Détection et ré-identification de personnes dans un réseau de caméras

D.N. Truong Cong, Catherine Achard, L. Khoudour

Détection et ré-identification de personnes dans un réseau de caméras

D-N. Truong Cong[1] C. Achard[2] L. Khoudour[1]

[1]Institut National de Recherche sur les Transports et leur Sécurité, LEOST
20 rue Elisée Reclus, F-59650 Villeneuve d'Ascq, France.
[2]Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie/CNRS
UMR 7222, BC 173, 4 place Jussieu, F-75005 Paris, France

truong@inrets.fr, catherine.achard@upmc.fr, louahdi.khoudour@inrets.fr

Résumé

Cet article présente un système de vidéo-surveillance complet permettant de ré-identifier des personnes qui se déplacent dans différents sites surveillés par des caméras dont les champs de vision ne se recouvrent pas. Le premier apport de cet article réside dans la détection robuste d'objets en mouvement dans des conditions difficiles. Celle-ci est réalisée en fusionnant les résultats de deux méthodes de détection complémentaires, ce qui permet d'obtenir des régions de bonne qualité, et un algorithme s'adaptant très rapidement au manque de stationnarité du fond. Malgré cet apport, beaucoup de bruit réside dans les régions extraites, c'est pourquoi nous avons mis en place un processus de classification de régions à partir d'un double critère de stabilité temporelle et colorimétrique. Le but de celui-ci est de sélectionner, parmi toutes les régions détectées, celles qui correspondent réellement à la silhouette. Ces régions, nommées régions clés, sont ensuite caractérisées par une nouvelle signature modélisant à la fois la distribution des couleurs et leur répartition spatiale. La ré-identification de personnes est alors réalisée en utilisant une approche qui analyse les propriétés spectrales du graphe de similarité. Le système a été testé sur deux bases de données réelles difficiles représentant le passage de plusieurs personnes dans différents sites. Les résultats montrent d'une part que de très bons taux de ré-identification sont obtenus et d'autre part que ceux-ci ne sont que très peu altérés par une détection automatique des silhouettes.

Mots Clef

Vidéosurveillance, Détection de mouvement, Ré-identification de personne, Signature colorimétrique.

Abstract

This paper presents an automatic system for detecting and re-identifying people moving in different sites monitored by cameras with non-overlapping views. The first contribution of this article is a robust algorithm for moving object detection in difficult conditions, which combines the results of two complementary detection methods : one allows us to extract the moving objects with highly precise silhouettes and another adapts quickly to lighting changes of the background. Despite the advantages of this approach, the extraction process still produces noisy results. Therefore, we carry out a classification process of the extraction results based on their temporal and colorimetric stability in order to select, among all detected regions, those that correspond to the silhouette. These regions, called key-regions, are then characterized by a new appearance-based signature including both color and spatial feature of silhouttes. People re-identification is now carried out by estimating the similarities of passages thanks to a graph-based approach. The global system was tested on two real and difficult data sets recorded in very different environments. The experimental results show that our proposed system leads to very satisfactory results.

Keywords

Surveillance system, Background subtraction, Motion detection, People re-identification, Color-based descriptor.

1 Introduction

Ces dernières années, les systèmes de surveillance visuelle ont connus un essor important de part l'augmentation de l'insécurité. Dans le but de diminuer les menaces telles que des agressions contre des personnes, du vandalisme ou des actes terroristes, de nombreuses caméras ont envahi les places publiques. La surveillance manuelle de ces écrans est une tâche fastidieuse en raison de la grande quantité d'information qui circule. Il est donc très intéressant d'automatiser cette procédure à partir de systèmes de traitement d'images, capables d'extraire l'information utile des vidéos, et de l'interpréter. Une des tâches les plus importantes consiste à établir des correspondances entre les personnes qui peuvent apparaître et réapparaître à différents instants dans le réseau de caméras.

De nombreux travaux ont été menés sur la reconnaissance de personnes à partir de modèles d'apparence. Nakajima et al. [1] ont proposé un système capable de reconnaître des personnes dans un environnement intérieur. Celui-ci uti-

lise des SVMs qui sont appris à partir de caractéristiques de forme et de couleur extraites des silhouettes. Gheissari et al. [2] proposent une signature temporelle invariante à la position du corps et à l'apparence dynamique des vêtements. Les travaux de Wang et al. [3] se concentrent sur la modélisation de la distribution spatiale de la couleur de différentes parties d'un objet afin d'extraire des descripteurs fortement distincts. Yu et al. [4] ont introduit un modèle d'apparence construit à partir d'une estimation par noyau. Une procédure de sélection d'images clé permet ensuite de représenter l'information contenue dans les vidéos et de comparer celles-ci.

Dans cet article, nous présentons un système capable de ré-identifier des personnes se déplaçant dans un site lorsqu'elles ont déjà été observées dans le champ de vue d'une autre caméra du réseau. Dans un premier temps, nous avons introduit une méthode robuste permettant d'extraire automatiquement les objets en mouvement dans la scène. Celle-ci combine une détection par modélisation du fond et une détection par différence d'images successives et permet d'obtenir un algorithme qui s'adapte très rapidement à la non-stationnarité du fond et qui fournit des régions de bonne qualité.. Un classement des régions extraites est ensuite nécessaire afin de déterminer celles qui correspondent à la silhouette de la personne. Cette étape est très délicate à cause de la difficulté des séquences étudiées : train en déplacement. Ainsi, le pur suivi temporel des régions s'est avéré inexploitable. Aussi, nous avons proposé d'étudier les régions sur une longue durée temporelle en utilisant deux critères : un de stabilité colorimétrique et un de stabilité temporelle. Les régions ainsi obtenues, nommées régions clés, sont caractérisées par des signatures spatio-colorimétriques quasi invariantes aux conditions d'éclairage. Toutes ces signatures sont ensuite exploitées avec l'analyse spectrale qui permettra d'extraire les informations utiles afin de procéder à la ré-identification.

L'article est organisé comme suit. La section 2 présente l'approche proposée pour extraire les régions en mouvement. La sélection des résultats d'extraction permettant de ne conserver que les régions correspondant à la silhouette est détaillée en section 3. Dans la section 4, nous introduisons la signature proposée pour caractériser les silhouettes, ainsi que les procédures de normalisation couleur mises en place pour obtenir l'invariance aux conditions d'éclairage. L'utilisation de l'analyse spectrale pour réaliser la ré-identification des personnes est ensuite explicitée dans la section 5. Des résultats sur les performances de système testé sur deux bases de données réelles sont introduits dans la section 6, juste avant la conclusion et les perspectives à court terme, section 7.

2 Détection des zones en mouvement

La première étape du système consiste à détecter des zones en mouvement afin de pouvoir les caractériser par la suite. Afin d'obtenir une détection robuste face aux conditions extrêmes des séquences que nous traitons (train en mouvement), une approche mixant deux méthodes est proposée : une détection par modélisation du fond et une approche par différence d'images successives. Cette fusion de méthodes permet de tirer partie des avantages de chacune d'elles : (i) des silhouettes de bonne qualité sont obtenues grâce à la soustraction de l'image de fond. (ii) Le système s'adapte très rapidement aux changements d'éclairage grâce à la considération des images successives. Le synopsis de l'algorithme de détection est présenté sur la figure 1.

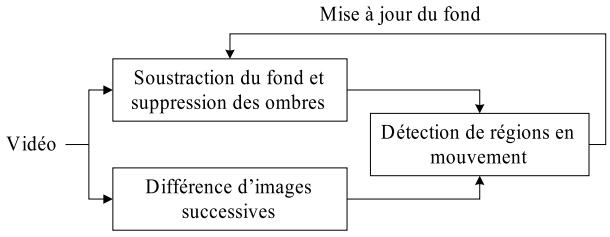


FIGURE 1 – Algorithme de détection des zones en mouvement.

La détection par soustraction de fond est réalisée de manière classique, en modélisant le fond par une mixture de gaussienne [5]. Cette méthode, très fréquemment utilisée dans la littérature, consiste à modéliser l'historique de chaque pixel par un mélange de gaussiennes. Ainsi, à chaque pixel est associée une somme pondérée de N_g gaussiennes ($2 \leq N_g \leq 5$). La probabilité d'observer la valeur courante du pixel $I^t(p)$ est estimée par :

$$p\left(I^t(p) \middle| \mathbf{I}_p\right) = \sum_{i=1}^{N_g} w_i^t . \eta\left(I^t(p); \vec{u}_i^t, \Sigma_i^t\right) \qquad (1)$$

où $\mathbf{I}_p = \left\{I^1(p), I^2(p), \ldots, I^t(p)\right\}$ est l'historique des valeurs prises par le pixel ; w_i^t est le poids de la gaussienne i à l'instant t ; $\vec{u}_i^t$ et Σ_i^t sont la moyenne et la matrice de covariance de la gaussienne. Si de plus les trois canaux couleur sont considérés indépendants, la matrice de covariance Σ_i^t devient une matrice diagonale avec la variance des canaux σ_i^t comme éléments. Le pixel courant $I^t(p)$ appartient à la gaussienne i s'il vérifie :

$$\left\|I^t(p) - \vec{u}_i^t\right\| < K\sigma_i^t \qquad (2)$$

où K est un paramètre de la méthode.

Les paramètres des gaussiennes qui valident cette condition sont mis à jour avec :

$$\begin{aligned} w_k^t &\leftarrow (1-\alpha_p)\, w_k^{t-1} + \alpha_p \\ \vec{u}_k^t &\leftarrow \left(1 - \tfrac{\alpha_p}{w_k^t}\right) \vec{u}_k^{t-1} + \tfrac{\alpha_p}{w_k^t} I^t(p) \\ {\sigma_{k,c}^t}^2 &\leftarrow \left(1 - \tfrac{\alpha_p}{w_k^t}\right) {\sigma_{k,c}^{t-1}}^2 + \tfrac{\alpha_p}{w_k^t}\left(I_c^t(p) - \mu_{k,c}^t\right)^2 \end{aligned} \qquad (3)$$

où α_p est le coefficient de mise à jour du pixel p.

Contrairement à l'approche proposée dans [5], nous faisons varier ce coefficient d'un pixel à l'autre en fonction d'une matrice de coefficient de mise à jour $\mathbf{A}$ estimée à la fin de

l'étape de détection de régions en mouvement. Ceci permet, lorsqu'une personne se déplace face à la caméra, de ne pas l'incorporer dans l'image de fond.
Pour les autres gaussiennes qui ne valident pas la condition (2), le poids est mis à jour avec :

$$w_k^t \leftarrow (1-\alpha_p)\, w_k^{t-1} \tag{4}$$

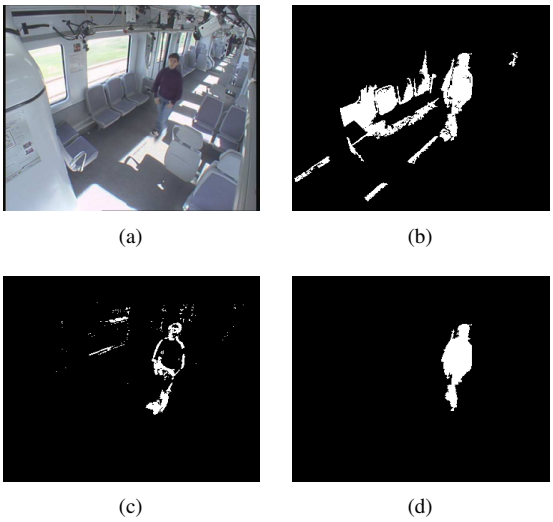
L'ensemble des gaussiennes est ensuite ordonné selon un ordre décroissant des valeurs de w_k^t/σ_k^t. Les C gaussiennes décrivant le fond sont alors sélectionnées :

$$C = \arg\min_c \left(\sum_{i=1}^{c} {w_i}^t > S \right) \tag{5}$$

où S est un seuil fixé.
Si le pixel courant $I^t(p)$ appartient à une de ces C gaussiennes, il sera considéré comme statique. Dans le cas inverse, il sera détecté en mouvement. On obtient ainsi une image de détection binaire. Afin d'améliorer cette détection, un algorithme de suppression des ombres [6] est ensuite mis en place.
Malgré les performances de cette méthode, les résultats sont encore très bruités, en particulier lors de changements soudain d'éclairage. Ils vont donc être fusionnés avec une détection de mouvement par différence d'images successives. Celle-ci s'obtient en considérant trois images successives :

$$\begin{aligned} M^t(p) = & \left(\frac{\left|I^t(p) - I^{t-1}(p) - \mu_1\right|}{\sigma_1} > S_M \right) \\ & \cup \left(\frac{\left|I^{t-1}(p) - I^{t-2}(p) - \mu_2\right|}{\sigma_2} > S_M \right) \end{aligned} \tag{6}$$

où μ_1 et σ_1 sont la moyenne et l'écart type de $\left|I^t(p) - I^{t-1}(p)\right|$ et S_M est un seuil de différence. Les deux détections de mouvement sont fusionnées en recherchant, parmi les régions issues de la soustraction de fond, celles qui sont en mouvement et représentent donc les objets réellement mobiles de la scène (la silhouette dans notre application). Elles correspondent aux régions dont un fort pourcentage de pixels a été détecté en mouvement par la seconde méthode. Une illustration de ce résultat final de détection est présentée sur la figure 2. A cause d'un fort changement d'éclairage dû au déplacement du train, de nombreuses régions sont détectées en mouvement après la soustraction de fond, la suppression des ombres et des petites régions (figure 2(b)). En effet, le fond n'a pas encore eu le temps de se mettre à jour. La détection par différence d'images successives, qui s'adapte très vite aux changements, permet de supprimer ces sur-détections, comme on peut le constater figure 2(d).
Le résultat final de détection est utilisé pour réactualiser les coefficients de la matrice de mise à jour du fond $\mathbf{A} = [\alpha(p)]$ en affectant une faible valeur à $\alpha(p)$ lorsque le pixel est détecté en mouvement et une forte valeur dans le cas inverse. Ceci permet de ne pas intégrer la personne dans le fond lorsque celle-ci se déplace face à la caméra.



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

FIGURE 2 – Les différentes étapes de la détection de régions en mouvement.

(a) Image originale. (b) Détection avec modélisation par mixture de gaussiennes, suppression des ombres et des petites régions. (c) Détection de mouvement par différence d'images successives. (d) Résultat final combinant les deux détections.

Malgré cette détection robuste, de nombreuses régions parasites persistent ainsi que des problèmes de détections (silhouette coupée en deux par exemple). Une étape de plus haut niveau est donc mise en place afin de ne conserver que les régions correspondant à des silhouettes entières, régions que nous nommons régions clés dans la suite de cet article.

3 Sélection des régions clés

Le but de cette étape est de sélectionner, parmi toutes les régions issues de la détection, celles qui correspondent réellement à la silhouette de la personne. Les difficultés des séquences étudiées rendent les algorithmes de suivi temporel inefficaces. Ainsi, nous proposons de classer les régions en région clé ou non en réalisant une étude globale de la séquence exploitant deux critères : la stabilité spatiale et la stabilité colorimétrique des régions.

3.1 Classement selon un critère de stabilité spatiale

Cette première étape a pour but de rechercher, parmi les régions issues de la détection, celles qui présentent une forte stabilité spatiale. En effet, de manière générale, toutes les régions parasites introduites par le bruit, ne valident pas cette propriété. Soit $R = \left\{R^1, R^2, ..., R^T\right\}$, l'ensemble des régions détectées au cours d'une séquence de T images, avec $R^t = \{r_1^t, r_2^t, ..., r_N^t\}$. Les régions dont le centre de gravité est proche dans deux images successives sont regroupées en classes selon l'algorithme suivant :

Initialisation : $\forall\ j \in [1, N]\ ,\ C\left(r_j^1\right) \leftarrow j$,
tant que $t <= T$ **faire**
  pour *chaque region* r_j^t **faire**
    si $\exists\ i / dist\left(r_i^{t-1}, r_j^t\right) < S$ **alors**
    $C\left(r_j^t\right) = C\left(r_i^{t-1}\right)$
    sinon $C\left(r_j^t\right) =$ new class
    t=t+1

où $C\left(r_j^t\right)$ est la classe de la région r_j^t et $dist\left(r_i^{t-1}, r_j^t\right)$ est la distance entre les centres de gravité des régions r_i^{t-1} et r_j^t.

A titre d'illustration, nous présentons les résultats du processus de sélection de classes obtenus pour une séquence correspondant au passage d'une personne. 6 classes ont été construites dont un représentant est présenté figure 3. Parmi les classes obtenues, certaines classes (C2, C3, C5) correspondent à de mauvais résultats de segmentation dus à une forte ressemblance entre le pantalon de la personne et le fond de l'image et un fort changement de luminosité, tandis que les classes C1, C4 et C6 épousent relativement bien la silhouette de la personne. Afin de regrouper ces classes qui ont été divisées suite à des problèmes de détection, un deuxième regroupement utilisant un critère colorimétrique est mis en place.

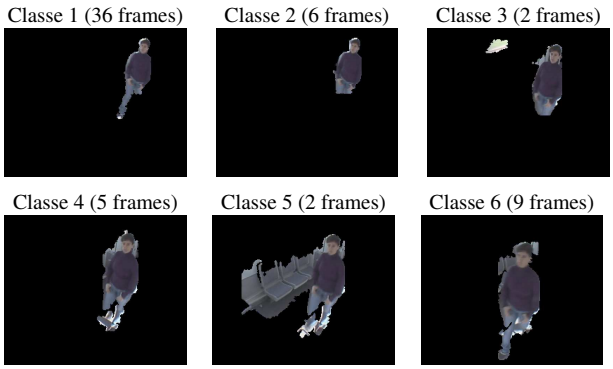


FIGURE 3 – Exemple de classes construites avec un critère de stabilité spatiale.

3.2 Classement selon un critère de stabilité colorimétrique

L'objectif de ce deuxième classement est de regrouper les classes composées de régions ayant les mêmes caractéristiques colorimétriques. En effet, celles-ci ont été introduites par des problèmes de segmentation mais correspondent physiquement au même objet. Il est donc nécessaire, dans un premier temps, de caractériser chaque région r_j^t détectée par un descripteur colorimétrique. Nous avons opté pour un descripteur qui allie des informations colorimétriques et spatiales et qui présente une bonne aptitude à décrire des régions [7]. L'algorithme "leader/suiveur" est ensuite utilisé afin d'obtenir le nouveau regroupement.

Initialisation : $G_1 \leftarrow C_1\ , n = 1$
pour *chaque* $C_k : j = \operatorname{argmin}_i \{dist\left(C_k, G_i\right)\}$ **faire**
  si $dist\ (C_k,\ G_j\ ) < S_C$ **alors** $G_j \leftarrow C_k$
  sinon créer une nouvelle groupe : $n = n + 1$;
  $G_n \leftarrow C_k$

Dans cet algorithme, la distance entre une classe C_k et un groupe G_i est définie par :

$$dist\left(C_k, G_i\right) = \min\left\{dist\left(C_k, C_{ib}\right), b = 1 \ldots B\right\} \quad (7)$$

où C_{ib} est une classe appartenant au groupe G_i. La distance entre classes $dist\left(C_k, C_{ib}\right)$ est définie par :

$$dist\left(C_p, C_q\right) = \frac{1}{P \times Q} \sum_{i=1}^{P} \sum_{j=1}^{Q} d\left(S_{pi}, S_{qj}\right) \quad (8)$$

où S_{pi} est le descripteur de la région i de la classe C_p, P et Q sont les nombres de régions des classes i et j respectivement.
Ces regroupements amènent, pour la même séquence que précédemment, à 4 nouvelles classes comportant respectivement aux regroupements C1+C4+C6 composé de 50 régions, C2 composé de 6 régions, C3 composé de 2 régions et C5 composé de 2 régions. Parmi ces nouvelles classes, celle qui correspond aux silhouettes de la personne est la plus stable et a donc le plus grand nombre de régions. En effet, tous les effets non désirés tels que le bruit ou les problèmes de segmentation apparaissent puis disparaissent au cours de la séquence et n'aboutissent pas à la création de classes dominantes. L'ensemble des régions clés retenu est donc celui correspondant à la plus grande classe en terme de nombre de régions.
La figure 4 représente quelques résultats finaux de régions clés qui représenteront la séquence. Celles-ci vont maintenant être caractérisées afin de représenter la séquence.

FIGURE 4 – Illustration de résultats finaux de régions clés.

4 Caractérisation des silhouettes

Comme la couleur perçue par la caméra dépend de nombreux facteurs tels que la réflectivité de la surface, la couleur de l'illuminant, la géométrie de la source par rapport à l'objet et la caméra, la réponse du capteur,..., une procédure de normalisation couleur doit être mise en place. En

effet, la couleur des régions change au cours d'une même séquence à cause des changements d'éclairage. Elle change aussi lors du changement d'environnement. Plusieurs méthodes ont été proposées dans la littérature pour remédier à ce problème, nous présentons dans cet article les trois invariances qui ont mené aux meilleurs résultats :
- la normalisation Greyworld [8] qui s'obtient, à partir de l'espace RVB, en divisant pour chaque canal, la valeur du pixel par la valeur moyenne de l'image (de la région correspondant à la personne en mouvement dans notre cas) :

$$I_k^* = \frac{I_k}{moyenne\left(I_k\right)} \tag{9}$$

- la normalisation à partir de l'égalisation d'histogramme [9] qui suppose que le rang des couleurs est préservé lors d'un changement d'éclairage. La mesure de rang, pour le niveau i et la canal k est obtenu avec :

$$M_k(i) = \sum_{u=0}^{i} H_k(u) / \sum_{u=0}^{Nb} H_k(u) \tag{10}$$

où Nb est le nombre de pas d'échantillonnage et $H_k()$ est l'histogramme du canal k.
- la normalisation affine [10] définie par :

$$I_k^* = \frac{I_k - moyenne(I_k)}{std\left(I_k\right)} \tag{11}$$

Ces normalisations couleur sont appliquées à l'intérieur de chaque région, avant d'estimer sa signature.
Différentes méthodes de caractérisation de la silhouette ont vu le jour dans la littérature. Le descripteur le plus utilisé à cause de sa simplicité est l'histogramme couleur. Même si l'histogramme est invariant en translation et en rotation autour de l'axe de vue, il est peu discriminant car il ne décrit que la distribution statistique des couleurs de la région étudiée et ne tient pas compte de la répartition spatiale de ses couleurs. Birchfield et Rangarajan [11] ont étendu le concept d'histogramme en introduisant des informations spatiales. Ainsi, le spatiogramme d'ordre deux contient la moyenne et la covariance spatiale des pixels pour chaque case de l'histogramme. Le descripteur couleur/path-length, proposé par Yoon et al. [4] inclut aussi des informations spatiales : chaque pixel de la région est représenté par ses trois composantes couleur et sa distance à un point de référence. La distribution de ces caractéristiques est estimée par un histogramme 4D.
Dans cet article, nous proposons un nouveau descripteur basé sur la modélisation de plusieurs parties de silhouette par sa distribution de couleur. L'idée de ce descripteur consiste à effectuer un découpage horizontal de la silhouette en K bandes équidistantes. Fixer le nombre de bandes plutôt que leur taille permet d'obtenir une invariance face à l'échelle de la personne. Chaque bande est ensuite caractérisée par sa distribution de couleur modélisée par un mélange de gaussiennes.

Soit $X = \{\mathbf{x}_1, \mathbf{x}_2, ., \mathbf{x}_M\}$ l'ensemble des pixels de la bande k de la silhouette, où $\mathbf{x}_m$ est un vecteur couleur de dimension d correspondant aux d composantes couleur. La densité de probabilité de couleur est modélisée par :

$$\mathrm{f}\left(\mathbf{x}\right) = \sum_{i=1}^{N} \alpha_i \, \mathrm{N}\left(\mathbf{x}; \vec{\mu}_i, \Sigma_i\right) \tag{12}$$

où $\vec{\mu}_i$, Σ_i et α_i sont respectivement la moyenne, la matrice de covariance et le poids de la gaussienne i. Les paramètres des gaussiennes sont estimés par l'algorithme Expectation-Maximisation (EM) [12] qui consiste à itérer, jusqu'à la convergence, les deux étapes suivantes :
- L'étape E :

$$p^{(t)}\left(n|m\right) = \frac{\alpha_n^{(t)} \, \mathrm{N}\left(\mathbf{x}_m; \vec{\mu}_n^{(t)}, \Sigma_n^{(t)}\right)}{\sum_{i=1}^{N} \alpha_i^{(t)} \, \mathrm{N}\left(\mathbf{x}_m; \vec{\mu}_i^{(t)}, \Sigma_i^{(t)}\right)} \tag{13}$$

- L'étape M :

$$\alpha_n^{(t+1)} = \frac{1}{M} \sum_{m=1}^{M} p^{(t)}\left(n|m\right) \tag{14}$$

$$\vec{\mu}_n^{(t+1)} = \frac{\sum_{m=1}^{M} \mathbf{x}_m p^{(t)}\left(n|m\right)}{\sum_{m=1}^{M} p^{(t)}\left(n|m\right)} \tag{15}$$

$$\Sigma_n^{(t+1)} = \frac{\sum_{m=1}^{M} \left(\mathbf{x}_m - \vec{\mu}_n^{(t+1)}\right)\left(\mathbf{x}_m - \vec{\mu}_n^{(t+1)}\right)^T p^{(t)}\left(n|m\right)}{\sum_{m=1}^{M} p^{(t)}\left(n|m\right)} \tag{16}$$

Cet algorithme utilise la variable cachée $\alpha_n^{(t+1)}$ relative à la probabilité de x_m d'appartenir à la gaussienne n.
Une silhouette est donc représentée par les paramètres $(\alpha_i, \vec{\mu}_i, \Sigma_i)$ des $K \times N$ gaussiennes.
Pour mesurer la dissimilarité entre deux silhouettes S et S^*, nous reconstruisons d'abord, pour chaque bande, la fonction de densité du mélange de N gaussiennes. La distance est ensuite définie par :

$$d\left(S, S^*\right) = \sum_{k=1}^{K} \sum_{\mathbf{x}} \left(\mathrm{f}_k\left(\mathbf{x}\right) - \mathrm{f}_k^*\left(\mathbf{x}\right)\right)^2 \tag{17}$$

où $\mathrm{f}_k\left(\mathbf{x}\right)$ est la fonction de densité du mélange de gaussiennes de la bande k.

5 Mesure de similarité entre passages

Dans la section précédente, nous avons présenté la signature colorimétrique invariante aux conditions d'éclairage, estimée sur chaque région clé. Cet ensemble de signatures, qui contient toutes les informations apparues au cours du

passage d'un individu devant une caméra, constitue les données d'entrée du processus de ré-identification. Dans cette section, nous présentons une approche basée sur la théorie des marches aléatoires sur un graphe permettant de faciliter l'évaluation de la similarité entre deux passages et de prendre la décision finale de ré-identification.

5.1 Rappels sur la théorie des graphes et des marches aléatoires

L'ensemble des signatures $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ extraites de m images appartenant à deux séquences est associé à un graphe de voisinage complet $G = (V, E)$ où chaque point de données S_i correspond à un sommet v_i du graphe. Deux sommets correspondant aux données S_i et S_j sont connectés par une arête qui est pondérée par la similarité W_{ij} entre les deux points de données. La similarité W_{ij} s'exprime à l'aide d'un noyau gaussien $W_{ij} = exp\left(-\frac{d(S_i,S_j)^2}{\sigma^2}\right)$, où $d(S_i, S_j)$ est la distance entre deux signatures et le paramètre σ est fixé à $\sigma = mean\left[d(S_i, S_j)\right]$, $\forall i, j = 1, \ldots, m$ $(i \neq j)$. La matrice $\mathbf{W} = [W_{ij}]_{i,j=1,\ldots,m}$ est appelée "matrice de similarité". Notons $\mathbf{D}$ la matrice diagonale composée des éléments $D_{ii} = \sum_j W_{ij}$ où D_{ii} est le degré du sommet $v_i \in V$. La probabilité de transition de sauter en une itération du sommet i au sommet j est donnée par $P_{ij} = W_{ij}/D_{ii}$. La matrice de transition $\mathbf{P} = [P_{ij}]_{i,j=1,\ldots,m}$ de la marche aléatoire est donc définie par :

$$\mathbf{P} = \mathbf{D^{-1}W} \tag{18}$$

L'ensemble des valeurs propres de $\mathbf{P}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, est usuellement appelé le spectre de $\mathbf{P}$ (ou le spectre du graphe associé G). Comme $\mathbf{P}$ est une matrice stochastique, et que $P_{ij} > 0$, $\forall i, j = 1, ..., m$, sa première valeur propre λ_1 est 1 avec le vecteur propre correspondant $\gamma_1 = [1, 1, ..., 1]^T \Big/ \sqrt{m}$. Dans le cadre de l'analyse spectrale pour la réduction de dimension, les q premiers vecteurs propres $\{\gamma_1, \gamma_2, \ldots, \gamma_q\}$ sont utilisés pour créer le nouveau système de coordonnées des points de données. On peut alors définir un opérateur de réduction de dimension $h : S_i \rightarrow u_i = [\gamma_1(i), \ldots, \gamma_q(i)]$ où $\gamma_k(i)$ est la ième coordonnée du vecteur propre γ_k.

Revenons à notre application, l'ensemble des signatures appartenant à deux passages est utilisé pour définir une matrice de transition $\mathbf{P}$ de la marche aléatoire sur le graphe G. L'ensemble des données est composé de deux classes connues, une pour chaque passage. Le problème de ré-identification consiste à évaluer si ces classes sont séparées ou non, ou, en d'autres termes, à mesurer la similarité entre les deux passages.

Dans le suivant, nous présentons une solution à ce problème utilisant les propriétés des valeurs propres et vecteurs propres de la matrice de transition $\mathbf{P}$.

5.2 L'analyse spectrale pour mesurer la similarité des séquences vidéo

Considérons tout d'abord le cas idéal où les points appartenant à une classe sont infiniment loin des points appartenant à l'autre classe. Supposons aussi que les points de données $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ sont ordonnés en fonction de la classe à qui ils appartiennent (les premiers points appartiennent à la classe a et les autres à la classe b). Comme les deux classes sont infiniment éloignées, la matrice de similarité est une matrice diagonale par blocs $\hat{\mathbf{W}} = \begin{bmatrix} \mathbf{W}^{(a)} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^{(b)} \end{bmatrix}$ où $\mathbf{W}^{(a)}$ et $\mathbf{W}^{(b)}$ sont les matrices de similarité intra-classe des classes a et b respectivement. La matrice de transition $\hat{\mathbf{P}}$ est aussi une matrice diagonale par blocs. Ses valeurs propres et vecteurs propres sont composés de l'union des valeurs propres et vecteurs propres de chaque bloc. Les deux premières valeurs propres de $\hat{\mathbf{P}}$ valent donc 1 et le matrice contenant les deux vecteurs propres correspondant est de la forme $\hat{\Upsilon} = \begin{bmatrix} \gamma_1^{(a)} & \vec{0} \\ \vec{0} & \gamma_1^{(b)} \end{bmatrix}$ où $\gamma_1^{(a)}$ et $\gamma_1^{(b)}$ sont des vecteurs constants. Les points $\hat{u}_i$, qui sont définis par la ième ligne de $\hat{\Upsilon}$, sont identiques pour tous les points S_i appartenant à la même classe ($\hat{u}_i = \begin{bmatrix} c^{(a)} & 0 \end{bmatrix}$ pour la classe a et $\hat{u}_i = \begin{bmatrix} 0 & c^{(b)} \end{bmatrix}$ pour la classe b). La classification des points $\hat{u}_i$ mène donc à deux classes qui correspondent aux deux vraies classes des données initiales.

De manière générale, les blocs hors diagonale de $\mathbf{W}$ et $\mathbf{P}$ ne sont pas nuls, i.e. les similarités inter-classes ne sont pas rigoureusement nulles. La différence $\mathbf{E} = \mathbf{P} - \hat{\mathbf{P}}$ est considérée comme une perturbation. La théorie sur les matrices de perturbation [13] montre que la stabilité des vecteurs propres de la matrice est déterminée par "les eigengaps" : l'écart entre deux valeurs propres successives. De manière plus formelle, les k premiers vecteurs propres de $\hat{\mathbf{P}}$ vont être stables aux perturbations de $\hat{\mathbf{P}}$ si et seulement si l'écart entre les valeurs propres $\xi_k = |\lambda_k - \lambda_{k+1}|$ est grand.

Appliquons cette propriété à notre problème : la matrice de transition $\mathbf{P}$ est généralement composée de blocs hors diagonal qui ne sont pas nuls et qui sont considérés comme des perturbations. Si deux passages sont plus similaires, les perturbations sont plus grandes. Les points X_i appartenant à ces deux séquences similaires ne sont pas bien séparés. Ceci se traduit par des points u_i qui ne sont pas bien séparés non plus. Les deux premiers vecteurs propres de $\mathbf{P}$ ne sont alors pas proches des vecteurs propres idéaux. Ils sont instables aux changements de $\mathbf{P}$ et la différence entre les valeurs propres $\xi_2 = |\lambda_2 - \lambda_3|$ est petite.

Si les deux séquences sont très différentes, la matrice de transition $\mathbf{P}$ est presque idéale avec des blocs hors diagonale proche de 0. Les points u_i correspondent aux points idéaux à une petite erreur près. Dans ce cas, les vecteurs propres sont stables et l'eigengap correspondant est grand.

Pour conclure, en construisant une marche aléatoire sur le graphe et en calculant le spectre du graphe associé, il est

possible de mesurer la similarité des séquences à partir du second eigengap $\xi_2 = |\lambda_2 - \lambda_3|$. Cette distance va servir à comparer les vidéos et à prendre la décision finale de ré-identification.

6 Résultats expérimentaux

Ce système a été entièrement évalué sur deux bases de données correspondant aux passages de plusieurs personnes devant des caméras disposées à différents endroits (Figure 5). La première base de données a été acquise dans les locaux de l'INRETS et contient des séquences vidéo de 40 personnes filmés à l'intérieur d'un bâtiment, avec une lumière naturelle à proximité de surfaces vitrées et à l'extérieur. La deuxième base acquise à l'intérieur d'un train en mouvement contient des séquences vidéo de 35 personnes filmées à deux endroits différents du train [14]. Cette base de données est très difficile, car l'acquisition de la vidéo est influencée par de nombreux facteurs tels que des variations rapides d'éclairage dus au déplacement du train, des changements irréguliers du fond, des reflets sur les vitres, des vibrations,... La figure 5 illustre ces deux bases de données.

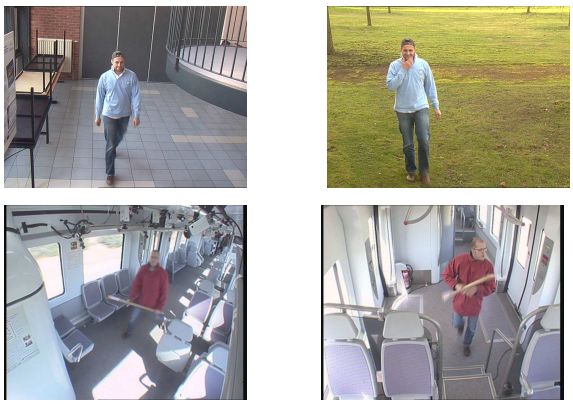


FIGURE 5 – Illustration de deux bases de données (Première ligne : base INRETS, Seconde ligne : base du train).

Les expérimentations ont été menées en deux temps. Dans un premier temps, nous évaluons l'approche de ré-identification de personnes indépendamment de la détection. Pour cela, les silhouettes ont été extraites manuellement dans toutes les séquences. Ceci permet de valider la signature proposée, comparativement aux autres signatures de la littérature ainsi que l'étape de ré-identification par analyse spectrale. La deuxième série d'expériences analysera l'influence de la détection automatique des silhouettes sur les résultats de ré-identification. Dans tous les cas, les taux de ré-identification présentés sont estimés à partir du seuil optimal tel que le taux de vraie ré-identification est égal au taux de vraie distinction.
Les résultats obtenus sur les deux bases de données, pour différentes signatures couplées à plusieurs invariants, sont résumés dans les tableaux 1 et 2. Nous constatons que le descripteur proposé amène aux meilleurs taux de ré-identification : 97.5% pour la base INRETS et 97.1% pour la base train. En comparant les signatures, nous constatons que les taux obtenus par l'histogramme sont toujours les plus faibles. Ceci montre le rôle important des informations spatiales dans le descripteur couleur. Notons aussi que l'introduction d'une étape de normalisation permettant d'obtenir une quasi-invariance aux conditions d'éclairage améliore considérablement les taux de ré-identification.

TABLE 1 – Performance des différentes signatures obtenus pour la base INRETS.

	RGB	Greyworld	RGB-rang	Affine
Histogrammes	77.5	82.5	87.5	90
Spatiogrammes	80	95	92.5	95
Couleur /Path-length	85	92.5	95	95
Descripteur proposé	87.5	**97.5**	95	**97.5**

TABLE 2 – Performance des différentes signatures obtenus pour la base du train.

	RGB	Greyworld	RGB-rang	Affine
Histogrammes	68.6	88.6	88.6	88.6
Spatiogrammes	77.1	94.3	94.3	91.4
Couleur /Path-length	82.9	91.4	94.3	94.3
Descripteur proposé	85.7	**97.1**	94.3	**97.1**

La deuxième série de tests permet d'évaluer l'étape d'extraction automatique de silhouettes. Pour cela, les performances de ré-identification avec une détection manuelle ou une détection automatique sont comparées. Les résultats obtenus sur les deux bases de données sont présentés dans le tableau 3. Pour la base INRETS, l'extraction automatique de silhouettes ne dégrade pas les performances de ré-identification, ce qui montre la robustesse de l'approche proposée sur cette base. Pour la base du train, les taux de ré-identification sont un peu dégradés par rapport à ceux obtenus par une extraction des silhouettes manuelle. Ceci est lié aux conditions réelles très difficiles de cette base : changements irréguliers du fond, variations lumineuses locales et globales, lentes et rapides, forte ressemblance entre les couleurs de vêtements des individus et du fond... Le meilleur taux de 88.6% obtenu par un processus complètement automatique est encourageant compte tenu de la difficulté de cette dernière base.

7 Conclusion

Nous avons présenté dans cet article un système complet de vidéo-surveillance permettant la ré-identification de personnes lorsqu'elles se déplacent à travers un réseau de ca-

TABLE 3 – Performance de la ré-identification avec une détection automatique.

	INRETS		Train	
	Manu.	Auto.	Manu.	Auto.
RGB	87.5	85	85.7	82.9
Greyworld	97.5	97.5	97.1	85.7
RGB-rang	95	95	94.3	82.9
Affine	97.5	97.5	97.1	88.6

méras. Celui-ci intègre toute la chaîne de traitements, allant des traitements bas niveau (détection) à des traitements de plus haut niveau comme la reconnaissance. Afin d'obtenir une détection robuste des objets en mouvement dans la séquence, nous avons mis en place une fusion de deux méthodes complémentaires : une détection par modélisation du fond par une mixture de gaussiennes et une détection par différence d'images successives. Ceci permet de tirer avantage de chacune d'elles : avoir des silhouettes de bonne qualité d'une part et s'adapter rapidement à des brusques changements d'éclairage d'autre part. Afin de décider si les régions extraites correspondent à la silhouette ou non, un processus de haut niveau, spécifique à cette application, a été introduit. En effet, les difficultés des séquences étudiées rendent l'utilisation des algorithmes de suivi impossible. Ainsi, nous avons proposé d'étudier la stabilité des régions de manière globale, sur toute la séquence correspondant à un passage. Une région extraite sera considérée comme une région clé (correspondant à la silhouette) si elle vérifie un double critère de stabilité spatiale et temporelle. Afin de caractériser chaque région clé, une signature robuste, quasi invariante aux conditions d'éclairage, intègrant à la fois des caractéristiques spatiales et des caractéristiques colorimétriques a été proposée. La similarité entres passages est ensuite estimée grâce à une approche utilisant la théorie des marches aléatoires sur un graphe construit à partir des signatures des régions clés.

Ce système complet a été testé sur deux bases de données obtenues dans des conditions différentes. La première a été acquise en laboratoire et représente des personnes observées dans deux lieux très distincts : dans un hall d'entrée et à l'extérieur. La deuxième base de données représente des personnes qui se déplacent dans un train en mouvement. Des variations d'éclairage très importantes apparaissent, rendant les étapes de détection et de caractérisation colorimétrique très délicates. Malgré ces difficultés, les taux de ré-identification obtenus par un processus complètement automatique sont très satisfaisants : 97.5% sur la première base et 88.6% sur la seconde.

Ces travaux ne demandent qu'à être améliorés en utilisant toutes les autres informations disponibles par le système : temps de transition entre caméras, direction des déplacements,... D'autre part, des scénarii plus complexes, où plusieurs personnes se déplacent dans le champ de vue des caméras, sont maintenant à l'étude.

Références

[1] C. Nakajima, M. Pontil, M. Heisele, and T. Poggio. Full body person recognition system. *Pattern Recognition*, 36(9) :1997–2006, 2003.

[2] N. Gheissari, T.B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, Washington, DC, USA, 2006.

[3] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[4] Y. Yu, D. Harwood, K. Yoon, and L.S. Davis. Human appearance modeling for matching across video sequences. *Machine Vision and Applications*, 18(3) :139–149, 2007.

[5] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 246–252, 1999.

[6] F.M. Porikli and O. Tuzel. Human body tracking by adaptive background models and mean-shift analysis. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2003.

[7] D-N. Truong Cong, L. Khoudour, C. Achard, and P. Phothisane. People re-identification by means of a camera network using a graph-based approach. In *Proceeding of IAPR Conference on Machine Vision Applications*, 2009.

[8] G. Buchsbaum. A spatial processor model for object color perception. *Journal of the Franklin Institute*, 310(1) :1–26, 1980.

[9] G.D. Finlayson, S. Hordley, G. Schaefer, and G. Yun Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38(2) :179–190, 2005.

[10] P. Gros, G. Mclean, R. Delon, R. Mohr, C. Schmid, and G. Mistler. Utilisation de la couleur pour l'appariement et l'indexation d'images. Technical report, Institut national de recherche en informatique et en automatique, 1997.

[11] S.T. Birchfield and S. Rangarajan. Spatiograms versus histograms for region-based tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2 :1158–1163, 2005.

[12] H. Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, pages 174–194, 1958.

[13] G.W. Stewart and J. Sun. *Matrix perturbation theory*. Academic Press, 1990.

[14] http ://www.multitel.be/boss