Suivi par ré-identification dans un réseau de caméras à champs disjoints

Boris Meden [1], **Frédéric Lerasle** [2], **Patrick Sayd** [1]

1. CEA, LIST
Laboratoire Vision et Ingénierie des Contenus
BP 94, F-91191 Gif-sur-Yvette
{boris.meden,patrick.sayd}@cea.fr

2. CNRS ; LAAS
Université de Toulouse ; UPS, LAAS
F-31077 Toulouse cedex 4
lerasle@laas.fr

RÉSUMÉ. *Cet article pose le problème du suivi automatique de piétons à travers les réseaux de caméras à champs de vue disjoints. Le suivi dans l'image est traité de manière locale par un algorithme de suivi par détections et ré-identification. Avec du filtrage particulaire à état continu et discret, nous introduisons la notion d'identité globale dans un algorithme de suivi multipiste pour caractériser les personnes au niveau du réseau et pallier les discontinuités d'observations. Ceci permet à chaque traqueur d'inclure l'identité de la cible qu'il est en train de suivre dans l'espace de recherche. Ce faisant, chaque traqueur maintient à jour une distribution de probabilité discrète sur l'identité de la piste qu'il est en train de suivre. La décision de ré-identification est renforcée par un schéma décisionnel haut niveau intégrant les hypothèses de chaque traqueur confrontées à la topologie du réseau. La composante suivi multipersonne et ré-identification est d'abord testée en contexte monocaméra. Nous évaluons ensuite notre approche complète sur un réseau de 3 caméras à champs de vue disjoints et un ensemble de 7 personnes. La seule connaissance a priori requise est la carte topologique du réseau.*

ABSTRACT. *This article tackles the problem of automatic multi-pedestrian tracking in non overlapping fields of view camera networks, using monocular, uncalibrated cameras. Tracking is locally addressed by a Tracking-by-Detection and reidentification algorithm. We propose here to introduce the concept of global identity into a multi-target tracking algorithm, qualifying people at the network level, to allow us to rebound observation discontinuities. We embed that identity into the tracking loop thanks to the mixed-state particle filter framework, thus including it in the search space. Doing so, each tracker maintains a mutli-modality on the identity in the network of its target. We increase the decision strength introducing a high level decision scheme which integrates all the trackers hypothesis over all the cameras of the network with previous reidentification results and the topology of the network. The tracking and reidentification module is first tested with a single camera. We then evaluate the whole framework on a 3*

Traitement du signal – n° 3-4-5/2012, 283-305

non-overlapping fields of views network with 7 identities. The only a priori knowledge assumed is a topological map of the network.

MOTS-CLÉS : *ré-identification, suivi de personnes, réseau de cameras, champs de vue disjoints, filtrage particulaire.*

KEYWORDS: *re-identification, pedestrian tracking, camera network, nonoverlapping fields of view, particle filtering.*

DOI:10.3166/TS.29.283-305

Extended abstract

Summary

This article tackles the problem of automatic multi-pedestrian tracking in large scale environnments. Material and economical reasons generally limit the number of cameras thus yielding discontinuities/blind spots in the network field of view. We use monocular, uncalibrated cameras.

Tracking is locally addressed by a Tracking-by-Detection and reidentification algorithm. We propose here to introduce the concept of global identity into a multi-target tracking algorithm. That way, we describe people at the network level, to allow us to be robust to observation discontinuities. We embed that identity into the tracking loop thanks to the mixed-state particle filter framework. The state vector recursively estimated is composed of real terms as well as an integer term (hence the mixed-state qualification) which represents the identity of the target out of a pool of identities. Doing so, we include the identity inference -or reidentification- in the filtering process, searching in a mixed space. That way, each tracker maintains a multi-modality on the identity in the network of its target.

We increase the decision strength introducing a high level decision scheme which integrates all the trackers hypotheses over all the cameras of the network with previous reidentification results and the topology of the network. The tracking and reidentification module is first tested with a single camera. We then evaluate the whole framework on a 3 non-overlapping fields of views network with 7 persons. The only *a priori* knowledge assumed is a topological map of the network.

Introduction

This reidentification problem is classically treated as a request in a database, inspired from web technologies, and puts the focus on the pedestrian appearance description to re-identify. Thus, (Gray, Tao, 2008) propose to train a classifier on the invariant parts during a camera change. (Farenzena *et al.*, 2010) adopt the same approach whithout any learning, proposing a robust fixed signature based on symmetry and asymmetry of the appearance and well positionned colorimetric features. These

methods are costly in terms of computation time and are well suited to *a posteriori* treatments.

For a camera network application, the reidentification module should allow real time processing. Here we target an online update of targets' *global identities*. A similar problem has been tackled by (Chen *et al.*, 2008 ; Lev-Tov, Moses, 2010 ; Zajdel, Kröse, 2005). However (Zajdel, Kröse, 2005) suppose to have single pedestrians passing in the network, (Chen *et al.*, 2008) do not report on their tracking process and (Lev-Tov, Moses, 2010) just simulate a NOFOV (for « Non Overlapping Field of View ») network and do not work on images. These works do not consider tracking and reidentification jointly, and thus occult the difficulties of multi-target tracking.

Mono-camera multi-target tracking is a largely tackled problem in the Computer Vision community: our approach is based on different associated assessments. First, particle filtering algorithms' interest for tracking (CONDENSATION) has been established since the initial work of Isard and Blake in (Isard, Blake, 2001), notably for multiple targets. Then, since (Okuma *et al.*, 2004), *tracking-by-detection* has emerged and particularly the temporal integration of *tracklets*, which robustness has been proven by Kaucic *et al.* in (Kaucic *et al.*, 2005). *Tracklets* optimisation has also been extended to two cameras presenting a disjoint field of view by (Kuo *et al.*, 2010). This method yet does not work online, as the optimisation is conducted on a temporal window.

In opposition to them, our approach places itself in the markovian formalism for the tracking module. Our approach is inspired by (Breitenstein *et al.*, 2010) and (Wojek *et al.*, 2010). Like (Breitenstein *et al.*, 2010), it is based on distributed particle filters enhanced by a reidentification component coming from a discrete identity variable also sampled. They are termed mixed-state particle filters. Then, in the vein of (Wojek *et al.*, 2010), we perform a *tracklet* temporal integration, but on the identities here, and not for cameras but on the whole network.

Tracking-by-Reidentification

In this article, we propose an extension to NOFOV networks of the tracking-by-detection algorithm proposed by Breitenstein *et al.* in (Breitenstein *et al.*, 2010), introducing the notion of *global identity* that we seek to retrieve for each target. We present first our implementation of (Breitenstein *et al.*, 2010) and how the use of mixed-state particle filtering for reidentification (Meden *et al.*, 2011) comes to extend that approach.

Topological Supervision

The previous reidentification strategy is integrated to the image plane tracking. That strategy has been established as superior to an exhaustive comparison to the database by (Meden *et al.*, 2011). Its limitation resides in the distributed aspect of the mixed-state filters. Indeed, the probability densities over the target identity are independent from one filter to another. Thus, two filters may produce the same identity

at the same time for two different targets. We wish here to constrain the process, so that it produces exclusive trackers/identities pairing. This is done at the network level.

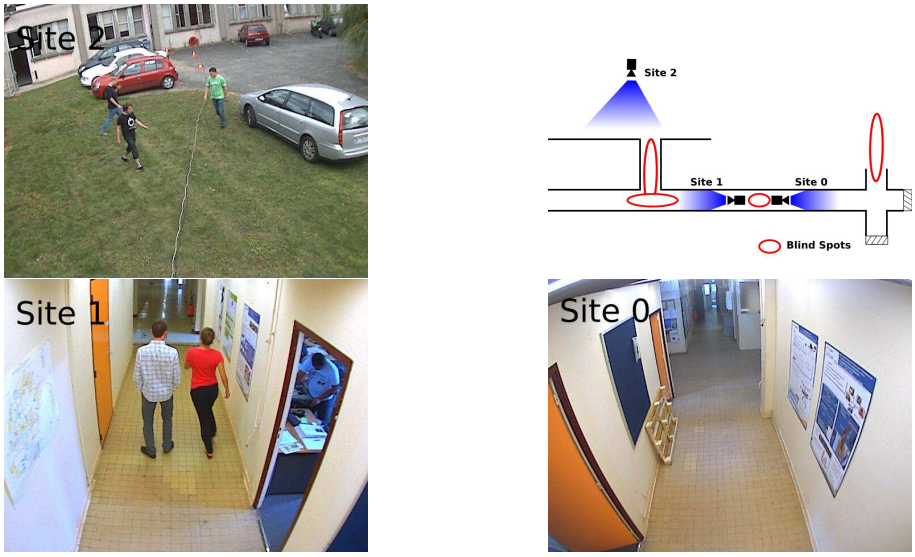


Figure 1. Exemple de réseau de caméras à champs disjoints

1. Introduction

Les travaux présentés dans cet article ont pour objectif le suivi de personnes en environnements intérieurs à large échelle. Les contraintes matérielles/économiques limitent en général le nombre de caméras et empêchent une couverture totale de l'espace, ce qui engendre des discontinuités dans le champ de vue du réseau. On parle de réseaux à champs de vue disjoints comme illustré en figure 1. Nous proposons ici une version étendue des travaux présentés dans (Meden *et al.*, 2012).

L'enjeu pour le processus de suivi est alors de gérer ces discontinuités dans le champ de vue du réseau pour assurer la cohérence spatiotemporelle. Outre le suivi des personnes dans chaque image, le système doit ré-identifier les personnes lors de leur apparition dans les différents lieux et donc dans les champs de vue des caméras associées.

Nous proposons ici d'intégrer le formalisme de filtrage particulaire à état mixte pour la ré-identification (Meden *et al.*, 2011) dans un algorithme de suivi-par-détection multipiste (Breitenstein *et al.*, 2010). Ceci permet une stratégie de ré-identification en ligne, intégrée au suivi, se basant sur une description colorimétrique des identités. La seconde contribution de cet article réside dans l'ajout d'un superviseur intégrant les résultats d'identification des traqueurs à la manière de *tracklet*, tout en les confrontant à la topologie du réseau. La figure 2 schématise l'ensemble de l'architecture

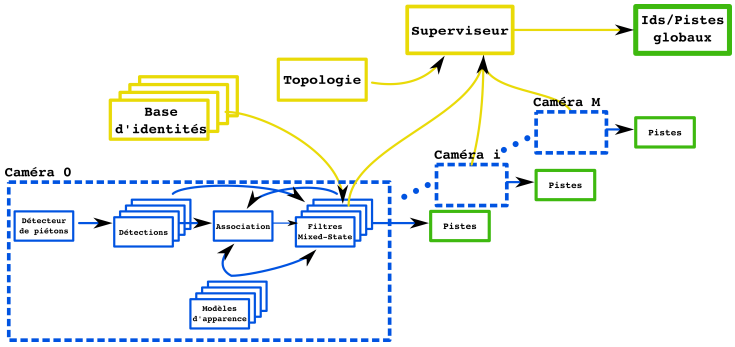


Figure 2. Architecture du système perceptuel. Les traitements « suivi et ré-identification dans l'image » sont localisés au niveau des caméras, alors que le superviseur a la connaissance de tout le réseau et confronte les distributions d'identités à sa topologie

proposée, avec les traitements locaux aux caméras, et la centralisation des résultats de ré-identification.

Nous présentons en section 2 un état de l'art des travaux du domaine. Puis nous détaillons le traqueur par ré-identification inhérent à chaque caméra dans la section 3. La supervision de ces ré-identifications de pistes est ensuite détaillée en section 4. Finalement, la section 5 décrit les évaluations quantitatives de la fonction de base de ré-identification, ainsi que l'apport de contraintes topologiques lorsqu'appliquées à un réseau.

2. État de l'art

Le suivi d'objets multiples (MOT pour *Multiple Object Tracking*) a été largement traité par les communautés Radar et Traitement du Signal, avec les travaux originels de (Reid, 1979) sur le *Multi-Hypothesis Tracker* (MHT) et ceux de (Bar-Shalom *et al.*, 1980) sur le filtre *Joint Probability Data Association* (JPDA). L'objectif est de filtrer un ensemble de détections en environnement encombré pour produire un suivi spatiotemporel des cibles ayant généré ces détections. Les travaux suivants ont cherché à limiter l'explosion combinatoire inhérente à ces algorithmes. En effet, ils explorent l'ensemble des assignations les plus probables avant de produire leur réponse. À ce titre, (Cox, Hingorani, 1996) ont proposé une implémentation plus efficace du MHT, à nouveau reformulée par (Danchick, Newnam, 2006). Suivant les idées de (Pasula *et al.*, 1999) sur les réseaux de capteurs, (Oh *et al.*, 2004) proposent d'explorer la combinatoire d'association à l'aide de *Monte Carlo Markov Chain Data Association* et ont proposé un algorithme MCMCDA et prouvé que la distribution converge vers l'optimum donné par les méthodes JPDA, offrant ainsi un schéma d'approximation aléatoire en temps polynomial. Cependant, les travaux pré-cités privilégient une logique

différée, *i.e.* travaillent sur une fenêtre temporelle et produisent un résultat après avoir accumulé les observations de cette fenêtre.

Le suivi d'objets multiples est aussi devenu un sujet de recherche majeur dans la communauté Vision par Ordinateur. Cependant les travaux initiaux relèvent de la logique séquentielle avec des modèles markoviens d'ordre 1. L'intérêt de ces algorithmes, *e.g.* de filtrage particulaire (CONDENSATION), a été établi par les travaux initiaux de (Isard, Blake, 1998a), et en particulier pour des cibles multiples dans (Isard, Blake, 2001). Ensuite avec (Okuma *et al.*, 2004), le *suivi-par-détection* est apparu. Avec l'augmentation de la puissance de calcul, la logique différée et l'intégration temporelle de *tracklet*, dont la robustesse a été prouvée par (Kaucic *et al.*, 2005), ont commencé à apparaître. (Xing *et al.*, 2009) ont prouvé dans l'intérêt de ces méthodes par rapport à des méthodes markoviennes pour une meilleure gestion des occultations de cibles. La décision est produite sur un horizon plus grand, mais l'un des coûts induits est la latence dans l'obtention du résultat, dépendante de la durée de la fenêtre temporelle.

Parmi les méthodes classiques de l'état de l'art en termes de suivi multipiste monocaméra, (Benfold, Reid, 2011) proposent une intégration temporelle de *tracklet* basée sur des détections HOG (Dalal, Triggs, 2005). L'implémentation GPU du détecteur (Prisacariu, Reid, 2009) rend la méthode temps réel, la logique différée imposant seulement une latence de quelques secondes dans la production de résultats. Par opposition, (Breitenstein *et al.*, 2010) évitent cette latence en ayant recours à une méthode markovienne. Le sujet du suivi et de l'identification conjoints dans des réseaux de caméras à champs de vue joints *e.g.* (Qu *et al.*, 2007) est très similaire : la mise en relation des différents flux vidéo des capteurs vient classiquement avec l'utilisation de la calibration du système, permettant de travailler dans un repère 3D commun. Ainsi l'identification des cibles d'une caméra aux autres se base uniquement sur un critère géométrique.

Lorsque les caméras du réseau ne partagent pas leurs champs de vue, la calibration est plus difficile à obtenir, et la discontinuité d'observation empêche tout suivi spatiotemporel classique. Ce problème prend le nom de ré-identification de personnes. Pour y répondre, des travaux récents tels que ceux de (Farenzena *et al.*, 2010 ; Gheissari *et al.*, 2006 ; Gray, Tao, 2008 ; Prosser *et al.*, 2010 ; Zheng *et al.*, 2011) ont cherché à construire un modèle de représentation d'identités, robuste par changement de caméra pour produire une fonction de similarité fiable. Le problème de la ré-identification a commencé à être traité pour lui-même, et des bases telles que VIPeR [1] (Gray *et al.*, 2007), ILIDS [2] (Zheng *et al.*, 2009), ETHZ [3] (Ess *et al.*, 2007) et CAVIAR [4] (Cheng *et al.*, 2011) sont apparues. Dans cette veine, (Gray, Tao, 2008) et (Gheissari *et al.*, 2006) on été les premiers à le formuler comme un problème de classement

1. http://vision.soe.ucsc.edu/node/178
2. http://www.ilids.co.uk
3. http://www.umiacs.umd.edu/~schwartz/datasets.html
4. http://www.lorisbazzani.info/code-datasets/caviar4reid

(ou *ranking* en anglais) et à l'évaluer avec des courbes d'association cumulées (CMC pour *Cumulative Matching Curves*). Pour chaque silhouette de la première caméra, la similarité à toutes les silhouettes de la seconde caméra est calculée, puis les réponses sont ordonnées et ainsi on obtient le rang de réponse de la silhouette vérité terrain. (Prosser *et al.*, 2010) ont même intégré cette notion de classement en utilisant une version dédiée du RankSVM. Ici, la ré-identification est traitée comme une requête dans une base de donnée, inspirée des technologies web. Ainsi (Gray, Tao, 2008) proposent d'entraîner un classifieur sur les éléments invariants par changement de caméras, en se basant sur un ensemble de paires de silhouettes vérité terrain. Le modèle appris met en avant les caractéristiques invariantes par changement de caméra. Par ailleurs (Zheng *et al.*, 2011) mettent l'accent sur la distance utilisée. Ils proposent une distance probabiliste entraînée pour minimiser la distance entre des paires de silhouettes appartenant à la même identité. Encore une fois, l'entraînement est réalisé sur des paires vérité terrain, et suppose donc avoir résolu le problème de la ré-identification pour construire cet ensemble d'entraînement labellisé. Par opposition à ces approches, (Farenzena *et al.*, 2010) adoptent une méthode non supervisée, ne nécessitant aucun apprentissage. Ils proposent d'utiliser les axes de symétrie et asymétrie de la silhouette pour pondérer les caractéristiques colorimétriques, une mise en correspondance de blobs de couleur moyenne extraits par *Maximal Stable Color Region* (Forssén, 2007), ainsi que des patches de texture locaux. Cependant, les méthodes listées ci-avant sont difficilement transférables à des applications en ligne du fait de leur important temps de calcul.

Dans un réseau de caméras, lorsque les trajectoires de cibles présentent des discontinuités dues à leur non-observabilité, *e.g.* entre différentes caméras non recouvrantes, l'application de la ré-identification devient une nécessité. Les travaux de (Huang, Russell, 1997) se situent parmi les premières formalisations de ce problème. Ils ont formalisé une ré-identification de véhicules en réalisant des associations dans un espace probabilisé construit sur des cibles de tailles et couleurs moyennes similaires. Ensuite, (Pasula *et al.*, 1999) ont prouvé la convergence de l'échantillonnage MCMC pour explorer la combinatoire d'association, tandis que (Kettnaker, Zabih, 1999) ont décomposé le problème multicaméra en un programme linéaire. (Zajdel, Kröse, 2005) ont adopté un formalisme proche de celui de Pasula *et al.* mais l'ont résolu à l'aide de réseaux de Bayes dynamiques (DBN pour *Dynamic Bayes Network*) pour évaluer la vraisemblance des hypothèses et un algorithme EM pour apprendre les paramètres du modèle. Des connaissances supplémentaires sur le réseau de caméras peuvent être apprises, telles que des correspondances spatiotemporelles entre zones d'entrées/sorties (Makris *et al.*, 2004) ou des fonctions de transfert de luminance reliant les apparences d'une personne dans plusieurs caméras (Javed *et al.*, 2005). Pour ce faire, (Makris *et al.*, 2004) ont investigué l'apprentissage non supervisé d'un modèle d'activité à partir d'un large ensemble d'observations, sans correspondance vérité terrain. (Javed *et al.*, 2005) ont montré que la fonction de transfert de luminance (BTF pour *Brightness Transfer Function*) d'une caméra vers une autre existe dans un sous-espace de petite dimension et ont montré comment l'approximer simplement par analyse en composantes principales probabiliste. De telles connaissances supplémentaires aident à la ré-identification en contraignant les associations possibles.

Quelques travaux récents commencent à transférer les descripteurs de ré-identification à des applications nécessitant une identification en ligne. (Oreifej *et al.*, 2010) abordent le MOT dans des images prises depuis un drone. Les mouvements de la caméra génèrent des changements de pose assimilables aux différentes poses de caméras d'un réseau NOFOV (pour *Non Overlapping Field Of View*), et empêchent tout suivi spatiotemporel. Dans leur contexte applicatif, ils réalisent le suivi multipiste en ré-identifiant les détections 2D obtenues dans les images. (Kuo, Nevatia, 2011) ont montré que, sous certaines conditions, les concepts de ré-identification peuvent même améliorer les performances d'un suivi multicible sur une caméra statique. En termes de réseaux NOFOV, à notre connaissance, (Matei *et al.*, 2011) et (Kuo *et al.*, 2010) sont les seuls à essayer d'unifier ré-identification et MOT dans un système complet. Matei *et al.* se positionnent sur le suivi de véhicules, ce qui induit un modèle de mouvement linéaire et la supposition d'une vitesse constante. Ces contraintes leur permettent de formaliser un MHT utilisant cinématique et apparence. Au contraire, Kuo *et al.* traitent le cas de piétons. Ils adoptent une ré-identification similaire à celle de (Gray, Tao, 2008) en entraînant un boosting sur des paires de silhouettes correctement appariées entre deux caméras. Le système proposé travaille en deux phases : tout d'abord, des contraintes spatiotemporelles d'apparition dans les caméras permettent de générer des appariements de silhouettes. Ensuite, un algorithme de type MIL-boost (pour *Multiple Instance Learning-boost*) apprend les caractéristiques invariantes dans le changement de caméra tout en tolérant du bruit dans la labellisation, *i.e.* dans les appariements de silhouettes donnés comme exemples positifs. Le suivi est réalisé par intégration temporelle de *tracklet*. A la fin de la séquence, les classifieurs de ré-identification sont utilisés avec l'algorithme Hongrois (Burgeois, Lasalle, 1971) pour apparier entre les caméras. La méthode est automatique, bien que la phase de labellisation repose sur des contraintes très faibles en contexte difficile. Par ailleurs, la ré-identification est vue comme un problème de maximum a posteriori (MAP), et résolu en énumérant toutes les possibilités d'associations. L'explosion combinatoire d'association est un frein à l'extensibilité d'une telle méthode.

Par opposition, notre approche se place en contexte markovien au niveau du module de suivi. Notre approche s'inspire de (Breitenstein *et al.*, 2010) et de (Wojek *et al.*, 2010). A l'instar de (Breitenstein *et al.*, 2010), elle repose sur des filtres particulaires distribués mais elle ajoute la composante de ré-identification *via* une variable discrète relative à l'identifiant de la cible. On parle alors de filtrage particulaire à état mixte. A l'instar de (Wojek *et al.*, 2010), une intégration temporelle (*tracklet*) sur les identités de pistes est mise en œuvre mais non à l'échelle d'une ou plusieurs caméras à champs joints mais au niveau du réseau de par notre problématique.

3. Suivi par ré-identification au sein d'une caméra

Dans cet article, nous proposons une extension aux réseaux à champs disjoints de l'algorithme de suivi-par-détection proposé par (Breitenstein *et al.*, 2010), en introduisant la notion d'*identité globale*. Nous présentons dans cette section notre implé-

mentation de (Breitenstein *et al.*, 2010) et comment l'utilisation du filtrage particulaire à état mixte pour la ré-identification (Meden *et al.*, 2011) vient étendre cette approche.

3.1. Description des cibles

3.1.1. Apprentissage des identités du réseau

Figure 3. Images-clés relatives à chaque ID (caméra 1)

Tout algorithme de ré-identification nécessite d'avoir vu une personne au préalable pour être capable de la ré-identifier. Nous supposons ici la phase de constitution d'une telle base acquise. Pour cela, nous extrayons une collection d'images-clés d'une des caméras du réseau (*e.g.* positionnée dans le hall d'entrée du bâtiment à surveiller), et utilisons celles-ci comme descriptions de nos identités. Le choix des images-clés est réalisé par K-means sur des séquences de suivi dans la caméra choisie à la manière de (Meden *et al.*, 2011). Ainsi, ces images-clés encodent la variabilité d'apparence obtenue pour cette identité au cours de son suivi initial. La figure 3 présente un exemple de base d'identités utilisé pour traiter le réseau de la figure 1, apprises ici dans la caméra 1.

3.1.2. Modélisation de l'apparence d'une piste

Nous utilisons le même modèle d'apparence que (Meden *et al.*, 2011) pour décrire les pistes de tracking ainsi que les identités de la base : des bandes horizontales de distributions couleur calculées dans l'espace RGB. La mesure de similarité entre deux descripteurs est la somme des distances de Bhattacharyya, évaluée sous un noyau gaussien. Ceci nous permet de calculer les similarités par rapport au modèle d'apparence d'un traqueur ainsi que par rapport à une identité de la base, notées respectivement $w_{App}(.)$ et $w_{Id}(.)$.

3.2. Gestion des détections

3.2.1. Association aux détections

Notre approche privilégie une stratégie « tracking-by-detection » *via* le détecteur classique proposé dans (Dalal, Triggs, 2005). Ces détections sont intégrées dans le processus de suivi par une étape d'association préalable de type algorithme glouton. À la fin de cette étape, chaque traqueur est potentiellement associé à une détection qui va servir à la mise à jour de ses particules. Pour ce faire, nous construisons une

matrice d'association entre les détections (lignes) et les traqueurs (colonnes). Le score de chaque paire détection d *versus* traqueur tr, donné par l'équation (1), fait intervenir :

– la distance des particules du traqueur à la détection évaluées sous une loi normale $p_{\mathcal{N}}(.) \sim \mathcal{N}(., \sigma^2)$;

– l'aire de la boîte du traqueur $\mathcal{A}(tr)$ relativement à celle de la détection aussi évaluée sous une loi normale ;

– l'évaluation du modèle d'apparence du traqueur en la détection ($w_{App}(.)$).

$$S(d,tr) = \underbrace{\sum_{p \in tr}^{N} p_{\mathcal{N}}(d-p)}_{\text{distance euclidienne}} \times \underbrace{p_{\mathcal{N}}\left(\frac{|\mathcal{A}(tr) - \mathcal{A}(d)|}{\mathcal{A}(tr)}\right)}_{\text{taille relative}} \times \underbrace{w_{App}(d,tr)}_{\text{modèle d'apparence}} \quad (1)$$

Ainsi, le traqueur et la détection doivent présenter simultanément une cohérence en termes de position, de taille et de contenu colorimétrique. Une fois cette matrice de similarité construite, on extrait itérativement les maxima, avec suppression de leurs lignes et colonnes. On itère tant que les maxima sont supérieurs au seuil d'appariement (fixé empiriquement à 0.002). Une telle heuristique est préférée à la solution optimale fournie par l'algorithme Hongrois (Burgeois, Lasalle, 1971), écarté pour sa complexité.

3.2.2. Initialisations/terminaisons automatiques de traqueurs

Toute détection récurrente temporellement donne lieu à l'instanciation d'un nouveau traqueur. Par ailleurs, tout traqueur n'ayant pas de détection associée sur un intervalle de temps supérieur au seuil de suppression (4 images en pratique) se voit arrêté.

3.3. Filtrage particulaire

3.3.1. Modèle de prédiction à état mixte

Chaque piste initialisée par une détection est suivie par un filtre à particules. Étant donné la base d'identités, nous avons des descripteurs de référence supplémentaires auxquels se comparer. Pour cela, à l'instar de (Meden *et al.*, 2011), nous utilisons des filtres de type Mixed-State CONDENSATION, introduits dans (Isard, Blake, 1998b). Nous cherchons à estimer un vecteur d'état mixte, ajoutant un terme discret aux paramètres continus, soit

$$\mathbf{X} = (\mathbf{x}, id)^{\mathsf{T}},\ \mathbf{x} \in \mathbb{R}^4,\ id \in \{1, \ldots, N_{id}\}$$

La partie continue de l'état $\mathbf{x} = [x, y, v_x, v_y]^{\mathsf{T}}$ se compose de la position dans le plan image $(x, y)^{\mathsf{T}}$ et du vecteur vitesse $(v_x, v_y)^{\mathsf{T}}$. La partie entière id renvoie à l'une des N_{id} identités de la base. Le suivi se passe dans le plan image, et la dimension des boîtes de suivi est fixée et mise à jour sur les détections associées à ces traqueurs. Le

modèle d'apparence est lui aussi mis à jour sur la détection associée. Étant donné ce vecteur d'état étendu, la densité du processus d'échantillonnage à l'image t peut être décomposée comme présenté dans (Isard, Blake, 1998b) :

$$\begin{aligned} p(\mathbf{X}_t|\mathbf{X}_{t-1}) &= p(\mathbf{x}_t|id_t, \mathbf{X}_{t-1}) \cdot P(id_t|\mathbf{X}_{t-1}) \\ P(id_t|\mathbf{X}_{t-1}) : &\quad P(id_t = j|\mathbf{x}_{t-1}, id_{t-1} = i) = T_{ij}(\mathbf{x}_{t-1}) \\ p(\mathbf{x}_t|id_t, \mathbf{X}_{t-1}) : &\quad p(\mathbf{x}_t|\mathbf{x}_{t-1}, id_{t-1} = i, id_t = j) = p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1}) \end{aligned}$$

où $T_{ij}(\mathbf{x}_{t-1})$ est la probabilité de transition de l'identité i vers j, appliquée au paramètre discret d'identité, et $p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1})$ est l'échantillonnage de la loi appliquée à la partie continue de l'état. La matrice de transition $T = [T_{ij}]$ est construite sur l'ensemble des images-clés. L'élément T_{ij} est la similarité $w_{id}(.)$ entre les identités i et j de la base, calculée entre les images-clés les plus dissemblables. Les particules sont propagées selon un modèle de mouvement d'ordre 1 :

$$p_{ij}(\mathbf{x}_t|\mathbf{x}_{t-1}) : \quad \begin{cases} (x,y)_t = (x,y)_{t-1} + (v_x, v_y)_{t-1} \cdot \Delta t + \epsilon_{(x,y)} \\ (v_x, v_y)_t = (v_x, v_y)_{t-1} + \epsilon_{(v_x,v_y)} \end{cases}$$

où les bruits $\epsilon_{(x,y)}$ et $\epsilon_{(x,y)}$ suivent des lois normales et où Δt est l'intervalle de temps séparant deux images.

3.3.2. Modèle d'observation intégrant les détections

Le poids $w_{tr}^{(p)}$ attribué à la p^{e} particule du traqueur tr est calculé en intégrant la distance de la particule à la détection d^* qui lui a été associée, la similarité colorimétrique au modèle d'apparence du traqueur $w_{App}(.)$ et la similarité colorimétrique à l'identité de la particule $w_{Id}(.)$. $Id(p)$ représente l'identité choisie par p. Il s'agit du terme mixte.

$$w_{tr}^{(p)} = \underbrace{\alpha \cdot \mathcal{I}(tr) \cdot p_{\mathcal{N}}(d^* - p)}_{\text{distance à la détection}} + \underbrace{\beta \cdot w_{App}(d, tr)}_{\text{modèle d'apparence}} + \underbrace{\gamma \cdot w_{Id}(d, id(p))}_{\text{identité}} \tag{2}$$

où α, β et γ sont des coefficients dont la somme est égale à 1, et $\mathcal{I}(tr)$ un booléen signifiant l'existence ou non d'une détection associée au traqueur. Comme dans (Meden *et al.*, 2011), l'introduction d'une similarité relative à l'identité dans la pondération de la particule permet de diriger le nuage de particules vers les identités les plus probables aux vues des observations reçues. En ce sens, chaque traqueur maintient une multimodalité sur les *identités globales* les plus vraisemblables pour la personne qu'il suit.

L'estimation de l'état est un processus en deux étapes. Nous commençons par calculer le MAP sur le paramètre discret par rapport à l'observation courante $\mathbf{Z}_t$ avec l'équation (3), *i.e.* l'identité la plus probable à l'instant t.

$$\hat{id}_t = \arg\max_j P(id_t = j|\mathbf{Z}_t) = \arg\max_j \sum_{p \in \Upsilon_j} w_{tr}^{(p)}(t), \tag{3}$$

$$\text{où } \Upsilon_j = \left\{ p | \mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, j) \right\}$$

Ensuite, les composantes continues sont estimées sur le sous-ensemble de particules $\hat{\Upsilon}$ qui possèdent l'identité la plus vraisemblable, selon l'équation (4).

$$\hat{\mathbf{x}}_t = \sum_{p \in \hat{\Upsilon}} w_{tr}^{(p)}(t) \cdot \mathbf{x}_t^{(p)} / \sum_{p \in \hat{\Upsilon}} w_{tr}^{(p)}(t), \quad (4)$$

$$\text{où } \hat{\Upsilon} = \{p | \mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, \hat{id}_t)^\mathsf{T}\}$$

De cette manière, en plus de l'estimation de la position image de la cible, chaque filtre fournit une distribution d'identités pour cette cible.

4. Supervision topologico-temporelle des ré-identifications

La section 3 a présenté une stratégie de ré-identification intégrée au suivi. Cette stratégie a été établie par (Meden *et al.*, 2011) comme supérieure à une comparaison exhaustive à la base. Sa limitation réside dans le caractère distribué des filtres à état mixte. En effet, les densités de probabilités sur l'identité de la personne suivie sont indépendantes d'un filtre à l'autre. Deux filtres peuvent produire simultanément la même identité. Nous souhaitons ici contraindre ceci de manière à produire un appariement filtre/ID exclusif *via* leur interaction au niveau du réseau.

4.1. Génération de tracklets sur les identités

4.1.1. Ajout de la topologie

Dans cette section, nous supposons disposer de la topologie du réseau sur lequel nous travaillons. Cette topologie est représentée par un graphe non orienté $G = (V, E)$ où les nœuds V représentent les zones d'entrées/sorties des caméras, et les arêtes E donnent les transitions possibles entre ces zones, comme présenté en figure 4. Cet *a priori* fixé ici pourrait être appris en ligne à l'aide de méthodes telles que celle proposée par (Chen *et al.*, 2008).

4.1.2. Intégration temporelle

Chaque traqueur produit à chaque instant une distribution discrète de probabilités sur l'ensemble des identités, calculée comme le ratio de particules dédiées à une identité. Ces probabilités sont aggrégées sur une fenêtre temporelle notée T dans un formalisme de type programmation dynamique. On parle alors de *tracklet* à l'instar de (Wojek *et al.*, 2010). Ce faisant, nous construisons une matrice d'association entre les traqueurs et les identités de la base selon l'équation (5).

L'utilisation de la topologie du réseau intervient à ce niveau. Elle permet de supprimer les associations traqueur/identité impossibles. Nous partons d'une localisation initiale des identités dans le réseau. À chaque terminaison de traqueur, nous mettons à jour cette localisation avec sa ré-identification. Nous utilisons cette localisation pour

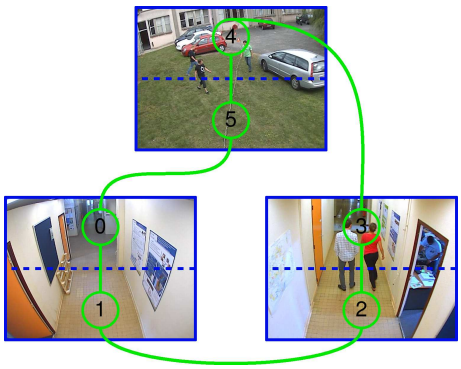


Figure 4. Modélisation de la topologie du réseau de caméras. Un graphe non orienté relie les zones d'entrées/sortie des caméras

mettre à zéro les associations incohérentes avec celle-ci. Une association est dite incohérente si la zone d'entrée/sortie dans laquelle se trouve le traqueur n'est pas connexe avec la dernière localisation enregistrée de l'identité testée.

$$S(tr_{t_0+T}, id_{t_0+T}) = p\left(id_{t_0+T}|zone(tr_{t_0})\right) \cdot \sqrt[T]{\prod_{t=t_0+1}^{t_0+T} \text{Card}\left(\Upsilon_{tr,id_t}\right)} \qquad (5)$$

$$\text{où } \Upsilon_{tr,id_t} = \left\{p|\mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, id_t)^\mathsf{T}\right\}$$

et où

$$p\left(id|zone(tr)\right) = \begin{cases} 1 & \text{si localization[id]=zone(tr)}\,; \\ 0 & \text{sinon.} \end{cases}$$

4.1.3. Exclusivité de l'association

Une affectation exclusive de même type que celle décrite en section 3, travaillant à partir de la fonction de similarité (5) permet d'obtenir une association exclusive traqueurs/identités en fin de fenêtre temporelle. La topologie et les ré-identifications précédentes interviennent pour supprimer des possibilités. Finalement, l'association traqueurs/détections impose une exclusivité entre les paires résultantes.

La gestion des identités au sein du suivi permet d'éviter les problèmes de combinatoire inhérents à la gestion de pistes multiples et de maintenir constamment à jour la répartition dans le réseau des *identités globales* évoluant à l'intérieur de celui-ci.

4.2. Optimisation des tracklets sur une séquence de suivi

Ces affectations supervisées interviennent à la fin de chaque fenêtre temporelle. Chaque fenêtre temporelle fournit une identification pour toute sa durée. Nous obtenons

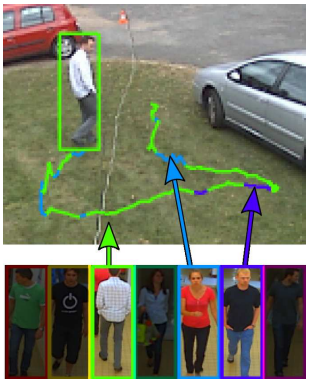


Figure 5. Tracklets d'identités au cours d'une séquence de suivi

ici des *tracklet* d'identités. La figure 5 présente les différentes *tracklet* d'identités inférées par le superviseur pour une séquence de suivi.

Pour ne pas biaiser le processus de ré-identification sur le début de la séquence de suivi, nous remettons la distribution des identités du filtre à état mixte à l'équiprobabilité à chaque fin de fenêtre temporelle. Ainsi, le processus de recherche d'identité de la cible propose à nouveau toutes les identités de la base, et converge à nouveau vers une en particulier, selon les observations qu'il reçoit.

Pour chacun des traqueurs actifs nous mémorisons ces ré-identifications dans un accumulateur indexé sur les identités de la base. Suivant les principes de la programmation dynamique, à chaque instant l'identité affectée à ce traqueur est le mode prédominant dans cet accumulateur (vote majoritaire). De la même manière, lorsqu'un traqueur s'arrête, il se voit affecter l'identité ayant eu le plus de votes dans son accumulateur, et la localisation dans le graphe topologique de cette identité est mise à jour.

5. Implémentation et évaluations associées

5.1. Implémentation

Notre réseau IP fournit une fréquence moyenne de 16 images par seconde pour le flux vidéo à traiter. Nous fixons donc $\Delta t = 1/16s$ dans le modèle d'évolution des filtres particulaires. Dans le modèle d'observation des particules, équation (2), nous fixons de manière empirique :

$$\begin{cases} \alpha = 0.90,\ \beta = 0.05 \text{ et } \gamma = 0.05 & \text{si } \mathcal{I}(tr) = 1 \text{ (Breitenstein } \textit{et al.}\text{, 2010)} \\ \alpha = 0.0,\ \beta = 0.8 \text{ et } \gamma = 0.2 & \text{sinon, (Meden } \textit{et al.}\text{, 2011).} \end{cases}$$

Dans le superviseur, la durée des intégrations temporelles est fixée à 7 images, ce qui correspond au temps moyen de convergence des filtres à état mixte sur leur identité.

5.2. Évaluations

5.2.1. Jeux de données

Nous évaluons les différentes composantes de notre approche sur deux jeux de données. Tout d'abord nous testons le traqueur dédié à chaque nœud/caméra, sans, puis avec le module de ré-identification actif, sur la séquence PETS'09 S2L1 [5]. Cette séquence publique, longue de 795 images, présente un espace ouvert, dans lequel évoluent 10 individus, avec croisements et entrées/sorties. Ayant labellisé ce jeu de données, nous sommes en mesure de quantifier les résultats de notre algorithme de suivi.

Au niveau du réseau à champs disjoints et étant donné l'absence de jeux de données publics associés, nous évaluons la composante de supervision sur une séquence privée notée NOFOVNetwork dans la suite. La séquence présente un ensemble de 7 personnes transitant entre 3 caméras. Il n'y a pas de champs de vue communs entre les caméras, deux sont placées en intérieur de bâtiment, alors que la troisième surveille un espace ouvert extérieur avec une configuration similaire à la séquence PETS'09. La séquence représente 837 images. Notre objectif est de rendre ces données publiques.

5.2.2. Critères et modalités évalués

Nous utilisons les métriques CLEAR MOT (Bernardin, Stiefelhagen, 2008) pour la quantification des résultats de suivi. Nous obtenons un score de précision : MOTP (*Multi-Object Tracking Precision*), calculé comme le rapport de l'intersection sur l'union des boîtes de suivi avec celles de la vérité terrain, et un score d'« accuracy » : MOTA (*Multi-Object Tracking Accuracy*) prenant en compte les faux positifs, les faux négatifs et les changements de cibles des traqueurs.

5.2.2.1. MOTP

La *Multiple Object Tracking Precision* se définit comme :

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$

Il s'agit de l'erreur totale dans les positions estimées pour les paires cibles-positions vérité terrain sur toutes les images de la séquence, moyennée par le nombre d'appariements faits. Cela met en avant la propension de l'algorithme à estimer précisément la position des objets, indépendamment de sa capacité à reconnaître une configuration de cibles, conserver des trajectoires consistantes…

5. http://www.cvg.rdg.ac.uk/PETS2009/a.html

5.2.2.2. MOTA

La *Multiple Object Tracking Accuracy* se définit comme :

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t},$$

où m_t, fp_t, et mme_t sont respectivement les nombres de cibles manquées, de faux positifs et de mauvaises associations, au temps t. Ce critère peut être vu comme dérivé des trois ratios d'erreurs suivants. Le ratio de cibles manquées dans la séquence, calculé par rapport au nombre total de cibles présentes dans l'ensemble de la séquence est donné par :

$$\bar{m} = \frac{\sum_t m_t}{\sum_t g_t},$$

le ratio de faux positifs est donné par :

$$\bar{fp} = \frac{\sum_t fp_t}{\sum_t g_t},$$

et le ratio de mauvaises associations par :

$$\bar{mme} = \frac{\sum_t mme_t}{\sum_t g_t}.$$

5.2.2.3. TRR

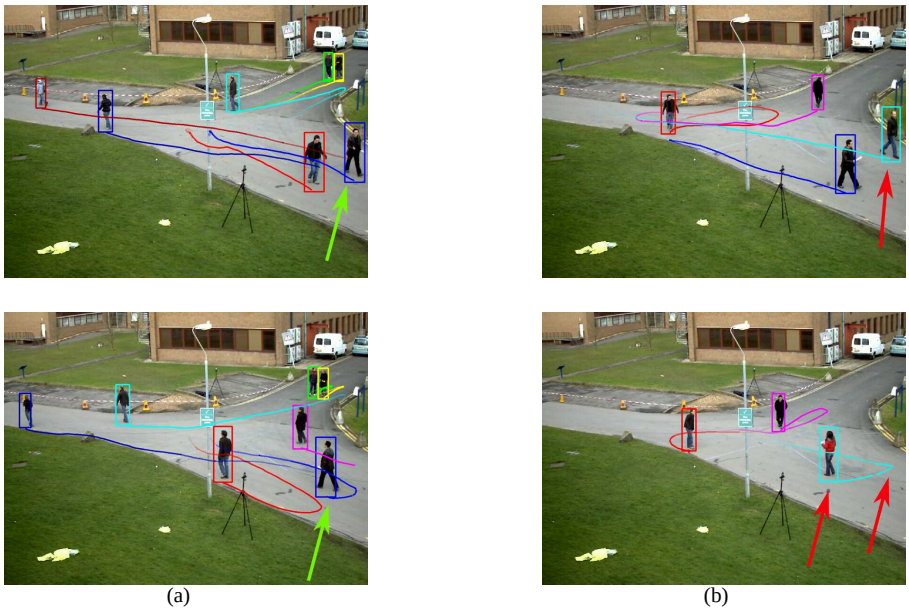
Nous évaluons les capacités de ré-identification par un taux de ré-identification correcte TRR (pour *True Re-identification Rate*), calculé comme le rapport du nombre de ré-identifications correctes sur le nombre de ré-identifications totales. Étant donné que le superviseur opérera sur une fenêtre temporelle, les critères TRR sont estimés en fin de fenêtre temporelle.

5.3. Performances au niveau caméra

*5.3.1. Notion d'*identité globale

Figure 6. Base des 10 identités de la séquence PETS

La figure 6 présente la base d'identités utilisée pour traiter la séquence PETS. Il s'agit ici d'un contexte monocaméra. Les images de la base sont donc issues de la caméra où est réalisé le suivi. Nous présentons ici quelques images de la séquence, pour chaque identité.



(a) (b)

Figure 7. Résultats issus de Breitenstein et al. *(a). Entre les images 204 et 241 la personne fléchée sort puis entre à nouveau. Un traqueur est maintenu et apparie la bonne personne, sur un critère spatial. (b) La même situation se produit entre les images 390 et 445. Mais cette fois, une nouvelle personne entre, un traqueur déjà existant lui est affecté, la trajectoire est reprise. Il s'agit d'une erreur de ré-identification*

Les figures 7 et 8, présentent la limitation d'une simple gestion d'*identités locales* et l'apport de notre modalité de ré-identification. Sur la figure 7, lorsqu'une personne sort et une personne différente entre, les trajectoires sont raboutées. Un simple critère spatial est utilisé dans (Breitenstein *et al.*, 2010). La figure 7 (b) met en exergue cette limitation lorsque la personne sortie n'est pas la même que celle qui entre. Le traqueur serait également pris en défaut si la personne suivie réapparaissait dans une autre région de l'image, typiquement dans un réseau de couloirs.

Dans notre cas (figure 8), à chaque instant, chaque traqueur propose une distribution de probabilité d'identité observée. Ceci permet d'accepter des périodes de non-observabilité comme une sortie de caméra puis de ré-initialiser le traqueur avec le bon identifiant. Lorsqu'une personne entre, le traqueur qui la suit va converger vers des identités de la base.

5.3.2. Performances quantitatives

Le tableau 1 présente nos résultats quantitatifs sur les deux séquences utilisées. PETS'09 nous a permis dans un premier temps de valider notre implémentation de

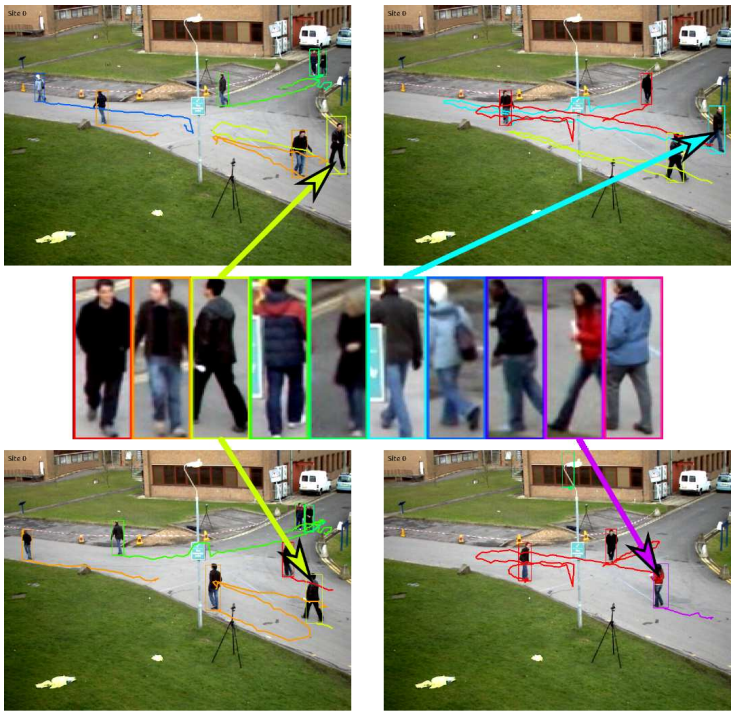


Figure 8. Intérêt de la ré-identification intégrée au suivi multipiste : en (a) comme en (b), le système ré-identifie la piste suivie par rapport à la base d'identités et permet de détecter que la personne est différente (b)

(Breitenstein *et al.*, 2010), dont certains aspects n'ont pas été implémentés (utilisation de la confiance du détecteur dans le modèle d'observation, modèle d'apparence de type Boosting Online).

Cependant, notre approche dispose d'une modalité supplémentaire avec la notion d'*identité globale*. Nous montrons d'abord que l'introduction du filtrage particulaire à état mixte ne dégrade pas les performances de suivi, en comparant MOTP et MOTA pour notre implémentation sans et avec le module de ré-identification actif. Puis cette modalité supplémentaire permet d'exprimer le taux de ré-identification pour la séquence. Finalement, nous comparons ces résultats de ré-identification seule à l'approche filtres supervisés, dans laquelle l'exclusivité entre les ré-identifications est imposée (section 4). Cette contrainte d'exclusivité induit des scores de ré-identification meilleurs.

L'aspect stochastique du filtrage particulaire est pris en compte : le tableau 1 présente les résultats moyens de chaque score, sur un ensemble de 10 répétitions.

Tableau 1. Résultats de suivi selon les métriques CLEAR MOT et taux de ré-identification sur la séquence monocaméra PETS'09 S2L1. Nous donnons ici les Multi-Object Tracking Precision (MOTP), Multi-Object Tracking Accuracy (MOTA), et True Re-identification Rate (TRR) définis en section 5.2

Séquence PETS'09	MOTP	MOTA	TRR
Suivi-par-détection (Breitenstein *et al.*, 2010)	**56,3 %**	**79,7 %**	-
Suivi-par-détection implémenté	42,7 %	77,9 %	-
Suivi-par-Réidentification	42,5 %	77,7 %	59,7 %
Suivi-par-Réidentification supervisé	42,4 %	75,9 %	**64,0 %**

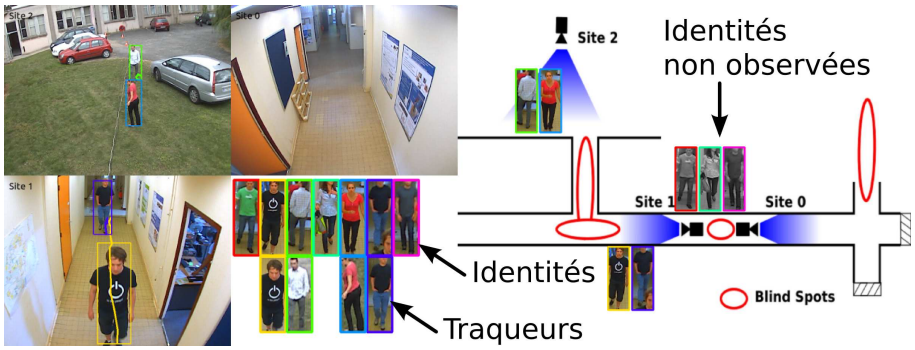


*Figure 9. Exemple de suivi dans le réseau avec maintien d'*identités globales *sur les pistes, permettant de les localiser dans la topologie du réseau*

5.4. Performances du superviseur

La séquence NOFOVNetwork n'étant pas annotée pour le suivi, nous ne présentons que des taux de ré-identifications pour cette séquence. Nous comparons ici la méthode se basant uniquement sur les informations de couleur introduites dans le filtrage particulaire (approche inspirée de (Meden *et al.*, 2011)), avec le système supervisé que nous proposons en section 4 et son ajout de contraintes topologiques.

Le tableau 2 présente les taux de ré-identification par caméra, puis au niveau du réseau global. La base étant construite à partir de la caméra 1, ceci explique les taux de ré-identification supérieurs dans cette caméra. Ces résultats illustrent l'apport du superviseur : chaque identité correctement ré-identifiée contraint le système dans la suite de la topologie.

Finalement, la figure 9 représente la sortie de notre méthode, avec les suivis dans les images et la localisation des identités dans la topologie du réseau. Les identités non observées, sont les identités de la base que les filtres considèrent ne pas être en train de suivre. Ces identités sont localisées dans les zones « blind spots » adjacentes à leur dernière zone d'observation pour affichage (représentées en niveaux de gris dans la

Tableau 2. Taux de ré-identifications correctes TRR pour chacune des caméras du réseau NOFOVNetwork : comparaison des approches sans, et avec superviseur sur le réseau

Séquence NOFOV	cam0	cam1	cam2	réseau
Suivi-par-Réidentification	43,7 %	67,3 %	55,5 %	54,6 %
Suivi-par-Réidentification supervisé	**67,7 %**	**76,9 %**	**63,8 %**	**68,2 %**

carte du bâtiment). Cette localisation est utilisée pour contraindre les ré-identifications futures (cf. section 4).

6. Conclusion et perspectives

Cet article traite de la surveillance de personnes évoluant dans des réseaux de caméras à champs disjoints, *i.e* localiser en ligne les pistes dans la topologie. Ceci passe par la notion d'*identité globale*. Nous avons présenté une méthode de suivi par ré-identification, travaillant à deux niveaux en se basant respectivement sur des signatures colorimétriques et sur des contraintes spatiotemporelles dans le réseau. Le niveau caméra est traité par un tracking-par-détection markovien inspiré de (Breitenstein *et al.*, 2010), enrichi par la notion d'*identité globale* prise en compte au sein des filtres particulaires avec le formalisme d'état mixte. Ainsi, chaque traqueur maintient une multimodalité sur l'identité qu'il est en train de suivre. Ce faisant, il intègre la capacité de ré-initialisation après disparition puis ré-apparition dans le champ de vue de la caméra. Ces distributions d'identités, considérées comme des *tracklet* sur les identités, sont filtrées spatiotemporellement au niveau du réseau par un superviseur. Celui-ci impose l'exclusivité des ré-identifications et s'assure de leur cohérence dans la topologie du réseau.

Une première extension vise l'apprentissage en ligne des images-clés et leur mise à jour au fur et à mesure du suivi. Il ne s'agit cependant pas là d'un problème simple, car incrémenter la représentation d'une identité dans différentes caméras suppose d'avoir résolu le problème de la ré-identification. Les travaux de (Kuo *et al.*, 2010) proposent une première solution à ce problème avec le recours à un « apprentissage flou » MILboost, autorisant les erreurs de labellisation dans les données lors de l'apprentissage. Un modèle d'apparence plus évolué entraîné en ligne sur la cible qu'il caractérise rendrait par ailleurs ceci plus simple. Finalement, une analyse dans le plan du sol et non image pour le traqueur serait à considérer.

Bibliographie

Bar-Shalom Y., Fortmann T., Scheffe M. (1980). Joint probabilistic data association for multiple targets in clutter. In *Proc. conf. on information sciences and systems*, p. 404–409.

Benfold B., Reid I. (2011). Stable multi-target tracking in real-time surveillance video. In *Proceedings of the international conference on computer vision and pattern recognition.*

Bernardin K., Stiefelhagen R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *Journal on Image and Video Processing*.

Breitenstein M., Reichlin F., Leibe B., Koller-Meier E., Van Gool L. (2010). Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE *Transactions on Pattern Analysis and Machine Intelligence*.

Burgeois F., Lasalle J.-C. (1971). An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*.

Chen K., Lai C., Hung Y., Chen C. (2008). An adaptive learning method for target tracking across multiple cameras. In *Proceedings of the international conference on computer vision and pattern recognition.*

Cheng D. S., Cristani M., Stoppa M., Bazzani L., Murino V. (2011). Custom pictorial structures for re-identification. In *Proceedings of the british machine vision conference.*

Cox I., Hingorani S. (1996). An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, n° 2, p. 138–150.

Dalal N., Triggs B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the international conference on computer vision and pattern recognition.*

Danchick R., Newnam G. (2006). Reformulating reid's mht method with generalised murty k-best ranked linear assignment algorithm. In *Radar, sonar and navigation, iee proceedings-*, vol. 153, p. 13–22.

Ess A., Leibe B., Van Gool L. (2007). Depth and appearance for mobile scene analysis. In *Proceedings of the international conference on computer vision*, p. 1–8.

Farenzena M., Bazzani L., Perina A., Murino V., Cristani M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the international conference on computer vision and pattern recognition.*

Forssén P. (2007). Maximally stable colour regions for recognition and matching. In *Proceedings of the international conference on computer vision and pattern recognition*, p. 1–8.

Gheissari N., Sebastian T., Hartley R. (2006). Person reidentification using spatiotemporal appearance. In *Proceedings of the international conference on computer vision and pattern recognition.*

Gray D., Brennan S., Tao H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. ieee international workshop on performance evaluation for tracking and surveillance (pets).*

Gray D., Tao H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the european conference on computer vision.*

Huang T., Russell S. (1997). Object identification in a bayesian context. In *Proceedings of the international joint conference on artificial intelligence.*

Isard M., Blake A. (1998a). Condensation-conditional density propagation for visual tracking. *International journal of computer vision*, vol. 29, n° 1, p. 5–28.

Isard M., Blake A. (1998b). A mixed-state CONDENSATION tracker with automatic model-switching. In *Proceedings of the international conference on computer vision.*

Isard M., Blake A. (2001). BraMBLe: a Bayesian multiple blob tracker. In *Proceedings of the international conference on computer vision.*

Javed O., Shafique K., Shah M. (2005). Appearance modeling for tracking in multiple non-overlapping cameras. In *Proceedings of the international conference on computer vision and pattern recognition.*

Kaucic R., Perera A., Brooksby G., Kaufhold J., Hoogs A. (2005). A unified framework for tracking through occlusions and accross sensor gaps. In *Proceedings of the international conference on computer vision and pattern recognition.*

Kettnaker V., Zabih R. (1999). Bayesian multi-camera surveillance. In *Proceedings of the international conference on computer vision and pattern recognition.*

Kuo C., Huang C., Nevatia R. (2010). Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *Proceedings of the european conference on computer vision.*

Kuo C., Nevatia R. (2011). How does person identity recognition help multi-person tracking? In *Proceedings of the international conference on computer vision and pattern recognition.*

Lev-Tov A., Moses Y. (2010). Path recovery of a disappearing target in a large network of cameras. In *Proceedings of the international conference on distributed smart cameras.*

Makris D., Ellis T., Black J. (2004). Bridging the gaps between cameras. In *Proceedings of the international conference on computer vision and pattern recognition.*

Matei B., Sawhney H., Samarasekera S. (2011). Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In *Proceedings of the international conference on computer vision and pattern recognition.*

Meden B., Sayd P., Lerasle F. (2011). Mixed-State Particle Filtering for Simultaneous Tracking and Re-Identification in Non-Overlapping Camera Networks. In *Proceedings of the scandinavian conference on image analysis (scia).* Ystad, Suède.

Meden B., Sayd P., Lerasle F. (2012). Suivi par ré-identification dans un réseau de caméras à champs disjoints. In *Actes du congrès francophone sur la reconnaissance des formes et l'intelligence artificielle (rfia).* Lyon.

Oh S., Russell S., Sastry S. (2004). Markov chain monte carlo data association for general multiple-target tracking problems. In *Cdc*.

Okuma K., Taleghani A., De Freitas N., Little J., Lowe D. (2004). A boosted particle filter: multitarget detection and tracking. In *Proceedings of the european conference on computer vision.*

Oreifej O., Mehran R., Shah M. (2010). Human identity recognition in aerial images. In *Proceedings of the international conference on computer vision and pattern recognition.*

Pasula H., Russell S., Ostland M., Ritov Y. (1999). Tracking many objects with many sensors. In *Proceedings of the international joint conference on artificial intelligence.*

Prisacariu V., Reid I. (2009). *fasthog - a real-time gpu implementation of hog*. Rapport technique n° 2310/09. Department of Engineering Science, Oxford University.

Prosser B., Zheng W., Gong S., Xiang T., Mary Q. (2010). Person Re-Identification by Support Vector Ranking. In *Proceedings of the british machine vision conference.*

Qu W., Schonfeld D., Mohamed M. (2007). Distributed bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. *Int. Journal EURASIP*.

Reid D. (1979). An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, vol. 24, n° 6, p. 843-854.

Wojek C., Roth S., Schindler K., Schiele B. (2010). Monocular 3D scene modeling and inferences: understanding multi-object traffic scenes. *Proceedings of the european conference on computer vision.*

Xing J., Ai H., Lao S. (2009). Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. *Proceedings of the international conference on computer vision and pattern recognition*, p. 1200-1207.

Zajdel W., Kröse B. (2005). A sequential bayesian algorithm for surveillance with nonoverlapping cameras. *International Journal of Pattern Recognition and Artificial Intelligence.*

Zheng W., Gong S., Xiang T. (2009). Associating groups of people. *Proceedings of the british machine vision conference*, vol. 7.

Zheng W., Gong S., Xiang T. (2011). Person re-identification by probabilistic relative distance comparison. In *Proceedings of the international conference on computer vision and pattern recognition.*

Boris Meden *effectue son doctorat en vision par ordinateur au département de la recherche technologique du Commissariat à l'Energie Atomique (CEA). En 2009, Boris Meden a obtenu son diplôme d'ingénieur de l'ISIMA et son master recherche en vision par ordinateur de l'Université Blaise Pascal, à Clermont-Ferrand.*

Frédéric Lerasle *est maître de conférence à l'Université Paul Sabatier depuis septembre 1997, et chercheur en vision pour la robotique au LAAS-CNRS de Toulouse. Son doctorat obtenu à l'Université Blaise Pascal de Clermont-Ferrand en 1997, effectué au LASMEA, traite de la capture de mouvements humains par capteurs optiques multiples. Ses recherches actuelles au LAAS-CNRS portent sur la vision pour la robotique, et plus particulièrement : (1) la détection, la reconnaissance et le suivi de personnes, tout comme l'interprétation de gestes et d'activités pour l'interaction homme-robot, (2) la détection/reconnaissance de marqueurs pour la navigation métrique et topologique de robots mobiles en environnements intérieurs.*

Patrick Sayd *a obtenu le grade de docteur en vision par ordinateur en 1996 à l'Université de Clermont-Ferrand (LASMEA) avec une thèse portant sur la reconstruction 3-D de formes complexes. Il a commencé sa carrière de chercheur au CEA en 1998 sur des sujets de reconstruction 3-D de sites industriels. Depuis 2001, il est en charge de plusieurs projets de recherche en reconstruction 3-D (ARCO IST-2000) et de vidéosurveillance (Clovis Eureka-2002, ISCAPS PASR-2004). Ses activités de recherche concernent plus particulièrement l'analyse de scène pour l'interprétation des activités humaines. (ITEA-2006 NUADU, FPT-Security SUBITO-2009, ANR Surtrain).*