures de performances sur cet exemple.

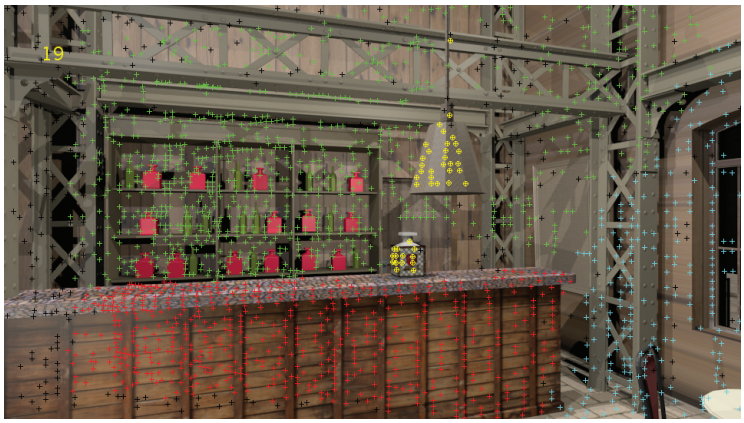


Fig. 4.18.: Segmentation des trajectoires par la méthode d'Elqursh [EE12]. Chaque groupe correspond à une couleur et le groupe dont les points sont affichés par des cercles est le groupe estimé mobile.

4.4 Synthèse

La méthode présentée dans ce chapitre permet un étiquetage *statique/mobile* éparse sur des points caractéristiques extraits et suivis dans des séquences vidéos réalisées avec une caméra mobile.

L'un des avantages de cette méthode repose sur le couplage des informations 2D et 3D qui permet de détecter des objets mobiles même dans le cas d'une scène complexe où les mouvements apparents peuvent être non uniformes. Contrairement

à la méthode présentée dans le chapitre précédent, les positions 3D des points ne sont pas projetées sur des plans.

La méthode est également plus robuste au bruit grâce au caractère temporel et spatial de la mise à jour des étiquettes. En effet, chaque point caractéristique est comparé à un ensemble stable de points caractéristiques statiques pour évaluer sa mobilité contrairement à la méthode précédente qui ne réalise qu'une comparaison locale pour chacun des points. De plus, l'introduction du troisième label, nommé *incertain*, permet de limiter le bruit provenant des points dont la valeur de confiance oscille autour de 0. La nature temporelle est quant à elle représentée par la valeur de confiance qui détermine l'étiquette des points caractéristiques en accumulant des valeurs qui déterminent l'état statique ou mobile de chaque points à chaque nouvelle image. Grâce à ces deux éléments, les étiquettes attribuées aux points caractéristiques sont plus stables au cours du temps que celles estimées avec la méthode du chapitre précédent.

L'exécution des étapes de contrainte 3D et de validation 2D est réalisée séquentiellement et sont donc totalement indépendantes l'une de l'autre. La suppression des étiquettes fausses positives par l'étape de validation 2D brise la temporalité établie par l'étape de la contrainte 3D en réinitialisant la valeur de confiance. Il sera donc nécessaire d'attendre plusieurs mises à jour de la valeur de confiance pour obtenir une nouvelle étiquette statique ou mobile.

Le type de séquence sur lequel notre approche est applicable est restreint à celles dont la caméra utilisée est perpétuellement en mouvement. Plus encore, il est nécessaire que le mouvement de la caméra contienne une composante translationnelle non nulle car dans le cas contraire il est impossible d'estimer la position 3D des points caractéristiques.

5 Conclusion et perspectives

Synthèse

Les travaux présentés dans cette thèse portent sur la détection d'objets mobiles dans le flux vidéo d'une caméra en mouvement et nous avons proposé plusieurs contributions à la résolution de ce problème.

Nous avons proposé de résoudre le problème de détection d'objets mobiles en analysant les déplacements dans l'espace 3D par opposition aux méthodes qui s'appuient sur des éléments 2D. Le mouvement 3D des points reconstruits dans le repère de la caméra mobile permet de discriminer plus facilement ceux qui appartiennent à un élément mobile de ceux appartenant à un élément statique. En effet, dans l'espace 3D la partie statique de la scène peut être apparentée à un corps rigide puisque les différents éléments qui la composent conservent leurs distances au cours du temps. Les éléments mobiles auront quant à eux des distances avec les éléments statiques qui varient au cours du temps. Les mouvements apparents des objets statiques observés dans le flux vidéo de la caméra ne garantissent pas l'uniformité et ce critère, généralement utilisé par les méthodes d'analyses des mouvements 2D, n'est alors plus suffisant pour détecter les objets mobiles.

Nous avons vu dans le chapitre 3 une structuration en plans de la scène sur laquelle est appliquée une contrainte géométrique 3D pour détecter les objets mobiles. Du fait de cette structuration, nous avons défini plusieurs seuils qui permettent de distinguer les mouvements des éléments mobiles de ceux statiques. L'efficacité de la contrainte 3D a été démontrée sur un cas extrême où les mouvements apparents de la partie statique de la scène sont opposés. Cependant, les résultats obtenus montrent que l'étiquetage est sensible au bruit car les mouvements sont analysés localement et l'étiquette attribuée à un point à l'instant t ne dépend pas des estimations précédentes.

Afin d'obtenir un étiquetage des points caractéristiques robuste et stable au cours du temps, nous avons proposé dans le chapitre 4 que chaque point soit comparé à un ensemble stable de points caractéristiques estimés statiques par opposition à l'analyse locale qui évalue le déplacement d'un point sans prendre en considération les autres points caractéristiques. De plus, les étiquettes sont définies d'après une

valeur de confiance mise à jour au cours du temps en fonction du résultat de l'analyse du mouvement 3D. L'aspect spatial et temporel de notre méthode permet de pallier le bruit qui apparaît ponctuellement. De plus, l'estimation des positions 3D des points caractéristiques sans la structuration en plans de la scène permet de ne définir qu'un seuil nécessaire à la détection d'objets mobiles. La majorité des étiquettes fausses positives générées par notre contrainte 3D sont supprimées par une étape de validation 2D qui analyse l'étiquetage par groupe de points caractéristiques. Les groupes sont formés d'après la similarité des mouvements apparents et chaque groupe formé alors ne peut contenir des points étiquetés statiques et d'autres mobiles. Du fait que le bruit a fortement tendance à étiqueter les points comme mobiles, ce sont ces points qui voient leurs valeurs de confiance réinitialisées. Les résultats présentés dans ce chapitre démontrent, qu'en comparaison avec la méthode d'Elqursh, notre étiquetage est plus stable et plus robuste et notamment sur un cas extrême où les mouvements apparents de la partie statique de la scène sont opposés.

Perspectives

Nous proposons dans cette section plusieurs pistes d'améliorations à la méthode proposée dans le chapitre 4 pour rendre la détection plus robuste et étendre le champ d'application de notre méthode à des mouvements de caméra plus complexes.

Qualité de la reconstruction

La stabilité de l'étiquetage des points caractéristiques via la contrainte géométrique 3D que nous avons définie dans le chapitre 4 est fortement dépendante de la qualité de la reconstruction des positions des points. Les erreurs de reconstruction occasionnelles peuvent être minimisées par l'aspect temporel de l'estimation des étiquettes, mais des erreurs trop fréquentes génèrent immanquablement des erreurs d'étiquetage qui se répercutent sur l'ensemble stable statique. Cet ensemble est la base de calcul de plusieurs estimations du processus de détection d'objets mobiles, il est donc nécessaire qu'il soit le plus stable possible. Les erreurs de reconstructions proviennent de mauvaises estimations du mouvement de la caméra. Nous proposons ici deux pistes à étudier pour pallier ces erreurs de reconstructions.

La première idée consiste à prendre en compte la similarité des deux reconstructions utilisées lors de l'application de la contrainte 3D à un instant t. Une évaluation de la proximité peut être réalisée sur l'ensemble stable de points caractéristiques en utilisant les formules de la contrainte 3D géométriques. Si l'ensemble stable conserve globalement (selon un seuil) les distances 3D entre les différents points qui le constituent entre les deux reconstructions, alors le calcul des valeurs de mise à

jour des valeurs de confiances reste inchangé. En revanche, si les reconstructions sont différentes, un coefficient pourra être appliqué aux valeurs de mise à jour afin de minimiser l'erreur.

La seconde idée consiste à obtenir un mouvement de caméra moins bruité en choisissant mieux les données en entrée de la méthode de Nister. Cette méthode utilise exactement cinq points caractéristiques pour réaliser l'estimation du mouvement caméra. Si les cinq points choisis ne sont pas suffisamment représentatifs du mouvement observé, le mouvement estimé est erroné. Pour le moment, ces cinq points sont choisis aléatoirement parmi l'ensemble stable. Il serait plus judicieux de choisir ces cinq points en fonction des mouvements apparents de la partie statique de la scène et qu'ils soient le plus uniformément répartis dans l'image. Les mouvements apparents peuvent être regroupés selon leurs amplitudes et leurs orientations de la même manière que nous l'avons présenté dans le chapitre 4 pour l'étape de la validation 2D. Les groupes qui contiennent le plus de points sont désignés comme représentants des mouvements apparents des éléments statiques. Ainsi, les mouvements issus de points dont le suivi dans les images a été bruité sont écartés. Les cinq points peuvent alors être choisis via une sélection uniforme parmi les groupes sélectionnés.

Suppression de la contrainte liée au mouvement

La détection par reconstruction que nous avons proposée est conditionnée par le mouvement de la caméra. En effet, pour estimer la position 3D d'un point à partir d'une seule caméra mobile, il est nécessaire que cette dernière effectue un mouvement dont la composante translationnelle est non nulle, sans quoi la triangulation du point devient impossible. Deux cas ne sont pas gérés par notre méthode : la rotation pure et l'arrêt total de mouvement.

Le cas d'arrêt de mouvement caméra peut être aisément détecté en analysant le flot optique de l'ensemble stable de points caractéristiques. Un flot optique nul ou quasiment nul sur la partie statique de la scène annonce un arrêt de mouvement. Pour poursuivre l'estimation des positions 3D via la triangulation, il est possible d'utiliser une image antérieure à l'arrêt du mouvement de la caméra. Cela permet d'avoir un mouvement de caméra entre les deux images utilisées pour la reconstruction alors que la caméra est à l'arrêt depuis quelque temps. Le choix de l'image antérieure est basé sur le flot optique de la partie de la scène qui doit être supérieur à un certain seuil pour pouvoir réaliser la reconstruction. Les images sont testées les unes après les autres en remontant dans le temps jusqu'à obtenir celle qui va satisfaire la contrainte du flot optique. Du fait que l'image antérieure peut être éloignée de l'image courante en termes de temps, les points caractéristiques présents à l'instant t

n'auront pas nécessairement tous des correspondants dans l'image antérieure. Ceci est d'autant plus vrai pour les points caractéristiques qui apparaissent à l'instant t. Ces nouveaux points proviennent généralement d'éléments mobiles qui viennent d'entrer dans le champ visuel de la caméra. Il ne sera pas possible de mettre à jour la valeur de confiance de ces points via la contrainte 3D puisqu'ils ne seront pas reconstruits, mais il est en revanche possible d'agir directement sur la valeur de confiance de ces points pour initialiser leur étiquetage comme mobile.

Le cas du mouvement rotationnel pur est plus complexe à traiter car bien qu'il y ait toujours un mouvement de caméra, celui-ci ne permet pas d'estimer les positions 3D des points caractéristiques. La détection de ce mouvement peut être réalisée par une analyse du flot optique de l'ensemble stable puisque tous les points caractéristiques ont la même amplitude de mouvement apparent et ce quelle que soit leurs distances à la caméra (cf. annexe A). La détection d'objets mobiles via la contrainte 3D peut être réalisée de la même manière que pour le cas d'arrêt de la caméra, i.e. en utilisant une image antérieure. Cependant, si la caméra effectue un mouvement de rotation trop ample, la partie de la scène visible dans l'image antérieure ne sera pas la même que celle visible dans l'image courante (cf. figure 5.1).

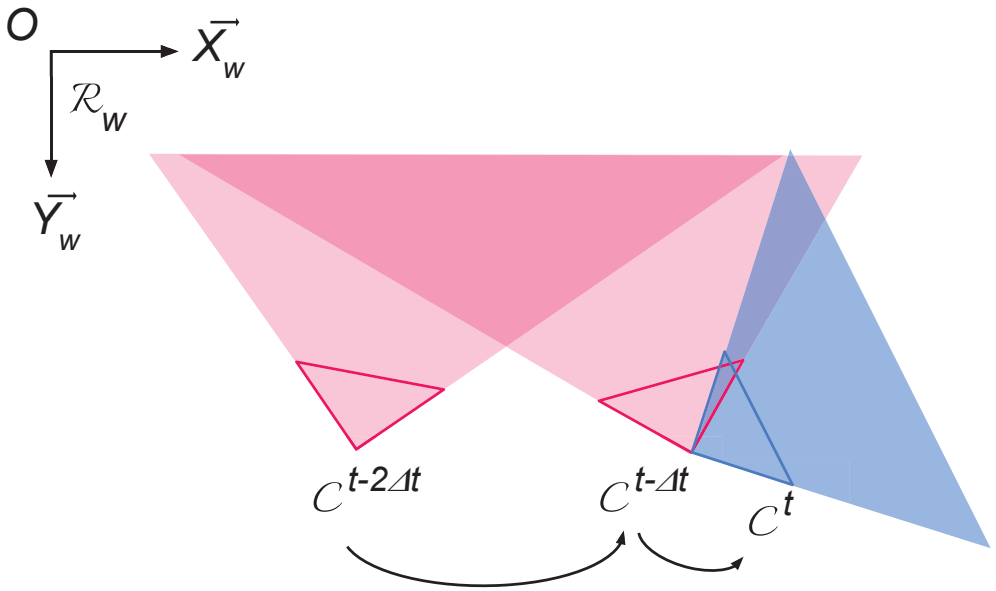


Fig. 5.1.: Schéma de la zone de recouvrement des champs de vision des caméras avec la caméra $\mathcal{C}^{t-2\Delta t}$ qui est en mouvement de translation et rotation et les caméras $\mathcal{C}^{t-\Delta t}$ et $\mathcal{C}^{t}$ qui sont uniquement en mouvement de rotation. Les caméras $\mathcal{C}^{t-2\Delta t}$ et $\mathcal{C}^{t-\Delta t}$ ont une grande partie de leur champ de vision en commun. Plusieurs points caractéristiques auront des correspondants entre les deux images de ces deux caméras. Les caméras $\mathcal{C}^{t-2\Delta t}$ et $\mathcal{C}^{t}$ n'ont aucune partie de leur champ de vision en commun donc aucun point caractéristique mis en correspondance.

Les points caractéristiques étiquetés statiques ou mobiles vont disparaître peu à peu au profit de nouveaux points étiquetés comme incertains. L'étiquetage de ces nouveaux points ne tendra ni vers un étiquetage statique ou mobile puisqu'ils ne pourront pas être reconstruits faute de correspondants dans l'image précédente utilisée. Pour palier cela, les valeurs de confiance des points qui n'ont pas d'estimation de leurs positions 3D peuvent être modifiées directement en analysant le flot optique. Les points dont le mouvement est cohérent avec un mouvement de rotation auront une valeur de confiance à 1 tandis que les autres seront à -1. Ainsi, l'ensemble stable de points caractéristiques continue d'être mis à jour pour pouvoir appliquer la contrainte 3D lorsque la caméra reprendra un mouvement de translation.

Intégration de l'analyse du mouvement apparent 2D dans la mise à jour de la valeur de confiance

La méthode que nous avons présentée dans le chapitre 4 réalise d'abord la contrainte 3D et la validation 2D dans un second temps. Toutefois, comme nous l'avons expliqué, il est plus cohérent que des points caractéristiques regroupés selon leurs mouvements apparents aient la même étiquette. De ce fait, il serait plus cohérent d'attribuer une valeur de mise à jour des points par groupes et non par individus. Pour ce faire, les valeurs de mises à jour seraient calculées dans un premier temps individuellement puis la valeur de mise à jour du groupe serait calculée sur la base des valeurs individuelles et appliquée à l'ensemble des points du groupe. En calculant cette valeur unique par une moyenne des valeurs individuelles (cf. équation 5.1), les valeurs bruitées seraient directement atténuées sans avoir besoin d'une seconde étape de validation qui réinitialise les valeurs de confiance et qui nécessite donc plusieurs mises à jour pour obtenir une étiquette statique ou mobile.

$$\begin{aligned} U_g^t &= \frac{1}{N}\sum_{i=0}^{N} U_i^t \\ Conf_i^t &= Conf_i^{t-1} + U_g^t, \text{ tel que } p_i^t \in g^t \end{aligned} \tag{5.1}$$

L'étiquetage des points caractéristiques via la contrainte 3D nécessite un suivi des points sur plusieurs images pour pouvoir réaliser les deux reconstructions nécessaires à la détection des objets mobiles. Dans le cas où les points caractéristiques apparaissent suffisamment longtemps sur plusieurs images consécutives, cela ne pose pas de problème. En revanche, les points qui apparaissent peu de temps dans les images n'auront pas le temps d'être étiquetés. Cela pose problème si ces points constituent la majeure partie de la scène statique visible par la caméra puisqu'ils ne pourront pas enrichir l'ensemble stable nécessaire à la réalisation de la contrainte

3D et cet ensemble tendrait à disparaître. En mettant à jour les valeurs de confiance via l'équation 5.1, les points suffisamment âgés pourraient influencer les points plus jeunes pour faire tendre plus rapidement leur étiquetage vers du statique ou du mobile. La figure 5.2 présente un exemple d'étiquetage obtenu avec notre méthode présentée dans le chapitre 4 ainsi que les groupes de points caractéristiques créés selon la similarité de leurs mouvements apparents. Les points caractéristiques trop jeunes affichés en blancs pourraient profiter des valeurs de mise à jour des autres points faisant partie du même groupe pour tendre vers un étiquetage statique/mobile plus rapidement.

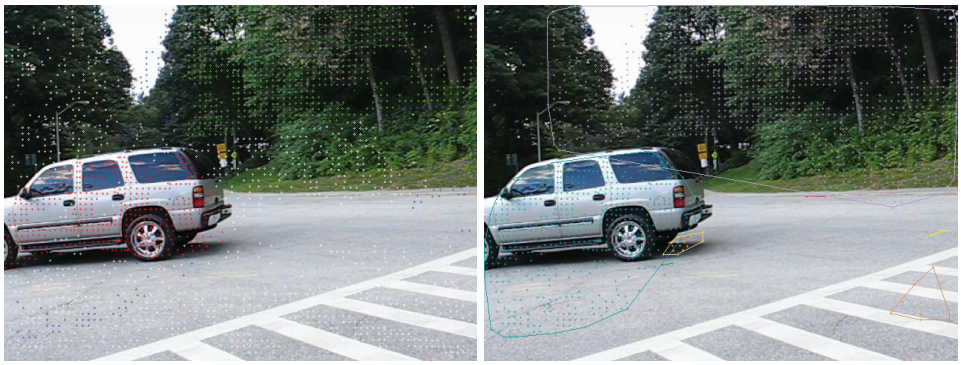


Fig. 5.2.: A gauche, les étiquettes avec les points verts étiquetés statiques, rouges étiquetés mobiles, bleus étiquetés incertains et en blanc les points trop jeunes pour avoir une étiquette. A droite, les points caractéristiques groupés selon la similarité de leurs mouvements apparents.

A

Notions fondamentales de la perception du mouvement

Sommaire

A.1 La perception du mouvement

Les êtres humains sont capables de détecter les objets/sujets en mouvement alors même qu'ils sont train de se déplacer. Dans ce chapitre, nous allons présenter les notions fondamentales de la perception du mouvement puis nous nous intéresserons aux différentes techniques utilisées pour capter le mouvement.

A.1.1 Le modèle de caméra

Le plus simple appareil qui permette de reproduire la vision est un trou dans un carton. De la même manière que les rayons lumineux traverse les différentes couches de l'œil pour venir produire une image renversée de l'environnement observé, le "trou" sert d'objectif pour former une image sur une surface plane (cf. figure A.1). On appelle ce dispositif un *sténopé*.

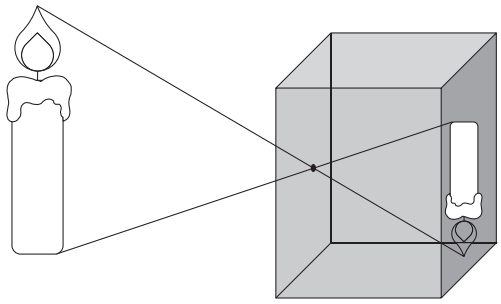


Fig. A.1.: Modèle de caméra sténopé.

Le centre optique - le "trou" - est situé sur le plan focal tandis que l'image se forme sur le plan image appelé aussi plan projectif. Les deux plans sont espacés d'une distance f appelée distance focale. Il existe trois repères : le repère monde qui correspond à l'environnement qui sera capté, le repère caméra qui correspond au centre optique et le repère image qui est généralement le coin supérieur gauche du plan image (cf. figure A.2). Le passage du monde 3D au repère image 2D peut être décomposé en trois transformations élémentaires : la transformation entre le repère monde et le repère caméra, la transformation entre le repère caméra et le repère capteur et la transformation entre le repère capteur et le repère image.

Transformation entre le repère monde et le repère caméra Un point 3D est localisé et orienté dans l'espace par une translation t et une rotation R par rapport au repère

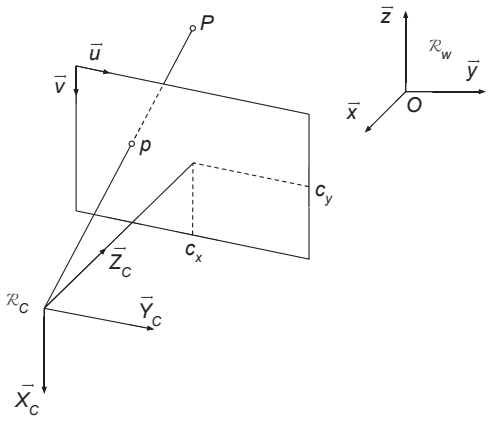


Fig. A.2.: Modèle de caméra projective.

monde. Soit t la translation et R la rotation qui permettent de situer le repère caméra dans le repère monde. Le changement de repère est défini par :

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = R \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + t \tag{A.1}$$

t_c et R_c sont appelés *paramètres extrinsèques* de la caméra et peuvent être regroupés en une seule matrice, la *matrice extrinsèque*.

Transformation entre le repère caméra et le repère capteur Un point de l'espace 3D $M = [XYZ]^\intercal$ se projette sur le plan projectif en un point 2D $m = [xyf]^\intercal$ d'après :

$$\begin{cases} x = f\frac{X_c}{Z_c} \\ y = f\frac{Y_c}{Z_c} \end{cases} \tag{A.2}$$

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{Z_c} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} \tag{A.3}$$

Transformation entre le repère capteur et le repère image Cette transformation convertit les coordonnées dans le repère capteur qui sont exprimées en mètres en coordonnées image exprimées en pixels :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_x & -k_x.\cos\Theta & c_x \\ 0 & k_y.\sin\Theta & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{A.4}$$

où c_x et c_y représentent la position du point principal de l'image en pixel (le centre), k_x et k_y représentent le nombre de pixels par unité de mesure en x et en y et Θ représente l'angle entre les lignes et les colonnes de l'image.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_x & 0 & c_x \\ 0 & k_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{A.5}$$

Ces paramètres sont les *paramètres intrinsèques* de la caméra et sont écris sous la forme d'une *matrice intrinsèque* dans l'équation A.4.

L'équation complète qui modélise la projection d'un point 3D du repère monde dans le plan 2D du repère image est :

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \sim \begin{bmatrix} k_x f & 0 & c_x & 0 \\ 0 & k_y f & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & -Rt \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{A.6}$$

A.1.2 La stéréoscopie et géométrie épipolaire

La sous section précédente détaille le processus de la projection perspective pour la formation d'une image sur le plan image d'une caméra et présente ainsi la relation qui existe entre le monde réel et l'image créée. Nous allons maintenant nous intéresser à la relation qui existe entre deux images prises de deux points de vues différents. Cette relation est modélisée par la *géométrie épipolaire*.

Soit $\mathcal{C}_1$ et $\mathcal{C}_2$ deux caméras situées à deux endroits différents dans le monde 3D. Soit P un point 3D visible par les deux caméras et p_1 et p_2 les projetés de P dans les deux caméras respectives. Le point P se projete en p_1 en suivant une droite de

projection. Soit l_1 le projeté de cette droite dans la seconde caméra (cf. figure A.3). l_1 est appelée *ligne épipolaire*.

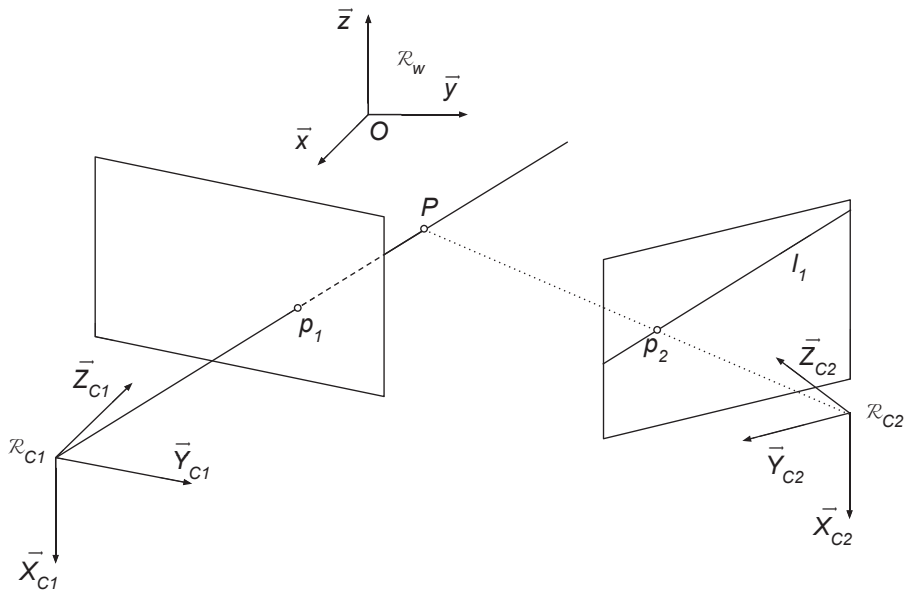


Fig. A.3.: Modèle stéréoscopique.

Puisque le point P est situé sur la droite de projection de la caméra $\mathcal{C}_1$, par projection le point p_2 doit se situer sur la droite épipolaire l_1. Cette relation géométrique qui lie les deux images est appelée *géoémtrie épipolaire* et la contrainte qui lie le projeté de P et l_1 est appelée *contrainte épipolaire*. Cette contrainte est souvent utilisée pour l'étape de mise en correspondance de caractéristiques entre deux images et est représentée par la *matrice fondamentale*.

A.1.3 Le mouvement apparent

Le mouvement apparent est le mouvement observé dans le flux vidéo de la caméra. Dans le cas d'une caméra en mouvement, il est composé du mouvement des objets/sujets en mouvement et du mouvement de l'ensemble des éléments composant l'environnement dû au mouvement de la caméra. Un *champ de mouvement* (motion field en anglais) est une construction idéale qui permet de décrire en 2D des mouvements 3D. Pour le moment, nous nous intéresserons aux mouvements apparents de la partie statique de la scène.

Le mouvement apparent d'un point de la scène entre deux images prises à l'instant t et $t+1$ est représenté par un vecteur 2D qui est en réalité la projection d'un vecteur de vélocité 3D provoqué par le mouvement de la caméra. Supposons que $R_c = I$ et $T = 0$ au temps t et que la caméra a fait un petit déplacement à $t+1$.

Soit $P = [X, Y, Z]^\intercal$ un point 3D dont la position 3D au temps $t + 1$ vaut $RP + T$. Le déplacement 3D de ce point est donc définit par $RP + T - P$.

$$V = -T - \omega \times P \tag{A.7}$$

où $T = [T_x, T_y, T_z]^\intercal$ représente ici un vecteur de vélocité - et non un vecteur de déplacement - et $\omega = [\omega_x, \omega_y, \omega_z]^\intercal$ la vélocité angulaire.

$$\begin{aligned} V_x &= -T_x - \omega_y Z + \omega_z Y \\ V_y &= -T_y - \omega_z X + \omega_x Z \\ V_z &= -T_z - \omega_x Y + \omega_y X \end{aligned} \tag{A.8}$$

D'après l'équation A.9, nous pouvons écrire que la projection perspective du point 3D P en un point 2D p est $p = \frac{fP}{Z}$, sans tenir compte des paramètres intrinsèques. En dérivant cette équation par rapport au temps, nous obtenons :

$$\frac{dp}{dt} = v = \frac{d\frac{fP}{Z}}{dt} \tag{A.9}$$

$$\frac{dp}{dt} = v = \frac{f}{Z^2}\left[\frac{dP}{dt}Z - P\frac{dZ}{dt}\right] \tag{A.10}$$

$$\frac{dp}{dt} = v = \frac{f}{Z^2}\left[V.Z - P.V_z\right] \tag{A.11}$$

$$\frac{dp}{dt} = v = f\frac{V}{Z} - p\frac{V_z}{Z} \tag{A.12}$$

En remplaçant V par l'équation A.9, l'équation du mouvement 2D devient :

$$v_x = \frac{T_z x - T_x f}{Z} - \omega_y f + \omega_z y + \frac{\omega_x xy}{f} - \frac{\omega_y x^2}{f} \tag{A.13}$$

$$v_x = \underbrace{\frac{T_z y - T_y f}{Z}}_{\text{Composante translationnelle}} \underbrace{+ \omega_x f - \omega_z x - \frac{\omega_y xy}{f} + \frac{\omega_x y^2}{f}}_{\text{Composante rotationnelle}} \tag{A.14}$$

Le mouvement apparent est dépendent du mouvement de la caméra ainsi que de la distance entre la caméra et le point 3D correspondant.

Translation Dans le cas où la caméra effectue une translation pure - sans rotation - parallèle au plan image de la caméra, le mouvement apparent est modélisé, d'après l'équation A.13, par l'équation suivante :

$$\begin{cases} v_x = -f\frac{T_x}{Z} \\ v_y = -f\frac{T_y}{Z} \end{cases} \tag{A.15}$$

Ainsi, tous les objets de la partie statique de la scène auront un mouvement apparent avec la même orientation mais avec des amplitudes différentes. D'après l'équation A.15, le seul paramètre qui dépend de la scène est Z. Plus un point 3D sera proche de la caméra, plus l'amplitude de son vecteur de flot optique sera grand et inversement. Dans le cas d'un mouvement de translation quelconque, les vecteurs des mouvements apparents seront tous parallèles les uns aux autres suivant la direction de la translation.

Prenons maintenant le cas général où les trois translations X, Y et Z, toujours sans rotation :

$$v_x = \frac{T_z x - T_x f}{Z} \tag{A.16}$$

$$v_y = \frac{T_z y - T_y f}{Z} \tag{A.17}$$

Soit un point p_0 tel que :

$$x_0 = f\frac{T_x}{T_z} \tag{A.18}$$

$$y_0 = f\frac{T_y}{T_z} \tag{A.19}$$

En remplaçant l'équation A.18 dans l'équation A.16, nous obtenons :

$$v_x = (x - x_0)\frac{T_z}{Z} \tag{A.20}$$

$$v_y = (y - y_0)\frac{T_z}{Z} \tag{A.21}$$

D'après l'équation A.20, le mouvement apparent issu d'une translation pure en Z est radial. Tous les vecteurs l'éloignent ou vont vers un même point qui est le point p_0. Dans le cas où les vecteurs s'éloignent, p_0 est appelé *focus d'expansion* (ou *focus of expansion* en anglais) et dans le cas contraire, p_0 est appelé *focus de contraction* (ou *focus of contraction* en anglais). Ce point représente la direction de la translation effectuée par la caméra. Ainsi, si la caméra effectue une translation pure en Z, p_0

sera le projeté du centre optique de la caméra sur le plan image qui est généralement le point principal. En plus d'être inversement proportionnels à leur profondeur, les vecteurs de flot optique des points 3D sont également proportionnels à la distance entre leur projeté 2D p et $p0$.

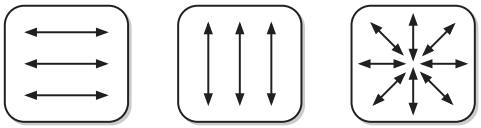


Fig. A.4.: Les trois types de mouvements apparents pour les trois types de translation. De gauche à droite : translation X, Y, Z.

Rotation Les mouvements apparents des rotations pures suivant l'axe des X et Y ressemblent à ceux des translations pures à la différence que la structure 3D de la scène n'intervient pas dans les calculs. Ainsi, les vecteurs de flot optique auront tous la même orientation et la même amplitude. Lorsqu'on ajoute la rotation selon l'axe des Z, chaque point de la scène décrit un mouvement circulaire autour d'un point O définit comme étant :

$$O_x = f\frac{\omega_x}{\omega_z} \tag{A.22}$$

$$O_y = f\frac{\omega_y}{\omega_z} \tag{A.23}$$

et où O est le point par lequel passe l'axe de rotation. Les mouvements observés résultent des composantes v_x et v_y qui sont des fonctions quadratiques de position dans l'image :

$$v_x = -\omega_y f + \omega_z y + \frac{\omega_x xy}{f} - \frac{\omega_y x^2}{f} \tag{A.24}$$

$$v_y = +\omega_x f - \omega_z x - \frac{\omega_y xy}{f} + \frac{\omega_x y^2}{f} \tag{A.25}$$

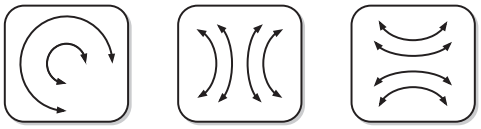


Fig. A.5.: Les trois types de mouvements apparents pour les trois types de rotation. De gauche à droite : rotation X, Y, Z.

Parallaxe du mouvement Lorsque la caméra se déplace, ce qui est statique apparait en mouvement. Nous avons vu que les mouvements observés dans le flux vidéo de la caméra dépendaient du mouvement mais aussi de la distance des objets à la caméra s'il y a translation. Dans ce cas, l'amplitude du vecteur de mouvement est inversement proportionnelle à la distance 3D entre le point et la caméra. Cette différence de mouvement est appelée la *parallaxe du mouvement*.

Annexe code

B

Sommaire

B.1 Présentation générale

Cette annexe présente les caractéristiques principales des deux codes qui ont permis de réaliser les expérimentations et de visualiser les différents résultats présents dans ce document de thèse. MoBDec (Moving Object Detection, prononcé *Moby Dick*) est le programme métier qui a permit de générer les résultats du chapitre 4. Il est organisé sous forme de modules qui utilisent une même structure de données généralisée à l'ensemble du programme. Achab est le second programme réalisé qui permet de visualiser de manière interactive les résultats obtenus avec MoBDec.

B.2 MoBDec

MoBDec a été développé en C++ afin d'utiliser les structures et les algorithmes présents dans la bibliothèque OpenCV 3.1.

Pour assurer la modularité du programme, les différentes étapes des méthodes que nous avons proposées ont été divisées en plusieurs modules qui réalisent chacun une tâche spécifique. Les données nécessaires à l'exécution d'une tâche sont stockées dans un seul objet, sobrement nommé *data* qui en plus de fournir des données en entrée, recueille les données produites par chacune des tâches.

L'ordre des modules a exécuter est fourni par l'utilisateur via un fichier de configuration qui contient toutes les informations nécessaires à l'exécution du programme : liste des modules, emplacement du dossier de la séquence d'image, numéro de la première image, de la dernière image, etc. Les différents seuils nécessaires à la méthode sont également spécifiés dans le fichier.

B.2.1 Enchaînement des modules

L'objet *LifeCycleManager* gère l'enchaînement des modules en lançant leur exécution de manière séquentielle durant plusieurs cycles. Un cycle correspond à un enchaînement de modules à un pas de temps donné. Un exemple de cycle de processus est présenté à la figure B.1. Les modules ne communiquent pas entre eux ce qui leur permet d'être totalement indépendants.

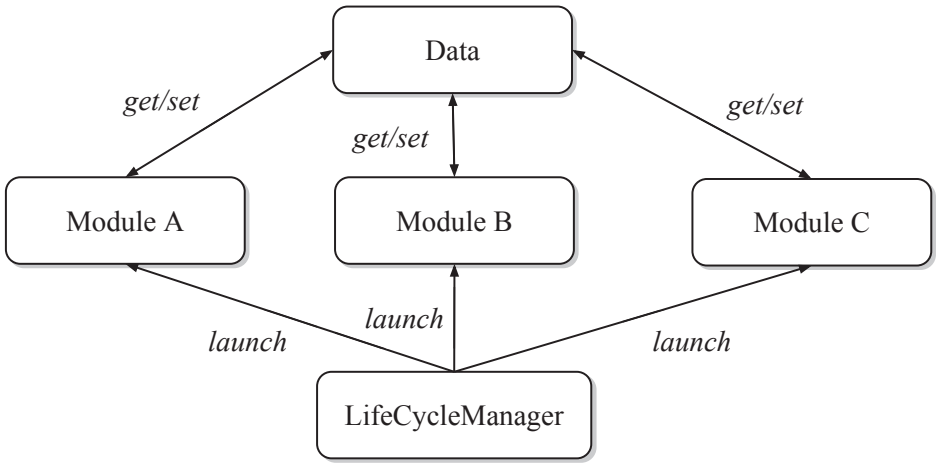


Fig. B.1.: Schéma d'un cycle de processus avec MoBDec.

B.2.2 Les modules

Chaque module réalise une tâche particulière en utilisant en entrée les données fournies par la structure de données générale et en l'enrichissant avec les nouvelles données obtenues à la fin du processus.

A l'heure de l'écriture de ce manuscrit, les modules sont répartis en quatre catégories :

— Détection de points caractéristiques

— Étiquetage

— Estimation du mouvement caméra

— Reconstruction

Le tableau B.1 présente une description des différents modules présents dans MoBDec qui ont été utilisés pour réaliser les résultats du chapitre 4. D'autres modules sont actuellement en cours de développement.

t_{start}	Module	Données en entrée	Données en sortie	Description
0	Ldof Brox [BM11]	Deux images	Points caractéristiques 2D	Extraction et suivi de points caractéristiques avec l'algorithme du LDOF de Brox [BM11]. La bibliothèque et le code d'utilisation fourni par les auteurs (https ://lmb.informatik.uni-freiburg.de/resources/software.php) a été adapté à notre structure.
$2\Delta t$	Labeling Section 4.2.3	Deux reconstructions	Étiquetage	Calcul des valeurs de mise à jour des tous les points caractéristiques en utilisant l'ensemble stable. Modification de la valeur de confiance des points.
$2\Delta t$	Validation2D Section 4.2.4	Étiquetage/Flot optique	Étiquetage	Création des groupes de points d'après le flot optique. Réinitialisation des valeurs de confiance des points étiquetés mobiles appartenant à un groupe comprenant des points étiquetés statiques.
Δt	EssentialMatrix Nistér [Nis04]	Points caractéristiques mis en correspondance	Rotation/Translation caméra	Estimation de la matrice essentielle d'après les points caractéristiques mis en correspondance entre deux images consécutives.
Δt	Scale Section 4.2.2	Deux reconstructions	Reconstruction remise à l'échelle	Estimation du facteur d'échelle entre deux reconstructions et application du facteur d'échelle à la seconde reconstruction.

Tab. B.1.: Présentation des modules de MoBDec. t_{start} correspond au numéro de la première image à laquelle le module peut être appliqué.

B.3 Achab

Le programme Achab permet de visualiser les différents résultats obtenus avec le programme MoBDec. Le code a été développé en C++ avec la bibliothèque Qt.

Achab est, au même titre que MoBDec, organisé sous forme de modules qui sont cette fois-ci graphiques. A chaque module de MoBDec correspond un module Achab qui permet de visualiser les données produites par le module. La figure B.2 présente un exemple de visualisation de données avec les modules de visualisation des points caractéristiques et d'étiquetage après validation 2D.

Fig. B.2.: Présentation Achab. En haut à gauche un menu déroulant pour sélectionner le projet à afficher, à gauche la liste des modules utilisés pour le projet sélectionné, au centre les résultats sélectionnés affichés.

La liste des projets réalisés avec le programme MoBDec est présente sous forme de liste déroulante nommée *MoBDec project* en haut à gauche. La liste des modules utilisée pour le projet sélectionné est affichée sur la gauche sous le nom *Results*. Les différents résultats produits par les différents modules sont affichables au centre de l'application en les sélectionnant via les cases à cocher. Il est possible de sélectionner un élément, ici un point caractéristique, sur l'un des affichages au centre afin de voir les données calculées par les modules qui s'affichent dans la partie *Results*. Le changement d'image se fait via les boutons situés en haut au centre. Lorsqu'un point est sélectionné, il est possible de suivre l'évolution des différents résultats qui le concernent en changeant d'image. L'avantage de cette application est d'avoir une

visualisation interactive qui permet d'analyser l'évolution de l'étiquetage des points caractéristiques au travers des différentes étapes de calculs.

Bibliographie

[Ade91] Edward H ADELSON. *Layered representation for image coding*. 1991 (cf. p. 23).

[AS95] Serge AYER et Harpreet S SAWHNEY. « Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding ». In : *Proceedings of IEEE International Conference on Computer Vision*. 1995, p. 777–784 (cf. p. 23).

[Bev+05] A BEVILACQUA, L Di STEFANO et P AZZARI. « An effective real-time mosaicing algorithm apt to detect motion through background subtraction using a PTZ camera ». In : *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* 2005, p. 511–516 (cf. p. 13).

[BFL06] Yuri BOYKOV et Gareth FUNKA-LEA. « Graph Cuts and Efficient N-D Image Segmentation ». In : *International Journal of Computer Vision* 70.2 (2006), p. 109–131 (cf. p. 28).

[Bha+00] K S BHAT, M SAPTHARISHI et P K KHOSLA. « Motion detection and segmentation using image mosaics ». In : *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*. T. 3. 2000, 1577–1580 vol.3 (cf. p. 13).

[BM10] Thomas BROX et Jitendra MALIK. « Object Segmentation by Long Term Analysis of Point Trajectories ». In : *Computer Vision – ECCV 2010 : 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V*. Sous la dir. de Kostas DANIILIDIS, Petros MARAGOS et Nikos PARAGIOS. Berlin, Heidelberg : Springer Berlin Heidelberg, 2010, p. 282–295 (cf. p. 27, 28).

[BM11] Thomas BROX et J MALIK. « Large Displacement Optical Flow : Descriptor Matching in Variational Motion Estimation ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.3 (2011), p. 500–513 (cf. p. 36, 49, 90).

[Bou+08] Thierry BOUWMANS, Fida EL BAF et Bertrand VACHON. « Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey ». In : *Recent Patents on Computer Science* 1.3 (2008), p. 219–237 (cf. p. 10).

[Bou14] Thierry BOUWMANS. « Traditional and recent approaches in background modeling for foreground detection : An overview ». In : *Computer Science Review* 11–12 (2014), p. 31–66 (cf. p. 9, 11).

[BS14] M BERGER et L M SEVERSKY. « Subspace Tracking under Dynamic Dimensionality for Online Background Subtraction ». In : *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, p. 1274–1281 (cf. p. 26).

[BZ14] Thierry Bouwmans et El Hadi Zahzah. « Robust PCA via Principal Component Pursuit : A review for a comparative evaluation in video surveillance ». In : *Computer Vision and Image Understanding* 122 (2014), p. 22–34 (cf. p. 11).

[Can+11] Emmanuel J Candès, Xiaodong Li, Yi Ma et John Wright. « Robust Principal Component Analysis ? » In : *J. ACM* 58.3 (2011), 11 :1–11 :37 (cf. p. 11).

[Cha+17] Marie-Neige Chapel, Erwan Guillou et Saida Bouakaz. « Coupled 2D and 3D Analysis for Moving Objects Detection with a Moving Camera ». In : *12th International Conference on Computer Vision Theory and Applications*. Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017). Porto, Portugal, 2017, p. 236–245 (cf. p. 65, 68).

[Che+08] Chung-Hao Chen, Yi Yao, David Page et al. « Heterogeneous Fusion of Omnidirectional and PTZ Cameras for Multiple Object Tracking ». In : *IEEE Transactions on Circuits and Systems for Video Technology* 18.8 (2008), p. 1052–1063 (cf. p. 12).

[Che+09] Chung-Chen Chen, Yi Yao, Anis Drira, Andreas Koschan et Mongi Abidi. « Cooperative mapping of multiple PTZ cameras in automated surveillance systems ». In : *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, p. 1078–1084 (cf. p. 12).

[Cui+12] Xinyi Cui, Junzhou Huang, Shaoting Zhang et Dimitris N Metaxas. « Background Subtraction Using Low Rank and Group Sparsity Constraints ». In : *Computer Vision – ECCV 2012 : 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I*. Sous la dir. d'Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato et Cordelia Schmid. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 612–625 (cf. p. 26, 28).

[Cui+14] Zhigao Cui, Aihua Li et Ke Jiang. « Cooperative Moving Object Segmentation using Two Cameras based on Background Subtraction and Image Registration. » In : *Journal of Multimedia* 9.3 (2014), p. 363–370 (cf. p. 13, 14).

[DN14] J DeGol et M Nam. « A clustering approach for detecting moving objects captured by a moving aerial camera ». In : *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, p. 6538–6542 (cf. p. 19).

[DP91] T Darrell et A Pentland. « Robust estimation of a multi-layered motion representation ». In : *Proceedings of the IEEE Workshop on Visual Motion*. 1991, p. 173–178 (cf. p. 23).

[ED07] E Eade et T Drummond. « Monocular SLAM as a Graph of Coalesced Observations ». In : *2007 IEEE 11th International Conference on Computer Vision*. 2007, p. 1–8 (cf. p. 29).

[EE12] Ali Elqursh et Ahmed Elgammal. « Online Moving Camera Background Subtraction ». In : *Computer Vision – ECCV 2012 : 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI*. Sous la dir. d'Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato et Cordelia Schmid. Berlin, Heidelberg : Springer Berlin Heidelberg, 2012, p. 228–241 (cf. p. 27, 28, 60, 62, 65, 68, 69).

[EE13] Ali ELQURSH et Ahmed ELGAMMAL. « Online Motion Segmentation Using Dynamic Label Propagation ». In : *Proceedings of the 2013 IEEE International Conference on Computer Vision*. ICCV '13. Washington, DC, USA : IEEE Computer Society, 2013, p. 2008–2015 (cf. p. 27).

[Eve+07] I EVERTS, N SEBE et G A JONES. « Cooperative Object Tracking with Multiple PTZ Cameras ». In : *14th International Conference on Image Analysis and Processing (ICIAP 2007)*. 2007, p. 323–330 (cf. p. 11).

[FP+15] Jorge FUENTES-PACHECO, José RUIZ-ASCENCIO et Juan Manuel RENDÓN-MANCHA. « Visual simultaneous localization and mapping : a survey ». In : *Artificial Intelligence Review* 43.1 (2015), p. 55–81 (cf. p. 29).

[Har+01] M HARVILLE, G GORDON et J WOODFILL. « Foreground segmentation using adaptive mixture models in color and depth ». In : *Proceedings IEEE Workshop on Detection and Recognition of Events in Video*. 2001, p. 3–11 (cf. p. 9).

[HE03] E HAYMAN et J O EKLUNDH. « Statistical background subtraction for a mobile observer ». In : *Proceedings Ninth IEEE International Conference on Computer Vision*. 2003, 67–74 vol.1 (cf. p. 13).

[Heu04] Stephan HEUEL. *Uncertain Projective Geometry : Statistical Reasoning For Polyhedral Object Reconstruction (Lecture Notes in Computer Science)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2004 (cf. p. 30).

[HJ07] Bohyung HAN et Ramesh JAIN. « Real-Time Subspace-Based Background Modeling Using Multi-channel Data ». In : *Advances in Visual Computing : Third International Symposium, ISVC 2007, Lake Tahoe, NV, USA, November 26-28, 2007, Proceedings, Part II*. Sous la dir. de George BEBIS, Richard BOYLE, Bahram PARVIN et al. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 162–172 (cf. p. 11).

[Hu+15] Wu-Chih HU, Chao-Ho CHEN, Tsong-Yi CHEN, Deng-Yuan HUANG et Zong-Che WU. « Moving object detection and tracking from video captured by moving camera ». In : *Journal of Visual Communication and Image Representation* 30 (2015), p. 164–180 (cf. p. 19).

[HZ03] Richard HARTLEY et Andrew ZISSERMAN. *Multiple view geometry in computer vision*. Cambridge university press, 2003 (cf. p. 51).

[IA98] Michal IRANI et P ANANDAN. « A unified approach to moving object detection in 2D and 3D scenes ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.6 (1998), p. 577–589 (cf. p. 20, 21).

[Jin+08] Yuxin JIN, Linmi TAO, Huijun DI, Naveed I RAO et Guangyou XU. « Background modeling from a free-moving camera by Multi-Layer Homography Algorithm ». In : *2008 15th IEEE International Conference on Image Processing*. 2008, p. 1572–1575 (cf. p. 23, 24).

[JS04] Boyoon JUNG et Gaurav S SUKHATME. « Detecting moving objects using a single camera on a mobile robot in an outdoor environment ». In : *in International Conference on Intelligent Autonomous Systems*. 2004, p. 980–987 (cf. p. 18).

[JS15] S JEEVA et M SIVABALAKRISHNAN. « Survey on Background Modeling and Foreground Detection for Real Time Video Surveillance ». In : *Procedia Computer Science* 50 (2015), p. 566–571 (cf. p. 11).

[Kad+13] Zulaikha KADIM, Marizuana Md DAUD, Syaimaa Solehah M RADZI, Norshuhada SAMUDIN et Hon Hock WOON. « Method to detect and track Moving object in non-static PTZ camera ». In : *Int MultiConf Eng Comput Sci* 1 (2013) (cf. p. 14).

[Kan+03] Sangkyu KANG, Joon-Ki PAIK, Andreas KOSCHAN, Besma ABIDI et Mongi A ABIDI. « Real-time video tracking using PTZ cameras ». In : *Quality Control by Artificial Vision*. International Society for Optics et Photonics. 2003, p. 103–111 (cf. p. 13).

[Kan+05] Jinman KANG, Isaac COHEN, Gérard MEDIONI et Chang YUAN. « Detection and tracking of moving objects from a moving platform in presence of strong parallax ». In : *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*. T. 1. IEEE. 2005, p. 10–17 (cf. p. 22).

[Keu+15] M KEUPER, B ANDRES et Thomas BROX. « Motion Trajectory Segmentation via Minimum Cost Multicuts ». In : *IEEE International Conference on Computer Vision (ICCV)*. 2015 (cf. p. 27).

[Kim+13] Soo Wan KIM, Kimin YUN, Kwang Moo YI, Sun Jung KIM et Jin Young CHOI. « Detection of moving objects with a moving camera using non-panoramic background model ». In : *Machine Vision and Applications* 24.5 (2013), p. 1015–1028 (cf. p. 16, 17).

[Kim+16] S KIM, D W YANG et H W PARK. « A Disparity-Based Adaptive Multihomography Method for Moving Target Detection Based on Global Motion Compensation ». In : *IEEE Transactions on Circuits and Systems for Video Technology* 26.8 (2016), p. 1407–1420 (cf. p. 19).

[KK02] Qifa KE et T KANADE. « A robust subspace approach to layer extraction ». In : *Workshop on Motion and Video Computing, 2002. Proceedings.* 2002, p. 37–43 (cf. p. 23).

[KM07] G KLEIN et D MURRAY. « Parallel Tracking and Mapping for Small AR Workspaces ». In : *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. 2007, p. 225–234 (cf. p. 29).

[Kun+10] Abhijit KUNDU, K Madhava KRISHNA et C V JAWAHAR. « Realtime Motion Segmentation Based Multibody Visual SLAM ». In : *Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing*. ICVGIP '10. New York, NY, USA : ACM, 2010, p. 251–258 (cf. p. 30).

[Kun+11] A KUNDU, K M KRISHNA et C V JAWAHAR. « Realtime multibody visual SLAM with a smoothly moving monocular camera ». In : *2011 International Conference on Computer Vision*. 2011, p. 2080–2087 (cf. p. 30).

[LDW91] John J LEONARD et Hugh F DURRANT-WHYTE. « Mobile robot localization by tracking geometric beacons ». In : *IEEE Transactions on Robotics and Automation* 7.3 (1991), p. 376–382 (cf. p. 29).

[LG09] Feng LIU et M GLEICHER. « Learning color and locality cues for moving object detection and segmentation ». In : *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, p. 320–327 (cf. p. 18).

[MB94] Don MURRAY et Anup BASU. « Motion tracking with an active camera ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.5 (1994), p. 449–459 (cf. p. 14).

[MH00] A MITTAL et D HUTTENLOCHER. « Scene modeling for wide area surveillance and image synthesis ». In : *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*. T. 2. 2000, 160–167 vol.2 (cf. p. 12).

[Mia08] Ajmal MIAN. « Realtime face detection and tracking using a single Pan, Tilt, Zoom camera ». In : *2008 23rd International Conference Image and Vision Computing New Zealand*. 2008, p. 1–6 (cf. p. 12).

[Mig+] Davide MIGLIORE, Roberto RIGAMONTI, Daniele MARZORATI, Matteo MATTEUCCI et Domenico G SORRENTI. *Use a Single Camera for Simultaneous Localization And Mapping with Mobile Object Tracking in dynamic environments* (cf. p. 30).

[Mou+06] E MOURAGNON, M LHUILLIER, M DHOME, F DEKEYSER et P SAYD. « Real Time Localization and 3D Reconstruction ». In : *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. T. 1. 2006, p. 363–370 (cf. p. 29).

[Nar+13] Manjunath NARAYANA, Allen HANSON et Erik LEARNED-MILLER. « Coherent Motion Segmentation in Moving Camera Videos Using Optical Flow Orientations ». In : *Proceedings of the 2013 IEEE International Conference on Computer Vision*. ICCV '13. Washington, DC, USA : IEEE Computer Society, 2013, p. 1577–1584 (cf. p. 27).

[Nis04] David NISTER. « An efficient solution to the five-point relative pose problem ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.6 (2004), p. 756–770 (cf. p. 51, 90).

[Non+13] Y NONAKA, A SHIMADA, H NAGAHARA et R I TANIGUCHI. « Real-Time Foreground Segmentation from Moving Camera Based on Case-Based Trajectory Classification ». In : *2013 2nd IAPR Asian Conference on Pattern Recognition*. 2013, p. 808–812 (cf. p. 27).

[OB11] Peter OCHS et Thomas BROX. « Object segmentation in video : a hierarchical variational approach for turning point trajectories into dense regions ». In : *IEEE International Conference on Computer Vision (ICCV)*. 2011 (cf. p. 28).

[Och+14] Peter OCHS, Jitendra MALIK et Thomas BROX. « Segmentation of Moving Objects by Long Term Video Analysis ». In : *IEEE Trans. Pattern Anal. Mach. Intell.* 36.6 (2014), p. 1187–1200 (cf. p. v, vi, 27, 28, 60).

[Oga+05] A S OGALE, C FERMULLER et Y ALOIMONOS. « Motion segmentation using occlusions ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.6 (2005), p. 988–992 (cf. p. 26).

[Oli+99] Nuria OLIVER, Barbara ROSARIO et Alex PENTLAND. « A Bayesian Computer Vision System for Modeling Human Interactions ». In : *Computer Vision Systems : First International Conference, ICVS' 99 Las Palmas, Gran Canaria, Spain, January 13–15, 1999 Proceedings*. Berlin, Heidelberg : Springer Berlin Heidelberg, 1999, p. 255–272 (cf. p. 10).

[PT03] Fatih PORIKLI et Oncel TUZEL. « Human body tracking by adaptive background models and mean-shift analysis ». In : *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. 2003, p. 1–9 (cf. p. 10).

[Rob+09] Lionel ROBINAULT, Stéphane BRES et Serge MIGUET. « Real Time Foreground Object Detection using PTZ Camera. » In : *VISAPP (1)*. 2009, p. 609–614 (cf. p. 14).

[Rom+14] Andrea ROMANONI, Matteo MATTEUCCI et Domenico G SORRENTI. « Background subtraction by combining Temporal and Spatio-Temporal histograms in the presence of camera movement ». In : *Machine Vision and Applications* 25.6 (2014), p. 1573–1584 (cf. p. 16, 17).

[Saw+99] Harpreet S. SAWHNEY, Yanlin GUO, J ASMUTH et Rakesh KUMAR. « Independent motion detection in 3D scenes ». In : *Proceedings of the Seventh IEEE International Conference on Computer Vision*. T. 1. 1999, 612–619 vol.1 (cf. p. 21).

[SG99] Chris STAUFFER et W Eric L GRIMSON. « Adaptive background mixture models for real-time tracking ». In : *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on*. T. 2. IEEE. 1999, p. 246–252 (cf. p. 9, 10).

[She+09] Y SHEIKH, O JAVED et T KANADE. « Background Subtraction for Freely Moving Cameras ». In : *2009 IEEE 12th International Conference on Computer Vision*. 2009, p. 1219–1225 (cf. p. 26–28).

[SL03] D SKOCAJ et A LEONARDIS. « Weighted and robust incremental method for subspace learning ». In : *Proceedings Ninth IEEE International Conference on Computer Vision*. 2003, 1494–1501 vol.2 (cf. p. 11).

[SM14] F SCHUBERT et K MIKOLAJCZYK. « Robust Registration and Filtering for Moving Object Detection in Aerial Videos ». In : *2014 22nd International Conference on Pattern Recognition*. 2014, p. 2808–2813 (cf. p. 18, 19).

[Smi+04] P SMITH, T DRUMMOND et R CIPOLLA. « Layered motion segmentation and depth ordering by tracking edges ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.4 (2004), p. 479–494 (cf. p. 23).

[ST94] Jianbo SHI et Carlo TOMASI. « Good features to track ». In : *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE. 1994, p. 593–600 (cf. p. 13).

[TP90] William B THOMPSON et Ting-Chuen PONG. « Detecting moving objects ». In : *International journal of computer vision* 4.1 (1990), p. 39–57 (cf. p. 25, 32).

[VB11] Parisa Darvish Zadeh VARCHEIE et Guillaume-Alexandre BILODEAU. « People tracking using a network-based PTZ camera ». In : *Machine Vision and Applications* 22.4 (2011), p. 671–690 (cf. p. 12).

[Vis+15] Amitha VISWANATH, Reena Kumari BEHERA, Vinuchackravarthy SENTHAMILARASU et Krishnan KUTTY. « Background Modelling from a Moving Camera ». In : *Procedia Computer Science* 58 (2015), p. 289–296 (cf. p. 15).

[WA93] John Y A WANG et Edward H ADELSON. « Layered representation for motion analysis ». In : *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1993, p. 361–366 (cf. p. 23).

[WA94] John Y A WANG et Edward H ADELSON. « Representing moving images with layers ». In : *IEEE Transactions on Image Processing* 3.5 (1994), p. 625–638 (cf. p. 23).

[Wan+03] Chieh-Chih WANG, C THORPE et S THRUN. « Online simultaneous localization and mapping with detection and tracking of moving objects : theory and results from a ground vehicle in crowded urban areas ». In : *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*. T. 1. 2003, 842–849 vol.1 (cf. p. 29).

[Wan+07] Chieh-Chih WANG, Charles THORPE, Sebastian THRUN, Martial HEBERT et Hugh DURRANT-WHYTE. « Simultaneous Localization, Mapping and Moving Object Tracking ». In : *Int. J. Rob. Res.* 26.9 (2007), p. 889–916 (cf. p. 29).

[Wan+14] Yanli WAN, Xifu WANG et Hongpu HU. « Automatic moving object segmentation for freely moving cameras ». In : *Mathematical Problems in Engineering* 2014 (2014) (cf. p. 17).

[WS05] Hanzi WANG et D SUTER. « A re-evaluation of mixture of Gaussian background modeling [video signal processing applications] ». In : *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* T. 2. 2005, ii/1017–ii/1020 Vol. 2 (cf. p. 10).

[WT02] Chieh-Chih WANG et Chuck THORPE. « Simultaneous localization and mapping with detection and tracking of moving objects ». In : *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*. T. 3. IEEE. 2002, p. 2918–2924 (cf. p. 29).

[Xu+08] Zhifei XU, Irene Yu-Hua GU et Pengfei SHI. « Recursive error-compensated dynamic eigenbackground learning and adaptive background subtraction in video ». In : *Optical Engineering* 47.5 (2008), p. 57001–57011 (cf. p. 11).

[Xue+13] Kang XUE, Yue LIU, Gbolabo OGUNMAKIN, Jing CHEN et Jiangen ZHANG. « Panoramic Gaussian Mixture Model and large-scale range background substraction method for PTZ camera-based surveillance systems ». In : *Machine Vision and Applications* 24.3 (2013), p. 477–492 (cf. p. 13).

[Yi+13] Kwang Moo YI, Kimin YUN, Soo Wan KIM et al. « Detection of Moving Objects with Non-stationary Cameras in 5.8ms : Bringing Motion Detection to Your Mobile Device ». In : *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013, p. 27–34 (cf. p. 17).

[Yua+07] Chang YUAN, Gérard MEDIONI, Jinman KANG et Isaac COHEN. « Detecting Motion Regions in the Presence of a Strong Parallax from a Moving Camera by Multiview Geometric Constraints ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.9 (2007), p. 1627–1641 (cf. p. 22).

[Zam+14] Daniya ZAMALIEVA, Alper YILMAZ et James W DAVIS. « A Multi-transformational Model for Background Subtraction with Moving Cameras ». In : *Computer Vision – ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Sous la dir. de David FLEET, Tomas PAJDLA, Bernt SCHIELE et Tinne TUYTELAARS. Cham : Springer International Publishing, 2014, p. 803–817 (cf. p. 24).

[Zha+05] Yunchu ZHANG, Zize LIANG, Zengguang HOU, Hongming WANG et Min TAN. « An adaptive mixture Gaussian background model with online background reconstruction and adjustable foreground mergence time for motion segmentation ». In : *2005 IEEE International Conference on Industrial Technology*. 2005, p. 23–27 (cf. p. 10).

[Zho+13] Xiaowei ZHOU, Can YANG et Weichuan YU. « Moving Object Detection by Detecting Contiguous Outliers in the Low-Rank Representation ». In : *IEEE Trans. Pattern Anal. Mach. Intell.* 35.3 (2013), p. 597–610 (cf. p. 18).

[ZMI02] L ZEINIK-MANOR et M IRANI. « Multiview constraints on homographies ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.2 (2002), p. 214–223 (cf. p. 23).

[ZY14] Daniya ZAMALIEVA et Alper YILMAZ. « Background subtraction for the moving camera : A geometric approach ». In : *Computer Vision and Image Understanding* 127 (2014), p. 73–85 (cf. p. 24, 25, 34, 35).

Colophon

This thesis was typeset with LaTeX 2$_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.