Détection et localisation d'objets stationnaires par une paire de caméras PTZ

Constant Guillot, Quoc-Cuong Pham, Sayd Patrick, Christophe Tilmant, Jean-Marc Lavest

Détection et localisation d'objets stationnaires par une paire de caméras PTZ

Constant Guillot[1] Quoc-Cuong Pham[1] Patrick Sayd[1] Christophe Tilmant[2] Jean-Marc Lavest[2]

[1] CEA, LIST, Laboratoire Vision et Ingénierie des Contenus,
Point Courrier 94, F-91191 Gif-sur-Yvette, France
[2] LASMEA UMR 6602, PRES Clermont Université/CNRS,
63177 Aubière cedex, France

constant.guillot@cea.fr

Résumé

Dans ce papier, nous proposons une approche originale pour détecter et localiser des objets stationnaires sur une scène étendue en exploitant une paire de caméras PTZ. Nous proposons deux contributions principales. Tout d'abord, nous présentons une méthode de détection et de segmentation d'objets stationnaires. Celle-ci est basée sur la réidentification de descripteurs de l'avant-plan et une segmentation de ces blobs en objets à l'aide de champs de Markov. La seconde contribution concerne la mise en correspondance entre les deux PTZ des silhouettes d'objets détectées dans chaque image.

Mots Clef

Objet stationnaire, caméra PTZ, appariement stéréo.

Abstract

In this paper we propose a novel approach for detecting and localising stationary objects using a pair of PTZ cameras monitoring a wide scene. Our contribution can be separated into two parts. First we propose a stationary object detection and labelling algorithm. It relies on the re-identification of foreground blocks and an MRF framework to detect and make a distinction between the stationary objects of the scene. Second we propose a geometric approach for robustly matching the stationary silhouettes from two PTZ cameras. Our system is tested on challenging sequences which prove its robustness to occlusions even in an unknown non planar 3D scene.

Keywords

Stationnary object, PTZ camera, stereo matching.

1 Introduction

Le développement des systèmes de surveillance des lieux publics s'accompagne d'un besoin croissant en analyse automatique capable de détecter les situations critiques. L'objectif est d'assister les opérateurs pour augmenter les capacités de détection de ce type d'installation. La détection d'objets stationnaires est la première étape de nombreuses fonctions telles que la détection de bagages abandonnés ou volés, ou encore de véhicules stationnés. La principale difficulté de cette tâche, s'appuyant souvent sur des techniques de soustraction de fond, est la robustesse aux changements de luminosité et aux occultations. Du fait du nombre limité de capteurs, les objets à détecter sont souvent de faibles résolution.

Dans ce papier, nous proposons un nouveau système pour détecter des objets stationnaires sur un espace étendu avec une seule paire de caméras Pan-Tilt-Zoom (PTZ), effectuant un tour de garde. Chaque caméra va scanner la zone à surveiller en parcourant indéfiniment un ensemble de positions prédéfinies de façon à couvrir l'ensemble de la zone à haute résolution. Chaque position, que nous appellerons *vues* dans la suite, peut être considérée comme une caméra fixe fonctionnant à une très faible cadence d'acquisition.

La première contribution de ce papier concerne la détection et la segmentation des objets stationnaires dans chaque *vue*. La détection est faite par réidentification de descripteurs de texture. La segmentation consiste à regrouper les blocs appartenant à un même objet par l'utilisation de champs de Markov.

La seconde contribution se rapporte à l'appariement des silhouettes détectées dans les différentes vues. Les difficultés principales de cette phase d'appariement sont la prise en compte de la géométrie de la scène, pas nécessairement plane et la configuration éloignée des caméras. En effet, pour obtenir une bonne précision de mesure sur des zones étendues et minimiser les risques d'occultations, l'écartement des deux caméras doit être important. La contrepartie de cette différence de point de vue est bien sûr des changements forts d'aspect entre les deux caméras qui vont exclure un critère d'apparence pour la phase d'appariement.

Le papier est organisé comme suit : le paragraphe 2 présente les travaux relatifs à notre problématique. La détection des objets stationnaires est décrite dans le paragraphe 3, la phase d'appariement dans le paragraphe 4. Les résultats expérimentaux sont présentés dans le paragraphe 5.

2 État de l'art

Ces dernières années, de nombreux travaux se sont intéressés à la détection d'objets stationnaires. La robustesse aux occultations est une difficulté importante dans cette tâche. Mathew *et al*.[14] utilisent un mélange de Gaussiennes pour modéliser le fond et les objets. Les objets stationnaires sont détectés à partir de l'analyse des transitions des gaussiennes de l'état objet à fond. Guler *et al*.[10] proposent de suivre les objets mobiles de la scène et définissent pour chaque objet une probabilité d'endurance qui augmente quand l'objet ne correspond pas au fond. Porikli *et al*.[16] utilisent un modèle de fond court terme et un modèle long terme. Il est supposé que les objets stationnaires vont être intégrés au modèle court terme mais pas au modèle long terme. L'analyse de ces deux modèles permet d'alimenter une table d'accumulation qui va être incrémentée pour les pixels classés comme stationnaires. Liao *et al*.[12] utilisent 6 images binaires résultats de la détection des objets qui représentent une fenêtre de 30s d'analyse. Un ET logique entre ces masques donnent les objets stationnaires. Bien que cette méthode soit considérée comme une des plus efficaces [3], elle provoque de nombreuses fausses alarmes car un ou plusieurs objets en mouvement peuvent créer une zone objet sur chacune des 6 images. Ce travail a cependant été étendu récemment par Bayona *et al*.[4]. Les fausses alarmes sont évitées en couplant un masque binaire des zones en mouvement. Ce masque permet de détecter et d'écarter les régions générées par des objets en mouvement. Cependant, cette approche ne gère pas les cas où un objet stationnaire est occulté pendant une longue période puis de nouveau visible. La méthode ne crée pas de lien entre les deux périodes où l'objet a été visible ce qui est très pénalisant quand de nombreux objets mobiles sont présents et occultent régulièrement les objets stationnaires.
Dans notre approche, nous ré-identifions des descripteurs robustes qui combinés à une fonction de segmentation permettront la détection des objets stationnaires, notion plus riche qu'une simple carte binaire fond/forme.

Concernant la localisation d'objets par un système multi-caméras, Beynon *et al*.[5] font l'hypothèse d'objets déposés sur un sol plan afin de retrouver des coordonnées 3D à partir d'une seule image. Une fonction de coût basée sur la couleur, la forme du blob et sa position mesure la similarité d'observations 2D avec des observations 3D précédentes. Les auteurs utilisent un algorithme d'association linéaire pour un appariement optimal entre observations et objets suivis. Des heuristiques sont appliquées pour essayer de filtrer les objets dupliqués et gérer les occultations. Miezianko *et al*.[15] supposent également que la scène 3D est plane. Lorsqu'un objet est détecté, il est projeté sur le plan du sol grâce à une homographie. Les objets sont positionnés sur les maxima locaux de recouvrement dans l'orthoimage. Utasi *et al*.[17] définissent une énergie basée sur des primitives géométriques dépendant de la position et de la hauteur des objets. Cette énergie est maximale pour les configurations réalistes. La configuration optimale est recherchée en utilisant un processus stochastique itératif, appelé "Multiple Death and Birth Dynamics". Fleuret *et al*.[8] discrétisent le plan du sol à l'aide d'une grille régulière. La silhouette humaine est modélisée par un parallélépipède rectangle qui peut être placé sur chaque position de la grille et projeté dans l'image. L'analyse de la zone image ainsi obtenue permet d'évaluer la probabilité d'occupation de la position au sol par une personne. Khan *et al*.[11] introduisent une contrainte qui permet la fusion de l'information fond/forme extraites de plusieurs caméras à travers une contrainte d'occupation planaire. Cette contrainte apporte une robustesse aux occultations et permet la localisation des personnes dans le repère du plan du sol. Ces méthodes font souvent l'hypothèse d'un monde plan pour utiliser des homographies ou réduire la dimension de l'espace de recherche pour les processus d'optimisation.
Toutes ces approches nécessitent un grand nombre de caméras avec des points de vues très différents pour gérer les occultations. Pour notre part, notre système fonctionne avec seulement deux points de vue. Nous proposons un appariement direct entre régions 2D qui permet ensuite l'estimation de la position 3D et de la taille de l'objet stationnaire.

3 Détection d'objets stationnaires

La plupart des méthodes de détection d'objets stationnaires délivrent pour chaque pixel une réponse binaire "objet stationnaire" ou non. Notre méthode nuance cette réponse en associant à un pixel classé objet l'instant de la première apparition du descripteur courant. Ce paramètre va non seulement permettre de différencier un objet stationnaire d'un objet mobile sur un critère temporel, mais aussi de regrouper les "pixels stationnaires" de même âge sous un même label conduisant à la notion d'objet stationnaire.

3.1 Détection de régions stationnaires

Nous utilisons la soustraction de fond décrite dans Guillot *et al*.[9] qui s'appuie sur le calcul d'un descripteur de texture (SURF) sur une grille régulière de l'image de résolution 8×8 pixels. Pour chaque bloc, nous conservons le dernier descripteur qui a été classé "fond" ainsi que les descripteurs ayant été classés "objet" depuis la dernière apparition du fond. Si un descripteur d'un bloc de l'image courante est apparié au descripteur du modèle de fond, ce dernier est mis à jour (en fait remplacé) et les descripteurs objet sont supprimés. Sinon, il est comparé aux descripteurs des objets. Si un appariement est trouvé, alors le modèle de l'objet est mis à jour. Si aucun objet déjà enregistré ne correspond, un nouvel objet est créé. Pour chaque objet de la liste, on conserve en mémoire l'instant de sa première apparition. Pour le fond, on conserve sa dernière apparition. Pour qu'un bloc soit considéré comme appartenant à un objet stationnaire, son descripteur objet courant doit avoir été vu pour la première fois depuis un temps supérieur au seuil

fixé par l'application.

3.2 Labellisation des objets stationnaires

La segmentation générique d'une scène en objets reste un sujet de recherche très ouvert lorsque peu d'a priori sont disponibles. Pour notre part, notre objectif est de séparer les objets stationnaires détectés en fonction de leur moment d'apparition. Ainsi si deux objets apparaissent simultanément dans un même blob, ils seront considérés comme un seul objet. Les blocs images classés objets stationnaires vont ainsi être labélisés lorsqu'ils apparaissent simultanément, y compris dans les cas d'une occultation partielle temporaire. Par exemple, nous devons gérer le cas d'une valise déposée et partiellement occultée par son propriétaire. Lorsque la personne quitte la scène, la segmentation doit attribuer le même label aux blocs de la valise. Une telle prise de décision au niveau d'un bloc est très délicate sans une prise en compte du voisinage. Nous construisons une stratégie pour qu'un même label soit affecté à tous les blocs compatibles. Nous utilisons pour cela un champ de Markov.

Rappel sur les champs de Markov. Soit $G = (V, E)$ un graphe représentant une image. Chaque noeud $v \in V$ représente un pixel de l'image, et chaque arc $e \in E \subseteq V \times V$ correspond à une relation de voisinage. Soit L un ensemble d'étiquettes, on cherche le meilleur étiquetage $x \in L^{|V|}$ pour l'image. Pour cela, on définit une fonction d'énergie $E : L^{|V|} \rightarrow \mathbb{R}$ qui sera minimisée pour obtenir l'étiquetage optimal.

$$E(x) = \sum_{i \in V} D_i(x_i) + \sum_{(i,j) \in E} V_{ij}(x_i, x_j) \qquad (1)$$

où $D_i(x_i)$, le terme d'attache aux données, représente le coût d'affectation de l'étiquette x_i au noeud $i \in V$, et $V_{i,j}(x_i, x_j)$ le terme de régularisation qui représente le coût associé à l'attribution d'étiquettes différentes à des pixels voisins. La minimisation de cette énergie peut être faite par l'algorithme d'Alahari *et al.*[2] qui garantit une convergence rapide vers un étiquetage proche de la solution optimale. La minimisation de l'énergie se fait par la recherche d'une coupe minimale. Parmi les solutions possibles, les segmentations proposant le plus petit périmètre sont favorisés, afin de lisser les frontières des régions.

Définition de l'énergie. Nous définissons une fonction d'énergie E construite telle que $\hat{x} = \arg\min_x E(x)$ est une labellisation de l'image correspondant aux objets stationnaires visibles. Soit $L = \{l_{BG}, l_1, \ldots, l_n\}$ l'ensemble des labels, avec l_{BG} le label des blocs fond ou objets mobiles et $l_1, \ldots, l_n$ les labels de n objets distincts stationnaires. Nous construisons un graphe représentant l'image issue de la ré-identification de l'avant-plan (paragraphe 3.1). Dans cette image, est associé aux pixels l'âge du bloc objet (temps écoulé depuis la première apparition du descripteur courant). Nous utilisons un voisinage 4-connexité. La fonction énergie est définie comme suit :

$$D_i(l_{BG}) = 0 \qquad (2)$$

L'équation 2 impose que le coût pour un label "non-stationnaire"est égal à 0. Ce label est la valeur par défaut, si les conditions d'une labellisation "stationnaire" ne sont pas remplies.

$$\begin{aligned} D_i(x_i \neq l_{BG}) \quad = \quad & C - age_i + pTemporal_i(x_i) \\ & + pIncompatibility_i(x_i) \end{aligned} \qquad (3)$$

avec $C > 0$ le laps de temps nécessaire pour considérer un objet comme stationnaire et age_i l'âge du bloc i. Les pénalités $pTemporal$ et $pIncompatibility$ sont définies par :

$$pTemporal_i(x_i) = max(t_{x_i} - t_i - C, 0) \qquad (4)$$

où t_{x_i} est l'instant de la première affectation du label x_i dans l'image, et t_i est l'instant de la première apparition du descripteur du bloc i. Ce terme pénalise un écart entre le moment d'apparition de l'étiquette et la première observation de l'objet stationnaire. On autorise toutefois une apparition plus tardive du descripteur pour autoriser une occultation partielle initiale.

$$pIncompatibility_i(x_i) = max(t_{i,\emptyset} - t_{x_i} + C, 0) \qquad (5)$$

où $t_{i,\emptyset}$ est l'instant de la dernière classification "fond" du bloc i. Ce terme pénalise l'attribution d'un label créé antérieurement à la première classification du bloc comme "objet stationnaire".
De l'équation 3, nous observons que $D_i(x_i \neq l_{BG}) < D_i(l_{BG})$ seulement si $age_i > C$. Autrement dit, pour assigner un label "objet stationnaire" à un bloc, la première condition est que le bloc soit présent depuis un certain temps. Ensuite, ce sont les pénalités (équations 4 et 5) qui vont permettre de choisir parmi les labels "objets stationnaires".
Le terme de régularisation est défini par :

$$V_{ij}(x_i, x_j) = \begin{cases} \lambda_1 + \lambda_2 \exp^{-|age_i - age_j|^2} & \text{si } x_i \neq x_j \\ 0 \text{ si } x_i = x_j \end{cases} \qquad (6)$$

avec λ_1 et λ_2 positifs. λ_1 pondère la pénalisation d'une labellisation différente de 2 blocs voisins. Le terme exponentiel pondéré par λ_2 pénalise l'attribution de labels différents pour 2 blocs "objets stationnaires" de même âge. L'idée est de considérer que deux objets différents ne sont probablement pas apparus au même instant.

Prise en compte des occultations. L'utilisation du formalisme des Champs de Markov permet d'obtenir une segmentation des objets stationnaires visibles en minimisant notre fonction d'énergie E. Nous créons une image binaire par label. Quand un label est associé à un bloc le masque du label est mis à jour. Quand un bloc est classé fond, ce bloc est mis à 0 dans chaque image de masque puisqu'à priori

il n'y a plus d'objet à cette position. Inversement, tant que le fond n'est pas revu le label est maintenu. Le bien-fondé de cette approche est montré sur la figure 1 où la forme des objets est préservé en dépit d'occultations.

FIGURE 1 – Un masque binaire par label est utilisé. Cela permet de gérer indépendamment les différents objets.

4 Appariement stéréo

L'objectif est de proposer une approche permettant l'appariement d'objets (ensemble de blocs portant le même label) entre deux vues. Nous nous plaçons dans le contexte d'une paire de caméras calibrées et où les points de vue sont très éloignés afin de réduire les risques d'occultation totale. La contrepartie est que l'appariement basé sur la mise en correspondance d'indices visuels n'est pas réalisable.
Pour une paire de caméras et un volume 3D, il existe au moins 2 points appelés *points frontières* [7] qui sont visibles par les deux caméras (sauf cas d'occultation). pour ces points les plans épipolaires sont tangents à la surface de l'objet. Lorsque la paire d'images stéréo est rectifiée, les points frontières correspondent aux points extrêmes hauts et bas de la projection de l'objet. Comme ces points sont visibles dans les deux images, ils peuvent être utilisés pour un objet comme critère d'appariement entre les deux vues.

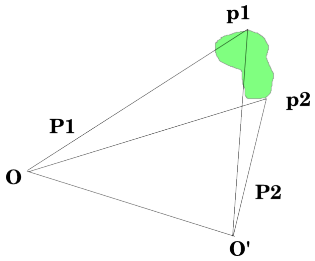


FIGURE 2 – Les plans épipolaires P_1 et P_2, tangents à l'objet, définissent les points frontières p_1 et p_2. O et O' sont les centres optiques des deux caméras.

4.1 Construction du graphe

Dans le cas où chaque objet est présent et est constitué d'une seule silhouette l'appariement est trivial. Mais, la

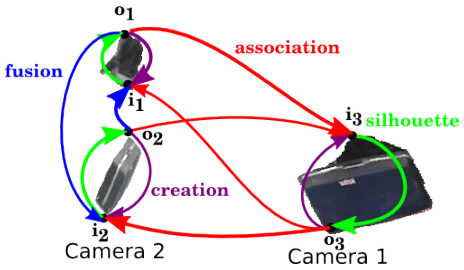


FIGURE 3 – Graphe orienté illustrant les quatre types d'arcs possibles pour un exemple avec trois silhouettes. Toutes les associations autorisées de points frontières sont ici représentées.

tâche est rendue compliquée par les erreurs de segmentation ou des occultations qui créent des silhouettes disjointes de l'objet. La présence de plusieurs objets spatialement proches génèrent des risques de confusion et des occultations. Un objet pouvant être représenté par plusieurs régions dans chaque image, il ne s'agit plus forcément d'apparier une silhouette avec une autre. Le problème revient à apparier un ensemble de silhouettes correspondant à un objet dans la caméra 1 à un ensemble de silhouettes pour cet objet dans la caméra 2.
Nous proposons d'apparier les silhouettes via leurs points frontières. Nous construisons un graphe orienté pour modéliser les associations autorisées de sorte qu'un objet soit représenté par un cycle reliant les différentes silhouettes constituant l'objet. Nous définissons 4 types d'arcs (illustré sur la figure 3) qui peuvent relier les points frontières, et auxquels sont attribués un coût.
Considérons deux silhouettes s_1, s_2 et leurs points frontière d'entrée et de sortie respectifs o_i et i_i.
Le coût d'un *arc de silhouette* est nul. Il permet de garantir l'unité de la silhouette.
Le coût d'un *arc d'association* entre deux points frontières est l'angle entre les plans épipolaires associés à chaque point. Comme illustré sur la figure 4. Son expression est :

$$c_{association} = |o_i - i_j| \tag{7}$$

Pour filtrer quelques erreurs d'appariements, le coût d'association de deux points frontières est mis à $+\infty$ si la triangulation de ces points donne un résultat incompatible avec la scène (seuil sur l'altitude). L'*arc de création* doit être sélectionné quand un objet est vu dans une seule caméra, et qu'il est donc considéré comme occulté dans l'autre caméra. Son coût est donné par :

$$c_{creation} = |o_i - i_i| \tag{8}$$

Le coût de *fusion* pour les silhouettes s_1 et s_2 est défini dans l'équation 9 :

$$c_{fusion} = (o_2 - o_1)^+ + (i_2 - i_1)^+ + (o_1 - i_2)^+ + d(s_1, s_2) \tag{9}$$

où $(.)^+ = \max(0, .)$ et $d(s_1, s_2)$ est la distance de Hausdorff entre les deux silhouettes dans les images rectifiées. Cette distance interdit la fusion de silhouettes éloignées dans l'image, en pénalisant les occultations trop larges. Ce coût est illustré sur la figure 5.

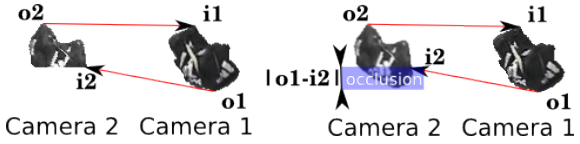


FIGURE 4 – Le coût d'association pour l'arc $o_1 \rightarrow i_2$. $|o_1 - i_2|$ représente le coût angulaire d'une occultation. La situation sur la gauche est interprétée comme étant la situation sur la droite.

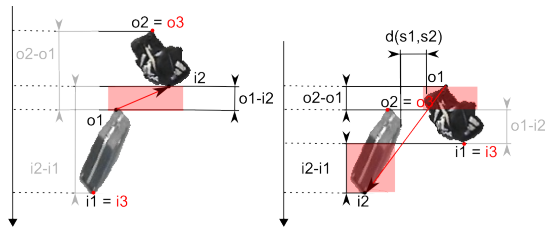


FIGURE 5 – Illustation du coût de deux *arcsde fusion* (en rouge) pour un paire de silhouette. Les coût non actifs (car nuls) sont grisés, les actifs sont mis en évidence par des rectangles rouges. Les points frontières entrant et sortant de la silhouette "virtuelle" résultante sont i_3 et o_3.

4.2 Phase de Pré-fusion

Pour une caméra donnée, les vues voisines d'un tour de garde se recouvrent. Un même objet peut apparaître dans plusieurs vues du tour de garde et ainsi générer plusieurs silhouettes dans une seule caméra. Pour simplifier la mise en correspondance inter-caméras nous procédons d'abord à la fusion de silhouettes se chevauchant entre les vues adjacentes d'une même caméra. Le critère pour cette pré-fusion de silhouettes est la consistance temporelle et la compatibilité de label. L'hypothèse est que les objets apparaissent dans le même ordre sur les deux vues et qu'un même label est assigné aux blocs d'un même objet.

4.3 Optimisation

Dans le paragraphe 4.1, les objets étaient représentés par un cycle dans le graphe des silhouettes et à chaque arc était assigné un coût. Le coût associé à un objet est la somme des coût des arcs constituant le cycle.
Nous avons défini un graphe orienté et pondéré dont les cycles correspondent à des associations de silhouettes. Par construction, on s'attend à ce que les associations ayant les poids les plus faibles correspondent aux observations des objets 3D réels. Par conséquent chercher les bonnes associations de silhouettes revient à chercher la partition de coût minimal en cycles disjoints de notre graphe. Comme le nombre possible de telles partitions croît exponentiellement avec le nombre des silhouettes dans le graphe, nous proposons une heuristique efficace. Un noeud du graphe est sélectionné aléatoirement puis le cycle de poids minimal passant par ce noeud est trouvé avec un algorithme de Dijkstra modifié. La modification consiste à contraindre le nombre d'arcs d'association à ne pas dépasser deux pour un unique cycle. Le type de configuration que nous souhaitons éviter peut survenir en particulier lorsque les silhouettes ne sont pas parfaitement alignées dans les deux caméras (voir figure 6) .

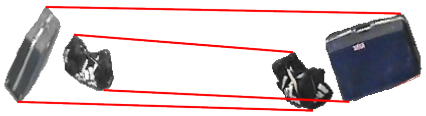


FIGURE 6 – Illustration d'un cycle ayant quatre arcs d'association. Pour éviter ce type de situation le nombre d'arc d'association est limité à 2 par cycle.

Dans un tel cas le cycle correspond à plusieurs objets mais l'on ne sait pas précisément comment apparier les silhouettes entre elles. Comme on suppose qu'une silhouette ne peut appartenir qu'à un seul objet, les noeuds du cycle ainsi trouvés sont supprimés du graphe. Le processus est alors répété sur le sous graphe restant jusqu'à obtention d'un graphe vide. Puisque cette approche dépend de l'ordre dans lequel les noeuds sont sélectionnés, elle est répétée plusieurs fois. Finalement c'est la partition de coût minimal qui sera conservée.

5 Expérimentations

Ce paragraphe consacré aux résultats expérimentaux se divise en deux parties. Tout d'abord, la détection d'objets stationnaires est évaluée dans le cas standard d'une caméra fixe. Ensuite, nous présentons les résultats obtenus avec une paire de caméras PTZ.
La détection d'objets stationnaires est d'abord testée sur les séquences publiques I-Lids AVSS2007 [1], puis sur des séquences plus complexes en terme d'occultation à gérer.
Pour la séquence ILids, nous considérons qu'un objet est stationnaire 60 secondes après s'être immobilisé. Les résultats et la vérité terrain peuvent être trouvé dans la Table 1. Les vérités terrain des séquences "Parked Vehicle" sont celles fournies sur le site, en revanche, nous avons dû adapter celles des séquences "bagages abandonnés" qui ne correspondent pas exactement au cas des "objets stationnaires". La Table 1 donne le moment de l'immobilisation de l'objet et le moment de sa suppression sur les séquences"Parked Vehicle". Un exemple de détection est présenté sur la figure **??**.
La fin de la détection "objet stationnaire" arrive systématiquement quelques secondes après le temps de fin de la vérité terrain. En effet, pendant quelques secondes le fond n'est pas ré-observé. Tant que ce n'est pas le cas, on maintient l'hypothèse que l'objet stationnaire est toujours pré-

Séquence	Tps debut VT	Tps debut estim	Tps finVT	Tps fin estim
AB Easy	2 :20	2 :20	3 :14	3 :18
AB Medium	1 :58	1 :58	3 :02	3 :03
AB Hard	1 :51	1 :52	3 :07	3 :11
PV Easy	2 :48	2 :48	3 :15	3 :21
PV Medium	1 :28	1 :28	1 :47	1 :56
PV Hard	2 :12	2 :12	2 :33	2 :35

TABLE 1 – Résultats de détection sur la base I-Lids [1]. Les détections d'objets stationnaires dépassent systématiquement la vérité terrain, car on attend de revoir le fond de la scène pour considérer l'objet comme disparu.

sent au cas ou celui-ci ne serait qu'occulté. Ce phénomène est visible sur la figure 7.

Les objets stationnaires détectés sur ces séquences hors des évènements de la vérité terrain sont causées par l'arrivée du train et à une personne assise très statique, avec sa valise. Ces "fausses alarmes" sont toutefois cohérentes avec la définition que nous avons donné d'un objet stationnaire (immobilité pendant 60 secondes). Les 8 objets stationnaires de la séquence AB et les 3 sur les séquences PV correspondent bien à des objets stationnaires. Nous n'avons pas constaté d'objets stationnaires non détectés.

FIGURE 7 – Lorsque l'objet est déplacé, il reste une ambiguïté tant que la zone est occultée (maintien de l'hypothèse).

Notre méthode est également confrontée à des séquences illustrant des configurations plus délicates à interpréter, les objets stationnaires y sont toujours occultés ou occultent pour un autre objet stationnaire.

La figure 8 montre la qualité de notre labellisation dans des cas d'occultations. De manière intéressante, le même label est attribué à deux composantes non adjacentes d'un même objet. Cependant, si deux objets sont détectés stationnaires au même instant ils porteront le même label. Notre approche ne distingue pas aujourd'hui ces deux scénarios.

FIGURE 8 – Gestion d'une occultation - un même label est attribué aux blocs correspondant à un même objet même si celui ci apparaît sous occultation partielle.

La figure 9 montre l'apport de la pénalité $pIncompatibility$ qui permet de segmenter correctement des bagages spatialement proches. arrivés à des instants différents. Toutefois, sur l'interface des deux objets, les deux labels sont équiprobables. C'est le terme de régularisation V_{ij} et plus particulièrement le facteur piloté par λ_2 qui intègre la consistance de l'âge qui permet d'obtenir une bonne segmentation. Sur la figure 9, les résultats de segmentation obtenu avec le terme piloté par $\lambda_2 > 0$ et $\lambda_2 = 0$. Dans ce dernier cas, l'algorithme d'optimisation privilégie le contour le plus court.

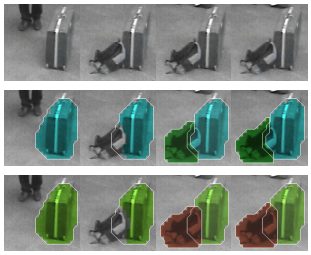
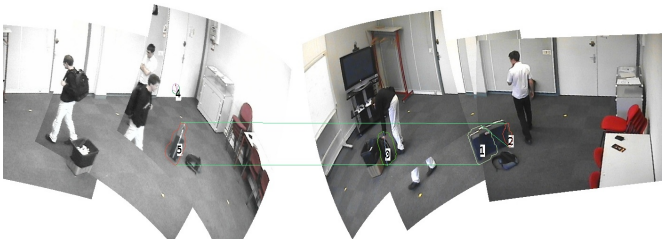


FIGURE 9 – Effet du terme de régularisation sur la segmentation. Haut : images originales. Milieu : $\lambda_2 = 0$, sans critère de consistance de l'âge, la segmentation de l'objet occultant est imparfaite. Bas : $\lambda_2 > 0$, avec le critère de consistance de l'âge la segmentation est correcte.

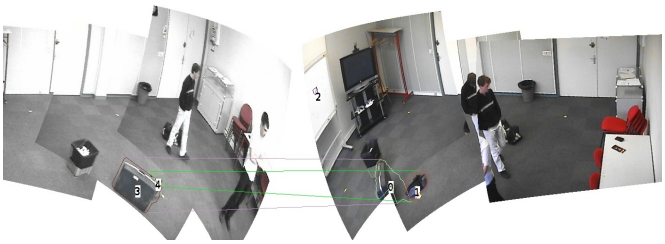
La seconde partie des expérimentations concerne l'évaluation de la fonction d'appariement. Les séquences sont acquises par deux caméras PTZ réalisant chacune un tour de garde. Pour toutes ces séquences la durée du tour de garde est de 15s, ce qui correspond à l'intervalle de temps entre deux acquisitions d'une même vue. Les deux tours de garde sont menés de façon asynchrone, ce qui explique les différences de positionnement des objets mobiles entre les deux panoramas et pourquoi la détection d'objets stationnaires n'est pas simultanée.

Le premier jeu de séquence a été acquis en intérieur, avec deux points de vue très différents pour les PTZ, et des occultations fortes comme le montre la figure 10. La figure 10(a) donne un exemple d'un objet détecté comme 2 silhouettes. Les silhouettes 1 et 2 sont reliées par un arc de fusion et correctement appariées avec la silhouette 5 dans l'autre caméra. Sur la figure 10(b), le sac formé des silhouettes 1 et 4 est presque complètement occulté dans la caméra gauche, malgré cela, l'appariement est réussi.

Le second jeu de séquences est acquis en extérieur et montre une scène non plane, puisque des objets sont placés au sol et sur un rebord de fenêtre (1,35m de haut). La baseline entre les 2 PTZ est de $13m$, et leur altitude est de 4,7m. Une phase de calibration permet d'estimer la géométrie épipolaire reliant les deux images panoramiques générer par les tour de garde. Les objets sont situés entre 15 et 20m des caméras. Chaque tour de garde est composé de 8 vues. La figure 11 montre les panoramas rectifiés et les points frontières associés. La figure 11(a) montre des appariements corrects dans un cas d'occultation. La valise



(a) Cas correct de fusion entre les silhouettes 1 et 2.



(b) En dépit d'une occultation quasiment totale les silhouettes 4 et 1 sont correctement appariées.

FIGURE 10 – Panoramas rectifiés d'une paire de caméras PTZ. Un identifiant est associé à chaque objet stationnaire. Les droites sont les arcs des cycles.

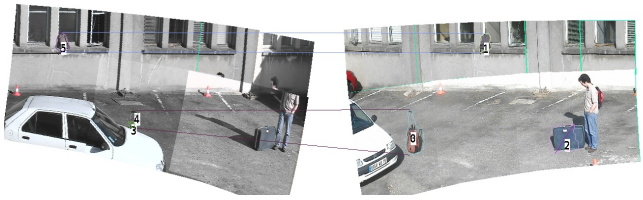
représentée par la silhouette 0 est fortement occultée dans la caméra gauche ; seulement le sommet et la poignée sont visible puisqu'une portion significative de la partie base de l'objet n'est pas visible. Du fait de ces occultations, la projection sur la plan du sol des détections faites dans chaque caméra ne s'intersecteraient pas. Cependant, notre approche basée sur l'exploitation d'un critère géométrique permet de trouver les associations expliquant au mieux les observations. Les figures 11(a) et 12 illustre la capacité de notre approche à gérer des altitudes différentes (scène non plane).

La figure 12 montre que toutes les ambiguïtés ne peuvent être levées. Si deux objets ont des points frontières sur les mêmes plans épipolaires le problème de l'association est insoluble sans autre a priori.

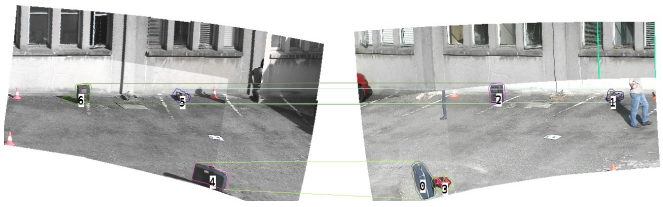
La table 2 montre les scores de rappel et précision sur les différentes séquences, tout d'abord en considérant les deux caméras comme indépendantes puis avec la mise en correspondances inter-caméra des silhouettes.

Pour être détectés par le système stéréo, les objets doivent être observés dans chaque caméra. De ce fait, l'approche stéréo n'augmente pas le taux de détection (même valeurs de rappel). En revanche, le filtrage géométrique réalisé par la phase d'appariement réduit le nombre de fausses alarmes et fait ainsi augmenter la précision. Cette caractéristique est essentielle pour un système de vidéo-surveillance où le taux de dérangement d'un opérateur doit être minimal.

Grâce à l'appariement entre les deux panoramas la position 3D et la taille des objets peuvent être reconstruites par triangulation des points frontières hauts et bas. Cette information 3D peut permettre de filtrer des données aberrantes issues de la phase de segmentation 2D. Ces informations

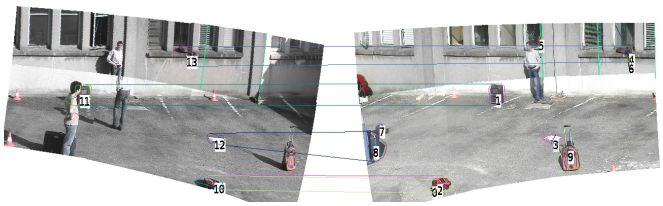


(a) Fusion correcte et appariement d'un objet non plat fortement occulté ($0 \leftrightarrow 3 \leftrightarrow 4$). Un objet situé au dessus du niveau du sol est également correctement apparié.

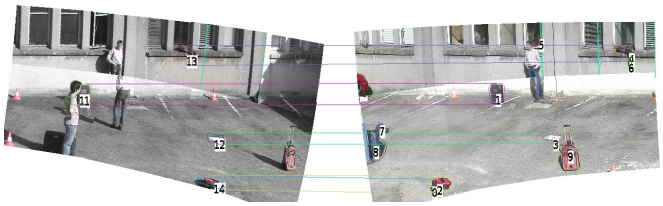


(b) Exemple d'appariements corrects.

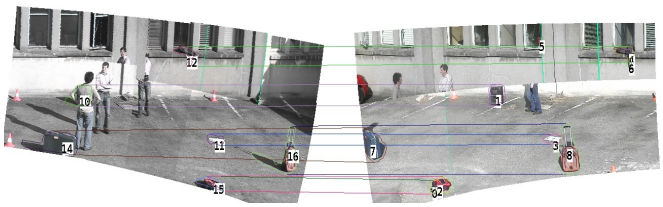
FIGURE 11 – Panoramas rectifiés pour une paire de caméras PTZ. Un identifiant est associé aux objets stationnaires. Les droites relient les points frontières des objets.



(a) Les choix $(12 \leftrightarrow 8, 3 \leftrightarrow 3)$ et $(12 \leftrightarrow 3, 8 \leftrightarrow 8)$ sont plausibles mais le cycle correspondant au faux appariement a le poids minimal.



(b) Une contrainte sur l'altitude maximale peut être ajoutée pour lever des ambiguïtés. L'association $12 \leftrightarrow 3$ est correctement selection-né et l'association $13 \leftrightarrow 4$ est conservée.



(c) Avec les mêmes paramètres que ce de la figure 12(a) l'ambiguïté est résolue lorsque chaque silhouette a un bon candidat à l'appariement dans la caméra opposée.

FIGURE 12 – Panoramas rectifiés d'images acquise par une paire de caméras PTZ. Illustrations des ambiguïtés d'appariement (Silhouettes 3, 8 and 12 figure 12(a)).

Séquence	Mono-caméra		Stéréo	
	Rappel	Précision	Rappel	Précision
Intérieur 1	0,99	0,63	0,99	0,88
Intérieur 2	1	0,63	0,93	0,80
Extérieur 1	0,95	0,86	0,95	0,92
Extérieur 2	0,95	0,81	0,91	0,81

TABLE 2 – Comparaison des statistiques calculées avec une approche mono-caméra à celles de l'approche stéréo.

Association	Altitude (m)	Taille estimée (m)	Taille réelle (m)
$12 \leftrightarrow 4$	$0,94$	$0,43$	$0,40$
$10 \leftrightarrow 1$	$-0,07$	$0,55$	$0,51$
$11 \leftrightarrow 3$	$-0,04$	$0,05$	$0,01$
$14 \leftrightarrow 7$	$0,09$	$0,48$	$0,53$
$16 \leftrightarrow 8$	$-0,02$	$0,97$	$1,05$
$15 \leftrightarrow 0$	$0,12$	$0,07$	$0,15$

TABLE 3 – Altitudes et tailles estimées des objets détectés figure 12(c). La colonne *Association* fait référence aux numéros des silhouettes dans la figure 12(c), chaque association correspond donc à un objet.

calculées sur l'image 12(c) sont données table 3.
Ces expériences ont été menées sur un PC quadricore de fréquence 2.4GHz et équipé de 3GB de RAM. L'étape de soustraction de fond prend 50 ms sur des images VGA, alors que les étapes de labellisation d'objets et d'appariement prennent chacune au maximum 10 ms.

6 Conclusion

Dans cet article, nous avons présenté une méthode de détection d'objets stationnaires à partir d'une paire de caméras PTZ. La première contribution concerne la détection et la labellisation des objets stationnaires dans chaque caméra. Cette méthode s'appuie sur une détection faite par ré-identification de descripteurs sur des blocs images et l'utilisation de champs de Markov pour assurer l'extraction d'objets à partir de la classification de ces blocs. La confrontation à de difficiles scènes issues de bases publiques ont montré la robustesse de notre approche. La seconde contribution porte sur une approche géométrique pour assurer l'appariement entre les détections faites dans une paire de caméras. Les contraintes épipolaires entre les points frontières des objets sont utilisées pour définir un graphe d'association. Les expérimentations ont montré la pertinence de notre approche vis-à-vis de scènes complexes pour la détection d'objets stationnaires incluant des scènes non planes et de fortes occultations.

Références

[1] i-lids dataset for avss 2007. http ://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html.

[2] K. Alahari, P. Kohli, and P. H. S. Torr. Reduce, reuse & recycle : Efficiently solving multi-label MRFs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[3] A. Bayona, J. SanMiguel, and J. Martinez. Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques. In *Advanced Video and Signal Based Surveillance.*, 2009.

[4] A. Bayona, J. SanMiguel, and J. Martinez. Stationary foreground detection using background subtraction and temporal difference in video surveillance. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010.

[5] M. D. Beynon, D. J. Van Hook, M. Seibert, A. Peacock, and D. Dudgeon. Detecting abandoned packages in a multi-camera video surveillance system. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, AVSS '03, 2003.

[6] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004.

[7] R. Cipolla, K. Astrom, and P. Giblin. Motion from the frontier of curved surfaces. In *Fifth International Conference on Computer Vision.*, 1995.

[8] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2008.

[9] C. Guillot, M. Taron, P. Sayd, Q.-C. Pham, C. Tilmant, and J.-M. Lavest. Background subtraction for ptz cameras performing a guard tour and application to cameras with very low frame rate. In *ACCV workshop on Visual Surveillance*, 2010.

[10] S. Guler, J. Silverstein, and I. Pushee. Stationary objects in multiple object tracking. In *Advanced Video and Signal Based Surveillance*, 2007.

[11] S. M. Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2009.

[12] H.-H. Liao, J.-Y. Chang, and L.-G. Chen. A localized approach to abandoned luggage detection with foreground-mask sampling. In *Advanced Video and Signal Based Surveillance.*, 2008.

[13] F. Lv, X. Song, B. Wu, V. Kumar, and S. R. Nevatia. Left luggage detection using bayesian inference. In *PETS*, 2006.

[14] R. Mathew, Z. Yu, and J. Zhang. Detecting new stable objects in surveillance video. In *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, 2005.

[15] R. Miezianko and D. Pokrajac. Localization of detected objects in multi-camera network. In *ICIP*, 2008.

[16] F. Porikli, Y. Ivanov, and T. Haga. Robust abandoned object detection using dual foregrounds. *EURASIP J. Adv. Signal Process*, 2008, January 2008.

[17] A. Utasi and B. Csaba. Multi-camera people localization and height estimation using multiple birth and death dynamics. In *ACCV workshop on Visual Surveillance*, 2010.

[18] B. Valentine, S. Apewokin, L. Wills, S. Wills, and A. Gentile. Midground object detection in real world video scenes. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007.