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
   prenom.nom@cea.fr
2. *Institut Pascal, UMR 6602 Université Blaise Pascal/CNRS/IFMA*
   F-63177 Aubière cedex
   prenom.nom@univ.bpclermont.fr

RÉSUMÉ. *Dans cet article, nous proposons une approche originale pour détecter et localiser des objets stationnaires sur une scène étendue en exploitant une paire de caméras PTZ. Tout d'abord, nous présentons une méthode de détection et de segmentation d'objets stationnaires. Celle-ci est basée sur la ré-identification de descripteurs de l'avant-plan et une segmentation de ces régions en objets à l'aide de champs de Markov. Notre méthode permet de dater l'avant-plan de la scène, et sous certaines conditions de segmenter différents objets contenus dans une seule composante connexe du premier plan. La seconde contribution concerne la mise en correspondance entre les deux caméras PTZ des silhouettes d'objets détectées dans chaque image. L'appariement, effectué à partir de contraintes purement géométriques, permet d'associer directement des ensembles de silhouettes. Notre système est finalement testé sur des séquences qui montrent sa robustesse aux occultations.*

ABSTRACT. *In this article we propose a novel approach for the detection and localisation of stationary objects using a pair of Pan-Tilt-Zoom (PTZ) cameras monitoring a wide scene. Our contribution is twofold. First we propose a stationary object detection and segmentation technique. It relies on the re-identification of foreground descriptors followed by a segmentation of these regions into objects, using Markov Random Fields. Our method allows the foreground to be dated and, under some conditions, to segment the different objects composing a single foreground blob. The second contribution concerns the matching of object silhouettes detected in each camera. This correspondence stage is only based on geometric constraints. Finally we tested our system on sequences which highlight its robustness to occlusions, even in the case of non planar scenes whose geometry is unknown.*

MOTS-CLÉS : *détection d'objets stationnaires, caméra PTZ, appariement stéréo.*

KEYWORDS: *stationary object detection, PTZ camera, stereo-matching.*

DOI:10.3166/TS.29.307-332

Extented abstract

In this article we propose a novel approach for the detection and localisation of stationary objects using a pair of Pan-Tilt-Zoom (PTZ) cameras monitoring a wide scene.

The context in which this work has been done is the detection of abandoned luggage, for security applications. Detection of stationary objects is a primary step to this aim. The main difficulty to this task, which often relies on background subtraction techniques, is the robustness to changes in illumination and to occlusions. Furthermore because of often limited the number of sensors, the resolution of the detected objects is low.

For these reasons we propose to use a pair of PTZ cameras, going through a guard tour, to monitor wide areas. Each camera independently monitors the scene by going through a predefined set of positions (pan, tilt, zoom) in order to cover the area at an adapted resolution, as illustrated in figure 1. Each of these positions, which we will refer to as a *view*, can be seen as an independent stationary camera with a very low frame rate.

The contribution of this paper is twofold. First, stationary objects are detected and labeled independently in each *view*. The labeling phase allows, in some cases, the distinction of several objects which are part of a single blob. This is done through the re-identification of the foreground through time and the minimisation of an energy function within an MRF framework. The second contribution consists in matching the silhouettes of the detected blobs from one camera to the other. The main difficulty stands in dealing with an arbitrary 3D scene (not necessarily planar) and with the large baseline between the two cameras. Indeed, in order to keep good precision in 3D measurements in a wide area, the baseline between the two cameras has to be large. In such a situation, the view points of the cameras on the objects might be so different that appearance cannot be used as a matching criterion.

First, in order to detect robustly the stationary regions independently in each *view*, a model of the background and foreground is created and updated through time, based on (Guillot *et al.*, 2010). The original image is decomposed in blocks of 8×8 pixels on which overlapping descriptors are computed. The background subtraction therefore generates an image whose pixels can be assimilated to the blocks of the original image.

To this aim, a descriptor is computed at each block. If it does not match the background model then it is checked against the foreground model. If a match is found in the foreground model then it is updated, otherwise a new foreground component is created and its time of creation is recorded. The foreground model at a specific block is emptied when background is observed. Thus, the output of the background subtraction stage is an image whose pixels contain 0 when background is observed, or the age of the foreground descriptor.

Although we could simply threshold the age of the foreground to find the stationary regions, we also want to label the different stationary regions of the scene, based

on their time of appearance. Segmenting unknown stationary objects is a very difficult problem which we will not try to address in the general case. For instance, if two objects appear at the same time and are detected as a single blob in the image, we do not try to separate them. What we want to do is to give different labels to objects appearing at different times while giving a single label to an object appearing under partial occlusion (for instance, a person partially occludes an item of baggage then leaves). This is not an easy task since at a block level (ie. ignoring the neighbourhood) it is impossible to state whether we are observing an object or an occluder. This goal is achieved by using a Markov Random Field (MRF) framework with $n + 1$ labels. One label is reserved for the background and non stationary objects, while the n others are for n stationary objects. Under this framework an energy function is defined, it quantifies the compatibility and the stationary object labels for each pixel of the image, taking into account the neighbouring pixels. Its unary part is the sum of three quantities. The first one indicates whether the pixel corresponds to a stationary region, the second one ensures the stability of the labelling, and the third one indicates if a label is not compatible with a pixel. This third compatibility criterion is based on the time the background was last seen at the pixel, and is an objective criterion that can ensure the presence of two distinct objects. The binary term of the energy is composed of two terms, the first one tends to give neighbouring pixels the same label, while the second one tends to give the same label to neighbouring pixels whose ages are similar . At the end of this segmentation stage, the output of the algorithm is a set of silhouettes representing the stationary regions detected in each view.

Second, the silhouettes from all the views from the two cameras are matched robustly to occlusion and segmentation errors. However the context we consider is not the standard stereo configuration where the baseline is small compared to the distance of object to the camera. In our case, objects may have a very different appearance in the two cameras and therefore it cannot be used as a matching criterion. Thus, when multiple objects are present in the scene the association of the blobs from one camera to the blobs from the other camera is not a trivial issue. As a consequence we propose to match particular points of the silhouettes instead of the silhouettes themselves. These points, called *frontier points* (Cipolla *et al.*, 1995), are visible by the two cameras and therefore naturally define a geometric criterion. In practice, in rectified images the two frontier points we use are the top and bottom point of each silhouette. The greater the angle between the two epipolar planes in which two candidate points for matching lie, the less likely they are to originate from the same 3D point. Since all frontier point associations are not possible we define a weighted directed graph to model them. This graph is constructed in such a way that its cycles are candidate silhouette associations. Finally, finding the best matching between silhouettes from the two cameras is equivalent to finding the vertex disjoint cycle cover of the graph of minimal cost, the cost of a cycle cover being the sum of the cost of all cycles. We propose a simple heuristic to find quickly an approximate solution. A random node n is selected and the shortest cycle starting from this node is computed using a modified version of the Dijkstra algorithm. The modification consists in constraining the number of association arcs in a cycle not to be greater than 2. Without this constraint,

cycles with 4 or more association arcs might be selected, which would mean there is more than one object in the cycle. Nodes selected in the cycle are then removed from the graph since we assume a silhouette belongs to only one object. The process is repeated until all nodes are selected. The final cycle cover depends on the order with which the nodes were selected; therefore to avoid local minima it is executed several times and the cycle cover of minimal cost is finally selected.

Then, in the experiment section, we show that our approach is capable of detecting and assigning stationary regions a coherent label, even in the case of challenging occlusions. The second part of the experiments shows that our algorithm allows the groups of silhouettes to be matched successfully and that the use of a second camera increases the accuracy of the detection.

1. Introduction

Le développement des systèmes de surveillance des lieux publics s'accompagne d'un besoin croissant d'outils d'analyse automatique capables de détecter les situations critiques. L'objectif est de fournir une assistance aux 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 observés à une faible résolution dans les images.

Dans cet article, 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, comme illustré en figure 1. Chaque caméra va parcourir la zone à surveiller en effectuant indéfiniment un ensemble de positions prédéfinies de façon à couvrir l'ensemble de la zone, à haute résolution. Chaque position, que nous appellerons *vue* par la suite, peut être considérée comme une caméra fixe fonctionnant à une très faible cadence d'acquisition.

Notre première contribution concerne la détection et la segmentation des objets stationnaires dans chaque *vue*. La détection est faite par ré-identification 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'occultation, l'écartement des deux caméras doit être important. La contrepartie de cette différence de point de vue repose

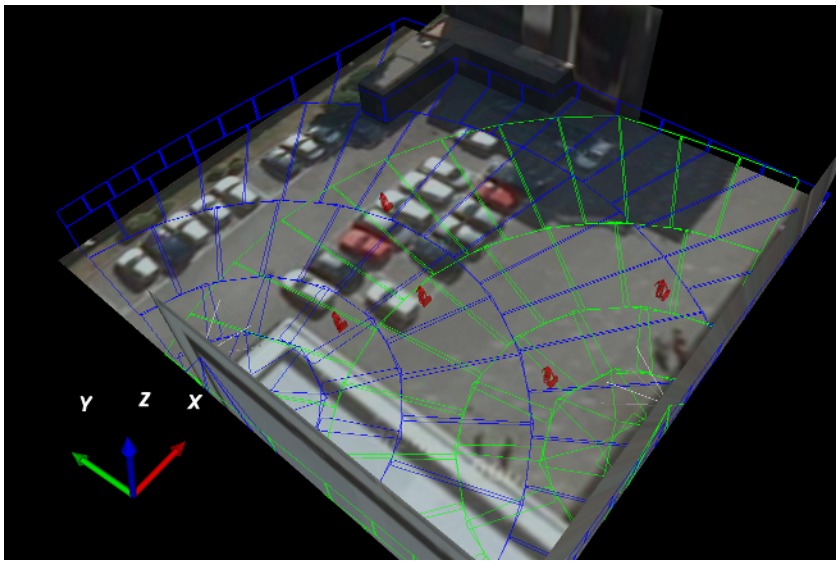


Figure 1. Exemple d'une scène surveillée par deux caméras PTZ. Les différentes vues sont représentées en gris et en noir

sur des changements importants de l'aspect des objets, d'une caméra à l'autre, qui vont interdire l'utilisation d'un critère d'apparence lors de la phase d'appariement.

Cet article [1] est organisé comme suit : la section 2 présente les travaux relatifs à notre problématique. La détection des objets stationnaires est décrite dans la section 3, la phase d'appariement dans la section 4. Les résultats expérimentaux sont présentés dans la section 5.

2. État de l'art

Ces dernières années, de nombreux travaux ont porté sur la détection d'objets stationnaires. La robustesse aux occultations est une difficulté importante dans cette tâche. (Mathew *et al.*, 2005) 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.*, 2007) 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.*, 2008) 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

1. Il s'agit d'une version étendue de (Guillot *et al.*, 2012).

terme. L'analyse de ces deux modèles permet d'alimenter une table d'accumulation qui est incrémentée pour les pixels classés comme stationnaires. (Liao *et al.*, 2008) utilisent six images binaires, résultats de la détection des objets, qui représentent une fenêtre de 30 secondes d'analyse. Un ET logique entre ces masques donnent les objets stationnaires. Bien que cette méthode soit considérée comme l'une des plus efficaces (Bayona *et al.*, 2009), elle provoque de nombreuses fausses alarmes car un ou plusieurs objets en mouvement peuvent créer une zone objet sur chacune des six images. Ce travail a cependant été étendu récemment par (Bayona *et al.*, 2010). 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 redevient 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 calculons sur des blocs de l'image des descripteurs robustes, ce qui nous permet de les dater et ré-identifier même après occultation. L'utilisation de l'âge des blocs de l'image dans une phase de segmentation permettra finalement la détection de différents objets stationnaires, notion plus riche qu'une simple carte binaire fond/forme.

Concernant la localisation d'objets par un système multicaméra, (Beynon *et al.*, 2003) 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 de la région 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, Pokrajac, 2008) 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, Csaba, 2010) 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.*, 2008) 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, Shah, 2009) introduisent une contrainte qui permet la fusion de l'information fond/forme extraite de plusieurs caméras à travers une contrainte d'occupation planaire. Cette contrainte apporte davantage de 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 vue 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

Bien que nous parlions de détection *d'objet* stationnaire, notre algorithme, de même qu'une grande majorité de ceux de la littérature, ne cherche pas réellement à détecter spécifiquement des *objets* mais plutôt des régions de l'image où le premier plan est stationnaire. Le terme *objet* est utilisé par abus car généralement seuls les objets sont effectivement stationnaires et l'application recherchée est bien leur détection.

La plupart des méthodes de détection d'objets stationnaires génèrent 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 une même étiquette conduisant à la notion d'objet stationnaire.

3.1. Détection de régions stationnaires

Nous utilisons la soustraction de fond décrite dans (Guillot *et al.*, 2010) qui s'appuie sur le calcul du descripteur de texture SURF (Bay *et al.*, 2008) sur une grille régulière de l'image de résolution 8×8 pixels. Pour chaque bloc de l'image, nous conservons le dernier descripteur qui a été classé « fond » ainsi que tous les descripteurs ayant été classés « premier plan » depuis la dernière apparition du fond. Ceci constitue donc un modèle du fond et un modèle du premier plan qui contient les différents objets observés.

Pour chaque bloc de l'image, si le descripteur courant est apparié à celui 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 du modèle du premier plan. 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 en ajoutant le descripteur courant à la liste correspondante. Pour chaque objet de la liste, on conserve en mémoire l'instant de sa première apparition. Pour le fond, on conserve sa dernière apparition. Ces informations seront utiles lors de la phase de détection et segmentation des objets stationnaires. Pour qu'un bloc soit considéré comme appartenant à un objet stationnaire, son descripteur objet courant doit avoir été vu pour la première fois à un temps supérieur au seuil fixé par l'application.

A chaque instant nous disposons donc de l'âge de chaque bloc du premier plan. Un seuil sur cet âge serait donc suffisant pour détecter les régions stationnaires. Nous

proposons cependant une segmentation plus riche, qui permet de différencier plusieurs objets de la scène, sous certaines conditions.

3.2. Étiquetage des objets stationnaires

La segmentation générique d'une scène en objets reste un sujet de recherche très ouvert lorsque peu *d'a priori* est disponible. Afin de rester générique au maximum nous avons décidé de n'utiliser que des informations temporelles dans le critère de segmentation. Ainsi si deux objets apparaissent simultanément et ne forment qu'une seule composante connexe, ils seront considérés comme un seul objet, et si deux objets apparaissent à des instants différents ils auront chacun une étiquette différente. Cependant cet étiquetage doit être capable de gérer notamment différents cas d'occultations. Par exemple, nous devons pouvoir gérer le cas d'une valise déposée partiellement occultée par son propriétaire. Lorsque la personne quitte la scène, la segmentation doit attribuer la même étiquette à tous les blocs de la valise, et ce même si l'ensemble des blocs de la valise n'ont pas le même âge. 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'une même étiquette soit affectée à tous les blocs compatibles. Nous utilisons pour cela un champ de Markov.

3.2.1. Rappel sur les champs de Markov

Soit $G = (V, A)$ un graphe représentant une image. Chaque nœud $v \in V$ représente un pixel de l'image, et chaque arc $a \in A \subseteq V \times V$ 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 $E : L^{|V|} \to \mathbb{R}$ à minimiser pour obtenir l'étiquetage optimal :

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

où $D_i(x_i)$, le terme d'attache aux données, représente le coût d'affectation de l'étiquette x_i au nœud $i \in V$, et $V_{i,j}(x_i, x_j)$, terme de régularisation, 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 proposé dans (Alahari *et al.*, 2008) 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 dans le graphe.

3.2.2. Définition de l'énergie

Nous définissons une fonction d'énergie E telle que $\hat{x} = \arg\min_x E(x)$ est un étiquetage de l'image correspondant aux objets stationnaires visibles. Considérons $L = \{l_{BG}, l_1, \ldots, l_n\}$ l'ensemble des étiquettes, avec l_{BG} l'étiquette des blocs fond ou objets mobiles et $l_1, \ldots, l_n$ les étiquettes de n objets distincts stationnaires. Nous

construisons un graphe représentant l'image issue de la ré-identification de l'avant-plan (section 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 d'énergie est définie comme suit :

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

L'équation (2) impose que le coût de l'étiquette « objet non stationnaire » soit égal à 0. Cette étiquette est la valeur par défaut, si les conditions d'une étiquette « objet stationnaire » ne sont pas remplies.

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

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

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

où t_{x_i} est l'instant de la première affectation de l'étiquette x_i dans l'image, et t_i est l'instant de la première apparition du descripteur du bloc i. Ce terme est positif pour les étiquettes créées après que le bloc soit considéré comme stationnaire. Il pénalise un écart entre le moment d'apparition de l'étiquette et la première observation de l'objet stationnaire. C'est ce terme qui permet donc de garantir qu'un objet se verra toujours affecter la même étiquette au cours du temps. On autorise toutefois que le descripteur apparaisse plus tardivement que l'étiquette, pour autoriser le cas où un objet apparaît partiellement occulté puis devient entièrement visible. Dans ce cas la pénalité est effectivement nulle.

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

où $t_{i,\emptyset}$ est l'instant de la dernière classification « fond » du bloc i. Ce terme pénalise l'attribution d'une étiquette créée antérieurement à la première classification du bloc comme « objet stationnaire » si le fond était observé à cet instant. En effet si un bloc appartient à un objet stationnaire mais qu'à un moment dans le passé on observait le fond au même endroit et qu'une étiquette e était alors déjà attribuée, alors à l'instant présent on sait que l'objet que l'on observe est différent de celui étiqueté e. Ce terme permet donc d'attribuer des étiquettes différentes à des objets dont on est certain qu'ils sont arrivés à leur place à des instants différents.

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 une étiquette « objet stationnaire » à un bloc,

la première condition est que le bloc soit présent depuis un temps supérieur à celui nécessaire pour qu'un objet soit considéré stationnaire. Ensuite, ce sont les pénalités (équations (4) et (5)) qui vont permettre de choisir parmi les étiquettes « objets stationnaires ». Le terme de régularisation est défini par :

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

avec λ_1 et λ_2 positifs. λ_1 pondère la pénalisation d'un étiquetage différent de deux blocs voisins. Le terme exponentiel pondéré par λ_2 pénalise l'attribution d'étiquettes différentes pour deux blocs « objets stationnaires » de même âge. Ce deuxième terme permet d'obtenir une meilleure segmentation au niveau de la zone de chevauchement lorsqu'un objet stationnaire en occulte un autre. Sans ce terme les blocs de cette zone peuvent se voir attribuer indifféremment l'étiquette de l'un ou l'autre objet. Une solution est donc d'utiliser le fait que les objets ne sont pas apparus au même instant.

3.2.3. Prise en compte des occultations

L'utilisation du formalisme des champs de Markov permet d'obtenir une segmentation des objets stationnaires visibles en minimisant la fonction d'énergie E. Nous créons une image binaire par étiquette. Quand une étiquette est associée à un bloc, le masque de l'étiquette est mis à jour. Quand un bloc est classé fond, ce bloc est mis à zéro dans chaque image de masque puisque *a priori* il n'y a plus d'objet à cette position. Inversement, tant que le fond n'est pas revu, l'étiquette est maintenue. Le bien-fondé de cette approche est montré sur la figure 2 où la forme des objets est préservée en dépit des occultations.

4. Appariement stéréo

L'objectif est de proposer une approche permettant l'appariement d'objets (ensemble de blocs portant la même étiquette) entre deux caméras. Nous nous plaçons dans le contexte d'une paire de caméras calibrées dont les points de vue sont très éloignés afin de réduire les risques d'occultation totale. La contrepartie est qu'un appariement reposant sur la mise en correspondance d'indices visuels est difficilement réalisable.

Pour une paire de caméras et un volume 3D, il existe au moins deux points appelés *points frontières* (Cipolla *et al.*, 1995) qui sont visibles par les deux caméras (sauf en 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 dans un critère d'appariement entre les deux caméras.

Figure 2. Les étiquettes objet sont maintenues jusqu'à réapparition du fond. Cette approche permet de maintenir une étiquette unique pour les blocs de la valise même après son occultation

Dans la suite les images utilisées sont donc rectifiées, de sorte que la différence d'ordonnées entre deux points correspond à l'angle entre les plans épipolaires auxquels ils appartiennent.

4.1. Construction du graphe

Dans le cas où chaque objet est présent et constitué d'une seule silhouette dans chaque caméra, l'appariement est trivial. Mais la tâche est rendue compliquée par les erreurs de segmentation ou des occultations qui peuvent segmenter un objet en plusieurs silhouettes. La présence de plusieurs objets spatialement proches génèrent des risques de confusion et des occultations. Un objet pouvant être représenté par plusieurs silhouettes 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 pour cela un graphe orienté et pondéré où les sommets représentent des points frontières et les arcs représentent les associations autorisées de points frontières (voir

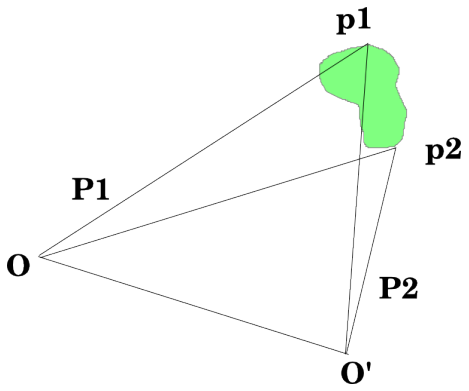


Figure 3. Les plans épipolaires P_1 et P_2, tangents à l'objet, définissent les points frontières p_1 et p_2. O et O' sont les centres optiques des deux caméras. En coordonnées rectifiées, les points haut et bas des silhouettes sont des points frontières et sont par conséquent visibles dans les deux caméras

figure 4). L'orientation du graphe permet d'obliger le passage par les deux points frontières d'une silhouette, et donc d'en conserver l'unité. On remarque que chaque silhouette a alors un point frontière entrant et un sortant (notés i et o), et leur position (bas ou haut) est inversée entre les deux caméras. Nous définissons quatre types d'arcs, dont les rôles et les coûts sont détaillés ultérieurement. De la façon dont les arcs sont construits il résulte que les objets (ou associations de silhouettes) sont représentés dans le graphe par un cycle. En choisissant les coûts des arcs de sorte qu'ils quantifient la vraisemblance de l'association, on définit alors naturellement le coût d'une association de silhouettes. Ce coût est la somme des coûts des arcs du cycle sélectionné. L'idée est alors de rechercher une partition de coût minimum du graphe en cycles dans une phase d'optimisation.

Nous allons maintenant définir les différents arcs du graphe et leur associer un coût reposant sur un critère purement géométrique. Par la suite, les coûts que nous utilisons sont des angles entre des plans épipolaires, ils peuvent donc être représentés par une distance selon l'axe vertical dans une paire d'images rectifiées. Les figures 4 et 5 montrent, sur un exemple, le graphe complet tel que nous le définissons ainsi que la partition en cycles disjoints attendue.

Considérons deux silhouettes s_1, s_2 et leurs points frontières d'entrée et de sortie respectifs o_i et i_i.

Un *arc de silhouette* relie le point frontière entrant au point frontière sortant d'une même silhouette et son rôle est d'en garantir l'unité. En conséquence, son coût est nul :

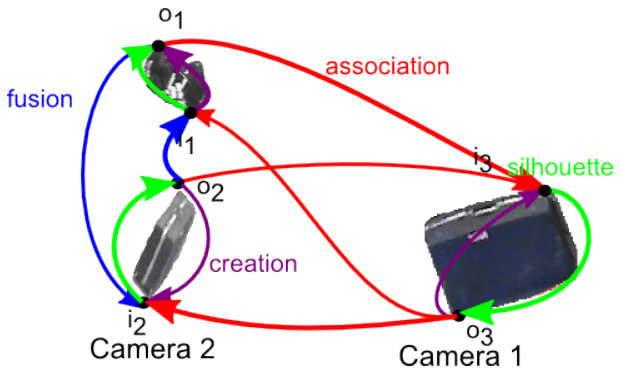


Figure 4. Graphe orienté illustrant les quatre types d'arcs possibles pour un exemple avec trois silhouettes. Toutes les associations autorisées de points frontières sont ici représentées. Les cycles du graphe correspondent aux différentes associations possibles de silhouettes

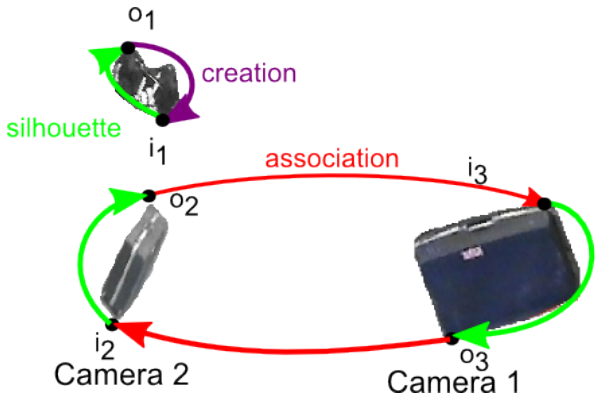


Figure 5. Partition en cycles disjoints souhaitée, calculée sur le graphe de la figure 4. Elle permet d'illustrer le rôle des arcs sur un exemple simple. L'arc de silhouette, de coût nul, est sélectionné pour chaque silhouette. Il en garantit l'unité. L'arc de création, n'est sélectionné que lorsqu'une silhouette ne peut être appariée

$$c_{silhouette} = 0 \tag{7}$$

L'*arc d'association* relie le point frontière sortant d'une silhouette au point frontière entrant d'une silhouette de l'autre caméra. Si cet arc est sélectionné, les deux points frontières seront considérés comme les points frontières hauts (ou bas) de l'objet. Le coût d'un tel arc est l'angle entre les plans épipolaires auxquels les points appartiennent. Ceci est illustré en figure 6. Son expression est :

$$c_{association} = |o_i - i_j| \tag{8}$$

Pour filtrer quelques erreurs d'appariements, le coût d'association de deux points frontières est mis à l'infini si la triangulation de ces points donne un résultat incompatible avec la scène (seuil sur l'altitude).

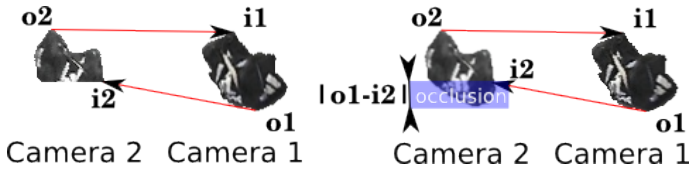


Figure 6. Coût d'association pour l'arc $o_1 \rightarrow i_2$. $|o_1 - i_2|$ représente le coût angulaire d'une occultation. En effet la situation sur la gauche est interprétée comme étant la situation sur la droite. Une association de coût non nulle n'est pas forcément mauvaise, puisqu'elle peut s'expliquer par une occultation dans l'une des caméras

L'*arc de création* permet de créer des cycles ne contenant qu'une seule silhouette. Cela correspond par exemple au cas où l'objet est observé dans une seule caméra, comme illustré en figure 5. Le coût de cet arc est donc l'angle entre les plans épipolaires portant les deux points frontières de la silhouette. Son expression est donc :

$$c_{creation} = |o_i - i_i| \tag{9}$$

L'*arc de fusion* relie le point frontière sortant au point frontière entrant de deux silhouettes d'une même caméra. De même que pour les autres arcs, son coût permet de quantifier des occultations ou des erreurs de segmentation. Son expression est cependant un peu plus complexe car deux cas sont à prendre en compte, ce qui est illustré en figure 7. Le coût de *fusion* pour deux silhouettes s_1 et s_2 est défini dans l'équation (10) :

$$c_{fusion} = (o_2 - o_1)^+ + (i_2 - i_1)^+ + (o_1 - i_2)^+ + d(s_1, s_2) \tag{10}$$

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 trop éloignées

dans l'image, en pénalisant les occultations trop larges. Ce coût est illustré sur la figure 7.

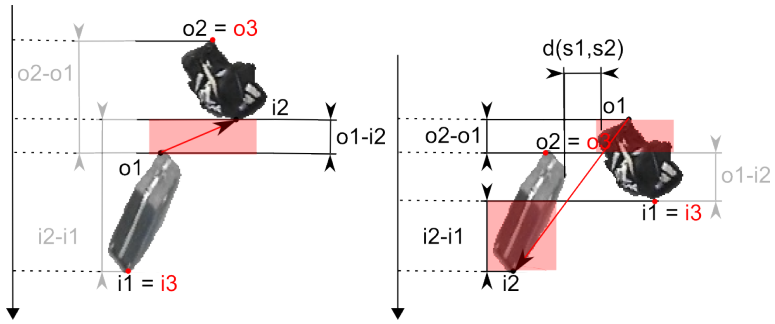


Figure 7. Illustation du coût de deux arcs de fusion pour une paire de silhouettes dans deux situations distinctes, à gauche et à droite. Les coûts non actifs (car nuls) sont grisés, les coûts actifs sont mis en évidence par des rectangles. Les points frontières entrant et sortant de la silhouette « virtuelle » résultant de la fusion sont i_3 et o_3

4.2. Phase préliminaire de 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 dans les vues adjacentes d'une même caméra. Le critère pour cette préfusion des silhouettes est la cohérence temporelle et la compatibilité d'étiquette. L'hypothèse est que les objets apparaissent dans le même ordre sur les deux vues et qu'une même étiquette est assignée aux blocs d'un même objet.

4.3. Optimisation

Dans la section 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 coûts 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 du graphe. Comme le nombre possible de telles partitions croît exponentiellement avec le nombre des silhouettes dans le graphe,

nous proposons une heuristique efficace. Un nœud du graphe est sélectionné aléatoirement puis le cycle de poids minimal passant par ce nœud 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 8) .

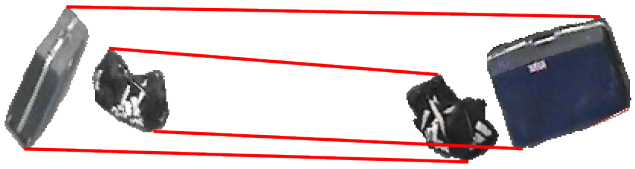


Figure 8. Illustration d'un cycle ayant quatre arcs d'association. Cela peut arriver lorsque les silhouettes sont mal alignées entre les deux caméras (erreur d'étalonnage ou de segmentation par exemple). Pour éviter ce type de situation le nombre d'arcs d'association est limité à deux par cycle

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 nœuds 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 nœuds sont sélectionnés, elle est répétée plusieurs fois. Finalement, c'est la partition de coût minimal qui est conservée.

5. Expérimentations

Cette section consacrée 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 afin d'exploiter des bases de données expertisées. 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 (*i-Lids dataset for AVSS 2007*, s. d.), puis sur des séquences plus complexes en termes d'occultation à gérer.

Pour la séquence I-Lids, nous considérons qu'un objet est stationnaire 60 secondes après s'être immobilisé. Les résultats et la vérité terrain peuvent être trouvés dans le tableau 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 ». Le tableau 1 donne le moment de l'immobilisation de l'objet et le moment de sa suppression sur les séquences « Parked Vehicle ».

Tableau 1. Résultats de détection sur la base I-Lids. 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

Séquence	Début vérité terrain (s)	Début détecté (s)	Fin vérité terrain (s)	Fin détectée (s)
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
(Guler *et al.*, 2007)		2:46		3:18
(Venetianer *et al.*, 2007)		2:52		3:16
PV Medium	1:28	1:28	1:47	1:56
(Guler *et al.*, 2007)		1:28		1:54
(Venetianer *et al.*, 2007)		1:43		1:47
PV Hard	2:12	2:12	2:33	2:35
(Guler *et al.*, 2007)		2:13		2:36
(Venetianer *et al.*, 2007)		2:19		2:34

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 réobservé. Tant que ce n'est pas le cas, on maintient l'hypothèse que l'objet stationnaire est toujours présent au cas où celui-ci ne serait qu'occulté. Ce phénomène est visible sur la figure 9.

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

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 10 montre la qualité de l'étiquetage obtenu dans des cas d'occultations. Une valise est déposée et est toujours partiellement occultée. Au bout du temps C elle commence à être partiellement détectée (à cause des occultations). De manière intéressante, le même étiquetage est attribué à deux composantes non adjacentes d'un même objet. Cependant, si deux objets sont détectés stationnaires au même instant, ils porteront la même étiquette. Notre approche ne distingue pas aujourd'hui ces deux situations.

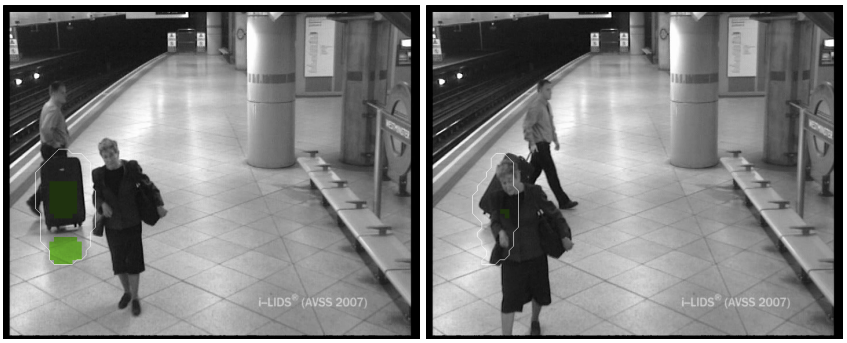


Figure 9. Lorsque l'objet est déplacé, il reste une ambiguïté tant que la zone est occultée (maintien de l'hypothèse). Pour cette raison la fin des alarmes est souvent détectée avec un léger retard

Figure 10. Gestion d'une occultation. Cette séquence montre une valise qui apparaît partiellement occultée. Tous les blocs de l'image qui la représentent n'ont donc pas le même âge. Notre système de pénalité permet cependant d'affecter une étiquette unique à la valise au fur et à mesure qu'elle se découvre

La figure 11 montre l'apport de la pénalité $pIncompatibility$ (équation (5)) qui permet de segmenter correctement des bagages spatialement proches arrivés à des instants différents. Toutefois, sur l'interface des deux objets, les deux étiquettes sont équiprobables. C'est le terme de régularisation V_{ij} et plus particulièrement le facteur piloté par λ_2 qui intègre la cohérence de l'âge qui permet d'obtenir une bonne segmentation. Sur la figure 11, les résultats de segmentation obtenus avec le terme piloté par $\lambda_2 > 0$ et $\lambda_2 = 0$ sont montrés. Dans ce dernier cas, l'algorithme d'optimisation privilégie le contour le plus court.

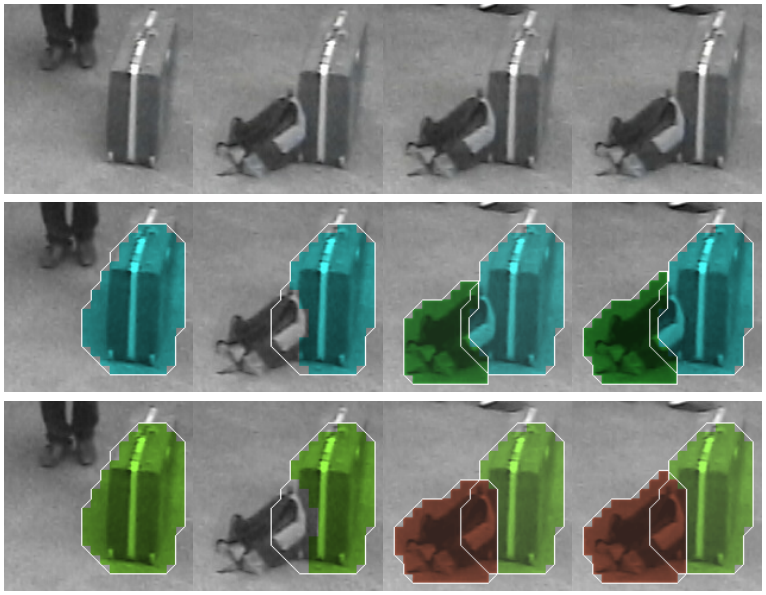
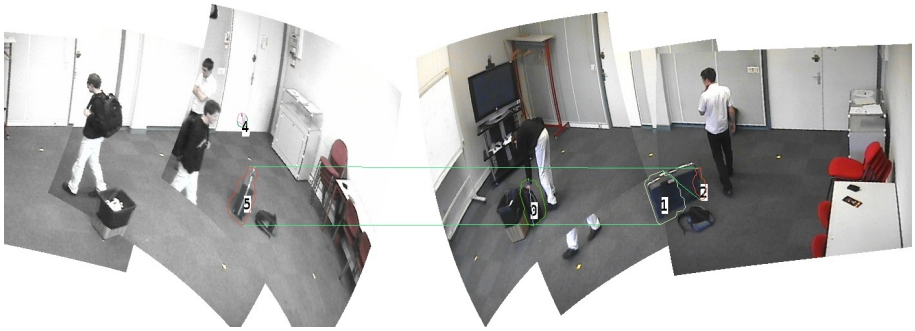


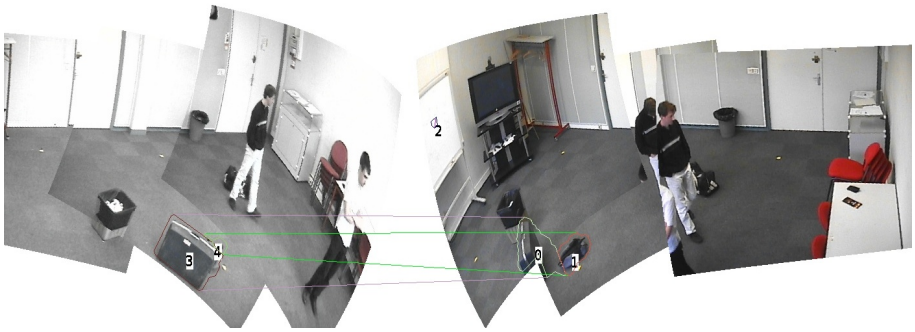
Figure 11. Effet du terme de régularisation sur la segmentation. Haut : images originales. Milieu : $\lambda_2 = 0$, sans critère de cohérence de l'âge, la segmentation de l'objet occultant est imparfaite. Bas : $\lambda_2 > 0$, avec le critère de cohérence de l'âge la segmentation est correcte

La seconde partie des expérimentations concerne l'évaluation de l'appariement de silhouettes entre une paire de caméras. 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 15 secondes, 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 entre les deux caméras.

Les paires d'images panoramiques présentées par la suite sont rectifiées, c'est-à-dire que les lignes horizontales correspondent aux droites épipolaires. Concrètement ces images sont construites par passage aux coordonnées sphériques. Un déca-



(a) Les silhouettes 1 et 2 sont correctement fusionnées, de sorte que la valise est détectée comme l'association des silhouettes 1, 2 et 3.



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

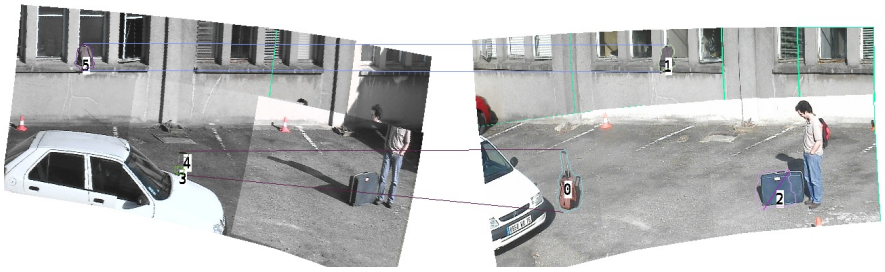
Figure 12. Panoramas rectifiés d'une paire de caméras PTZ. Un identifiant est associé à chaque objet stationnaire. Les droites sont les arcs des cycles

lage horizontal ou vertical en pixel constant correspond donc à un décalage angulaire constant.

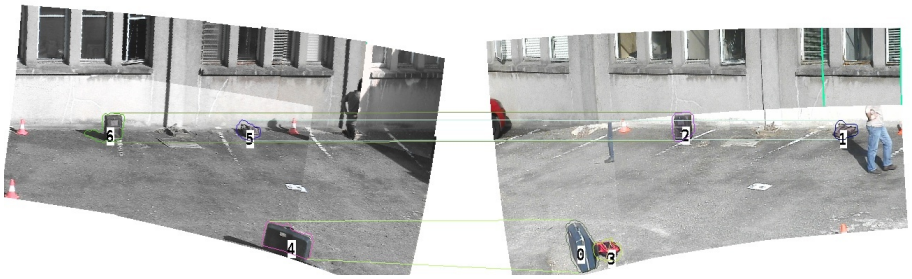
Le premier jeu de séquences a été acquis en intérieur, avec deux points de vue très différents pour les PTZ et des occultations fortes, comme le montre la figure 12. La figure 12(a) donne un exemple d'un objet détecté comme deux silhouettes. Les silhouettes 1 et 2 sont reliées par un arc de fusion et correctement appariées avec la silhouette 5 dans l'autre caméra. Sur la figure 12(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 (1,35 mètres de haut). La base stéréo entre les deux PTZ est de 13 mètres, et leur altitude est de 4,7 mètres. Une phase d'étalonnage permet d'estimer la géométrie épipolaire reliant les

deux images panoramiques générées par les tours de garde. Les objets sont situés entre 15 et 20 mètres des caméras. Chaque tour de garde est composé de 8 vues. La figure 13 montre les panoramas rectifiés et les points frontières associés. La figure 13(a) montre des appariements corrects dans un cas d'occultation. La valise représentée par la silhouette 0 est fortement occultée dans la caméra gauche, seulement le sommet et la poignée sont visibles puisqu'une portion significative de la partie basse de l'objet n'est pas visible. Du fait de ces occultations, la projection sur le plan du sol des détections faites dans chaque caméra ne s'intersecterait pas. Cependant, notre approche basée sur l'exploitation d'un critère géométrique permet de trouver les associations expliquant au mieux les observations. Les figures 13(a) et 14 illustrent la capacité de notre approche à gérer des altitudes différentes (scène non plane).



(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.

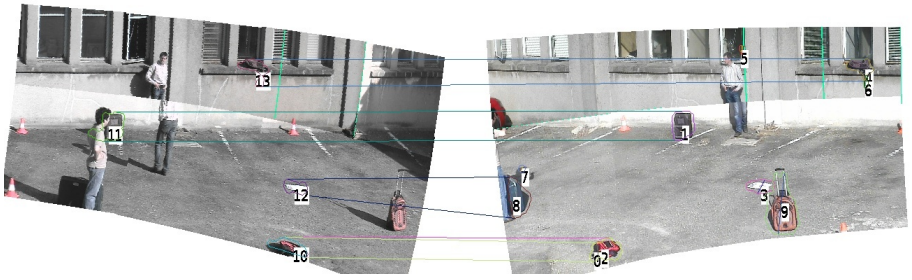


(b) Exemple d'appariements corrects.

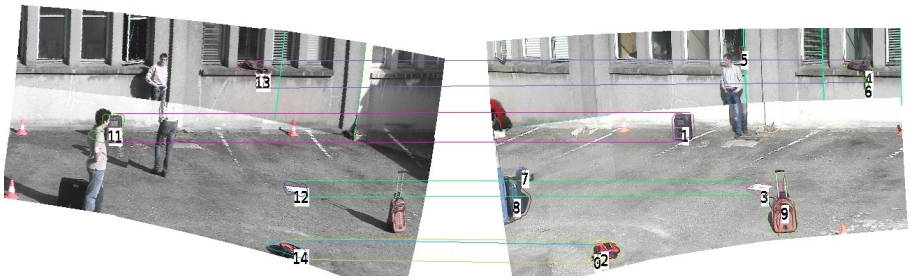
Figure 13. Panoramas rectifiés pour une paire de caméras PTZ. Un identifiant est associé aux objets stationnaires. Les droites relient les points frontières des objets

La figure 14 montre que toutes les ambiguïtés ne peuvent être levées. Si deux objets ont des points frontières sur les mêmes plans épipolaires alors le problème de l'association est insoluble sans autre a priori.

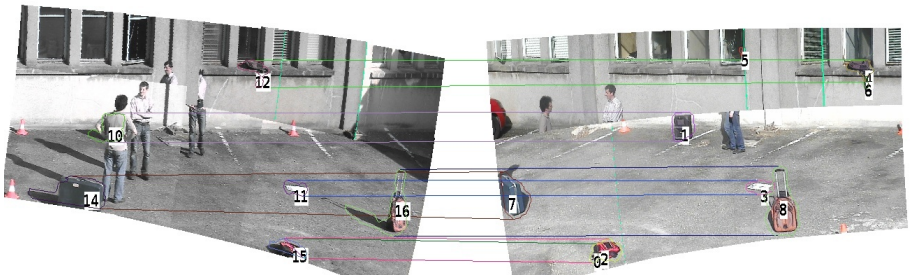
Le tableau 2 montre les scores de rappel et précision sur les différentes séquences, tout d'abord en considérant les deux caméras comme indépendantes puis avec la mise en correspondance intercaméra des silhouettes.



(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 sélectionnée et l'association $13 \leftrightarrow 4$ est conservée.



(c) Avec les mêmes paramètres que ceux de la figure 14(a) l'ambiguïté est résolue lorsque chaque silhouette a un bon candidat à l'appariement dans la caméra opposée.

Figure 14. Panoramas rectifiés d'images acquises par une paire de caméras PTZ. Illustrations des ambiguïtés d'appariement (Silhouettes 3*,* 8 *et* 12 *figure 14(a))*

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 (mêmes 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 vidéosurveillance où le taux de dérangement d'un opérateur doit être minimal.

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

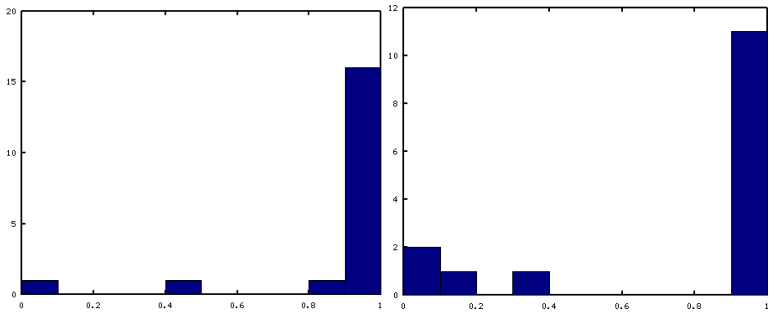
Séquence	Mono-caméra		Stéréo	
	Rappel	Précision	Rappel	Précision
Intérieur 1	0,99	0,63	0,99	0,88
Intérieur 2	1	0,63	0,93	0,80
Extérieur 1	0,95	0,86	0,95	0,92
Extérieur 2	0,95	0,81	0,91	0,81

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 permet de filtrer des données aberrantes issues de la phase de segmentation 2D. Ces informations calculées sur l'image 14(c) sont données tableau 3.

Tableau 3. Altitudes et tailles estimées des objets détectés figure 14(c). La colonne Association fait référence aux numéros des silhouettes dans la figure 14(c), chaque association correspond donc à un objet

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

La figure 15 représente les histogrammes des taux de détection des objets stationnaires des séquences *extérieur 1* et *extérieur 2*. Le taux de détection d'un objet est défini comme le nombre d'images où il a été détecté divisé par le nombre d'images où il était effectivement présent. On peut constater que sur les deux séquences, trois objets n'ont pas du tout été détectés par notre système. Cependant ceci n'est pas le résultat d'une erreur d'appariement ou de détection, mais vient du fait que ces objets ne sont visibles que dans une caméra. Ces histogrammes montrent donc que si un objet est visible dans les deux caméras, alors il est détecté par notre système. Ces expériences ont été menées sur un PC quadricore de fréquence 2,4 GHz et équipé de 3 GB de RAM. L'étape de soustraction de fond prend 50 ms sur des images VGA, alors que les étapes d'étiquetage d'objets et d'appariement prennent chacune au maximum 10 ms.



(a) Séquence *extérieur 1* (b) Séquence *extérieur 2*

Figure 15. Histogramme des taux de détection des objets à détecter

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 l'étiquetage 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. L'évaluation sur des scènes complexes issues de bases publiques a montré la robustesse de notre approche. La seconde contribution porte sur une approche géométrique qui assure 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. Ce travail constitue une première étape dans un système de détection d'objets abandonnés. Afin de pouvoir l'utiliser dans des conditions diverses, il sera intéressant de généraliser l'étape d'appariement à des systèmes comprenant plus de deux caméras.

Bibliographie

Alahari K., Kohli P., Torr P. H. S. (2008). Reduce, reuse & recycle: Efficiently solving multi-label MRFs. In *Proceedings of ieee conference on computer vision and pattern recognition.*

Bay H., Ess A., Tuytelaars T., Gool L. V. (2008). Surf: Speeded up robust features. In *Cviu.*

Bayona A., SanMiguel J., Martinez J. (2009). Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques. In *Advanced video and signal based surveillance.*

Bayona A., SanMiguel J., Martinez J. (2010). Stationary foreground detection using background subtraction and temporal difference in video surveillance. In *International conference on image processing.*

Beynon M. D., Van Hook D. J., Seibert M., Peacock A., Dudgeon D. (2003). Detecting abandoned packages in a multi-camera video surveillance system. In *Conference on advanced video and signal based surveillance.*

Bhargava M., Chen C.-C., Ryoo M., Aggarwal J. (2007, sept.). Detection of abandoned objects in crowded environments. In *Advanced video and signal based surveillance, 2007. avss 2007. ieee conference on*, p. 271 -276.

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

Cipolla R., Astrom K., Giblin P. (1995). Motion from the frontier of curved surfaces. In *Fifth international conference on computer vision.*

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

Guillot C., Sayd P., Pham Q.-C., Tilmant C., Lavest J.-M. (2012). Détection et localisation d?objets stationnaires par une paire de caméras ptz. In *Reconnaissance des formes et intelligence artificielle.*

Guillot C., Taron M., Sayd P., Pham Q.-C., Tilmant C., Lavest J.-M. (2010). Background subtraction for ptz cameras performing a guard tour and application to cameras with very low frame rate. In *Accv workshop on visual surveillance.*

Guler S., Silverstein J., Pushee I. (2007). Stationary objects in multiple object tracking. In *Advanced video and signal based surveillance.*

i-lids dataset for avss 2007. (s. d.). http://www.eecs.qmul.ac.uk/˜ andrea/avss2007_d.html.

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

Kohli P., Torr P. (2005). Efficiently solving dynamic markov random fields using graph cuts. In *Iccv*, p. II: 922-929.

Kolmogorov V. (2006, October). Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, p. 1568–1583. http://dx.doi.org/10.1109/TPAMI.2006.200

Liao H.-H., Chang J.-Y., Chen L.-G. (2008). A localized approach to abandoned luggage detection with foreground-mask sampling. In *Advanced video and signal based surveillance.*

Lv F., Song X., Wu B., Kumar V., Nevatia S. R. (2006). Left luggage detection using bayesian inference. In *Pets.*

Mathew R., Yu Z., Zhang J. (2005). Detecting new stable objects in surveillance video. In *Multimedia signal processing, 2005 ieee 7th workshop on.*

Miezianko R., Pokrajac D. (2008). Localization of detected objects in multi-camera network. In *Icip.*

Porikli F., Ivanov Y., Haga T. (2008, January). Robust abandoned object detection using dual foregrounds. *EURASIP J. Adv. Signal Process*, vol. 2008. http://dx.doi.org/10.1155/2008/197875

San Miguel J., Martinez J. (2008). Robust unattended and stolen object detection by fusing simple algorithms. In *Advanced video and signal based surveillance, 2008. avss '08. ieee fifth international conference on*, p. 18 -25.

Utasi A., Csaba B. (2010). Multi-camera people localization and height estimation using multiple birth and death dynamics. In *Accv workshop on visual surveillance.*

Valentine B., Apewokin S., Wills L., Wills S., Gentile A. (2007). Midground object detection in real world video scenes. In *Conference on advanced video and signal based surveillance.*

Venetianer P., Zhang Z., Yin W., Lipton A. (2007, sept.). Stationary target detection using the objectvideo surveillance system. In *Advanced video and signal based surveillance, 2007. avss 2007. ieee conference on*, p. 242 -247.

Constant Guillot est diplômé en 2008 de l'université de Manchester et de l'ENSIIE. Il a reçu son doctorat de l'université de Clermont-Ferrand en 2012 pour ses travaux sur la détection d'objets stationnaires réalisés au Laboratoire Vision et Ingénierie des Contenus du CEA LIST. Il est désormais ingénieur de recherche à R&D Vision.

Patrick Sayd a rejoint le CEA après son doctorat en Vision pour la Robotique préparé au Lasmea de l'Université Blaise Pascal de Clermont-Ferrand en 1996. Après avoir contribué aux problématiques de reconstruction d'environnement et de localisation 3D par vision, il s'est intéressé à l'analyse de scène dans les domaines de la vidéo-protection et les systèmes d'aide à la conduite automobile. Il est aujourd'hui responsable du Laboratoire Vision et Ingénierie des Contenus du CEA LIST.

Quoc Cuong Pham est chercheur en vision par ordinateur au CEA LIST. Il obtient le grade de docteur de l'Institut National Polytechnique de Grenoble en 2002 en imagerie médicale pour ses travaux sur les modèles déformables dynamiques. Il rejoint le CEA LIST en 2003. Ses activités de recherche au sein du LIST vont de la localisation 3-D à la détection, la classification et au suivi d'objets temps réel dans les flux vidéo, en particulier dans le cadre des applications de vidéosurveillance intelligente.

Christophe Tilmant est ingénieur ENSEEIHT (2001) et Docteur de l'université d'Auvergne (2004). Il rejoint en 2005, comme Maître de Conférences, le Laboratoire Institut Pascal de l'université Blaise Pascal. Ses activités de recherches portent sur la segmentation et le suivi dans des applications en vidéo-surveillance et en imagerie médicale.

Jean-Marc Lavest est Professeur à l'Université d'Auvergne. Chercheur à l'Institut Pascal UMR CNRS, son domaine d'expertise se focalise sur la métrologie par vision. Expert auprès d'Oséo, il a participé à la création de deux start-up Poseidon et Dxo-labs en application directe de ses travaux de recherche. Directeur de l'IUT de Clermont depuis 2009, membre du réseaux de APM, il milite pour le développement de l'entrepreneuriat et la création d'entreprises dans le secteur de la technologie.