Réidentification des personnes à travers un réseau de caméras

Dung Nghi Truong Cong[1], Catherine Achard[2], Louahdi Khoudour[1]

1. *Institut National de Recherche sur les Transports et leur Sécurité, LEOST 20 rue Elisée Reclus, F-59650 Villeneuve d'Ascq, France*
   {truong,louahdi.khoudour}@inrets.fr
2. *Institut des Systèmes Intelligents et de Robotique, Université Paris 6/CNRS UMR 7222, BC 173, 4 place Jussieu, F-75005 Paris, France*
   catherine.achard@upmc.fr

RÉSUMÉ. *Les travaux de recherche présentés dans cet article ont pour objet le développement d'un système automatique multicaméra de détection/réidentification de personnes évoluant dans différents sites surveillés. Nous proposons dans un premier temps un processus automatique d'extraction de l'individu du fond permettant d'avoir des silhouettes de bonne qualité. Par la suite, un nouveau descripteur est proposé afin de caractériser les silhouettes de plusieurs personnes. Une approche basée sur l'analyse spectrale de graphes de similarités est proposée pour réaliser la réidentification. Deux bases de données représentant le passage de plusieurs personnes devant des caméras ont été utilisées pour l'évaluation des algorithmes. Les résultats expérimentaux très satisfaisants permettent de valider la robustesse de l'approche proposée.*

ABSTRACT. *The research addressed in this paper consists in developing an automatic system for detecting/reidentifying moving people in different sites with non-overlapping views. Firstly, we propose an automatic process for silhouette extraction. Then, a new colour descriptor is proposed for characterizing the silhouettes of the detected people. Once all the appearance information of each passage is extracted, a graph-based algorithm is used to realize the comparison of passages of people in front of cameras and to make the final decision of reidentification. The global system is 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.*

MOTS-CLÉS : *vidéosurveillance, détection de personnes, réidentification de personne, signature colorimétrique, invariant couleur, réduction de dimension.*

KEYWORDS: *surveillance system, person detection, colour-based descriptor, person reidentification, colour invariant, dimension reduction.*

DOI:10.3166/TS.27.297-324

Extended abstract

In recent years, public security has been facing an increasing demand from the general public as well as from governments. An important part of the efforts to prevent the threats to security is the ever-increasing use of video surveillance cameras throughout the network, in order to monitor and detect incidents without delay. Existing surveillance systems rely on human observation of video streams for high-level classification and recognition. However, a large number of cameras makes this solution inefficient and, in many cases, unfeasible. Although the basic imaging technologies for simple surveillance are available today, their reliable deployment in a large network is still ongoing research.

The research presented in this paper is within the framework of BOSS European project (on BOard wireless Secured video Surveillance) which aims at developing a multi-camera vision system specified to monitor, detect and recognize abnormal events occurring on-board trains. We tackle the problem of people detection and re-identification in the extremely complex environments inside moving trains. The video sequences capturing moving people are analyzed in order to re-establish a match of the same person over different camera views located at different physical sites. In most cases, such a system relies on building an appearance-based model that depends on several factors, such as illumination conditions, different camera angles and pose changes.

The proposed system for automatically detecting and re-identifying people on-board trains consists of three main functions :

(1) Silhouette extraction: This first task is very challenging due to many factors, such as changes in illuminations, shadows and camera imaging noise, which contribute to extreme adverse effects on the performance. In order to obtain pertinent results of moving object extraction, we propose to combine two main techniques: an adaptive background subtraction algorithm and a motion detection module. This combination allows us to take advantage of both methods by extracting the moving objects with highly precise silhouettes on the one hand and by managing important lighting changes on the other hand. A classification module using the double criterion based on spatial and colorimetric coherence is also carried out in order to select, among all the regions resulting from the detection, only those that really represent silhouettes.

(2) People characterization: Once the silhouettes of people are extracted, they are characterized by a new descriptor, called "statistical color-position" signature, which includes both color and spatial features of object (silhouette in our case). The main principle of this descriptor consists in dividing the detected silhouettes in equal horizontal bounds and then, extracting the color statistical features of the pixels inside each bound using the kernel density estimation (Parzen, 1962). Fixing the number of bounds rather than their size allows us to obtain an invariance towards the scale of the person. Since the color acquired by cameras is heavily dependent on several factors, such as surface reflectance, illuminant color and lighting geometry, a color normalization procedure (Gevers *et al.*, 2004, Finlayson *et al.*, 2005) is also carried out in order to obtain invariant signatures.

(3) People re-identification: In order to characterize all the appearance information of each passage in a video sequence and to deal with the big quantity of data, we propose a graph-based approach that represents a set of signatures in an embedded non-linear manifold without losing the original information (Nadler *et al*., 2005, von Luxburg, 2007). This technique preserves the local proximity among data points by first constructing a graph representation for the underlying manifold with vertices and edges. The vertices represent the data points, and the edges connecting the vertices represent the similarities between adjacent nodes. This procedure helps us to obtain a robust distance between passages of people in front of cameras and to make the final decision of re-identification.

The proposed system is evaluated by using two real datasets. The first dataset collected on INRETS premises contains video sequences of 40 people captured in two different locations (indoors in a hall near windows and outdoors with natural light). The second dataset containing sequences of 35 people is collected by two cameras installed on-board a train at two different locations: one in the corridor and one in the cabin. This dataset is more difficult than the first one, since these two cameras are set up with different angles and the acquisition of the video is influenced by many factors, such as fast illumination variations, reflections, vibrations. The experimental results show that the proposed system leads to satisfactory results: 97.5% for the true re-identification rate for the first dataset and 91.4% for the second.

Since the re-identification problem has gained in interest in the last few years and since there is no common database for evaluation, a comparative study of similar works in literature for people re-identification is hard to carry out. Thus, in this article, we only compare the performance of our system with the others in literature based on the global re-identification rate. For instance, Gheissari *et al.* (Gheissari *et al*., 2006) have used a dataset with 44 people and have obtained a 60% true re-identification rate. Wang *et al.* (Wang *et al.*, 2007) achieve a matching rate of 82% on a dataset containing 99 individuals. In the system proposed by Yu *et al.* (Yu *et al.,* 2007*)*, the accuracy is 95% for a 30 people dataset captured indoors. Nakajima *et al.* (Nakajima *et al.,* 2003) have obtained a 100% accuracy but on a very reduced dataset (4 people). One can notice that our approach leads to better results if we take into account the large and diverse dataset used.

Furthermore, in order to make available to researchers a common database for comparison, the dataset captured by multiple cameras installed on board a moving train in the framework of the BOSS project can be found on the website (BOSS). We hope that it will be a reference database for evaluating similar works concerning people re-identification problem.

1. Introduction

De nos jours, il ne fait nul doute que la sécurité soit un souci majeur du grand public et des gouvernements. Chaque pays a mis en œuvre des mesures pour la renforcer en fonction de la spécificité des problèmes posés, des conditions locales et des traditions culturelles. Le but à atteindre est une détection automatique et rapide d'événements amenant à une atteinte à la sécurité des biens ou des personnes tels que le vandalisme, les actes de vols ou de terrorisme, le feu... Une partie importante des efforts mis en place pour empêcher ces menaces à la sécurité consiste à installer des caméras de surveillance dans des lieux publics, afin de surveiller et détecter des incidents. Cependant, une large couverture télévisuelle apporte de nouvelles difficultés pour gérer les grands volumes d'informations que produisent de tels systèmes. Ainsi, il n'est pas rare de voir une personne travaillant au service de sécurité contrôler en même temps les flux provenant de 20 à 40 caméras. Outre le côté fastidieux de ce travail, l'accroissement constant du nombre de caméras amène au développement de nouvelles technologies et fonctionnalités intelligentes afin de disposer de systèmes autonomes et plus performants pour une gestion efficace et préventive de la sécurité.

Les travaux de recherche présentés dans cet article contribuent au développement d'un système automatique multicaméras de détection/réidentification de personnes se déplaçant dans différents lieux. Dès qu'une personne passe sous le champ d'une des caméras d'un réseau, on souhaite être capable de la ré-identifier lorsqu'elle est vue par une autre caméra. Ces travaux s'inscrivent dans le cadre du projet européen BOSS (*on BOard wireless Secured video Surveillance*) qui consiste en particulier à mettre au point un système de vidéosurveillance à bord permettant de suivre le déplacement d'un individu lorsqu'il se présente devant les différentes caméras d'un train en marche. Néanmoins, la signature mise en place dans ces travaux est assez générique pour que le système puisse être tout aussi bien utilisé dans d'autres contextes, à l'intérieur de bâtiments par exemple, de stations de métro, de gares, ou à l'extérieur. La seule contrainte que nous nous sommes fixés dans un premier temps est que les personnes passent seules sous le champ de vue des caméras : nous ne gérons pas encore le cas multi-usager, ce travail représente une perspective à court terme. Le système proposé, utilisant une modélisation colorimétrique et spatiale de l'apparence des personnes, doit être robuste face aux conditions réelles extrêmement difficiles dans notre cas, telles que des changements brusques de luminosité, des points de vue différents pour chaque caméra, des variations de la pose des personnes, des vibrations liées au déplacement du train...

Plusieurs approches ont été proposées dans la littérature pour la réidentification de personnes basée sur l'apparence. Kettnaker et Zabih (Kettnaker *et al.*, 1999) exploitent conjointement la similarité des vues de personnes et la « plausibilité » du temps de déplacement d'une caméra à l'autre. Dans les travaux de Nakajima *et al.* (2003), des descripteurs colorimétriques sont utilisés comme signature puis introduits dans un algorithme de reconnaissance à base de SVM (*Support Vector Machines*) multi-classes. Javed *et al.* (2005) utilisent diverses caractéristiques spatio-temporelles (lieux d'entrée/sortie, temps de déplacement, vitesse...) combinées à un histogramme couleur afin de caractériser le passage d'un individu. Un cadre probabiliste est ensuite développé pour la réidentification. Gheissari *et al.* (2006)

proposent d'extraire une signature temporelle invariante à la position du corps et à l'apparence dynamique des vêtements d'une part, et d'utiliser un descripteur plus local basé sur les points d'intérêt d'autre part. Les travaux de Wang *et al.* (2007) 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. Récemment, Yu *et al.* (2007) ont introduit un nouveau descripteur utilisant à la fois des informations colorimétriques et spatiales. Ce dernier est utilisé pour sélectionner des images clés, puis pour comparer des séquences vidéo.

Dans cet article, nous proposons un système automatique de détection/réidentification de personnes dont le synoptique est présenté sur la figure 1. Il se décompose en trois parties principales :

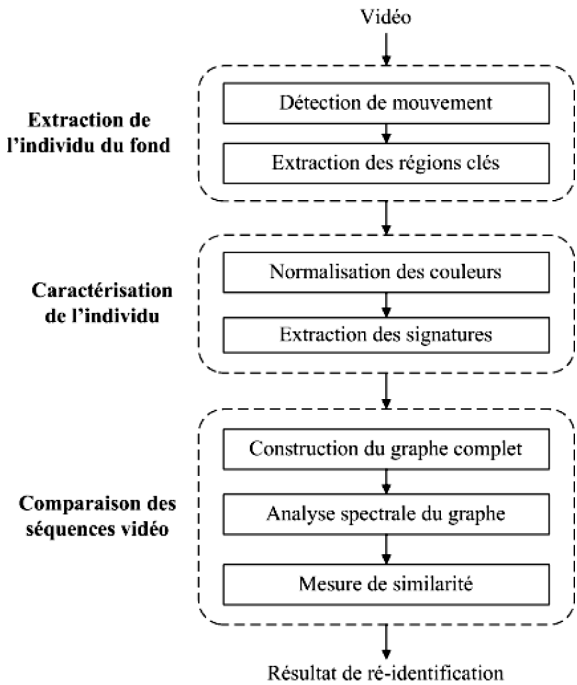


Figure 1. Synoptique du système de détection/réidentification

L'extraction de l'individu du fond. Cette première phase, qui a été conçue spécifiquement pour cette application, a pour but d'extraire les silhouettes des personnes dans les séquences vidéo. Afin d'obtenir des régions de bonne qualité, une double étape de détection est proposée. Elle combine les résultats d'une détection de mouvement par modélisation du fond avec une mixture de gaussiennes et ceux obtenus par la différence d'images successives. Cette combinaison permet de tirer avantage des deux méthodes en obtenant des silhouettes pleines d'une part et en gérant les forts changements d'éclairage d'autre part. Par la suite, nous avons introduit une étape de sélection de régions afin de ne conserver que les régions

correspondant aux personnes en mouvement (nommées régions clés par la suite). Celle-ci utilise à la fois des informations spatiales et colorimétriques.

La caractérisation de l'individu par sa signature colorimétrique. Afin de pouvoir comparer les séquences vidéo, il est nécessaire de réaliser un codage de chaque silhouette (région clé). Nous proposons à cette fin une nouvelle signature utilisant à la fois des informations spatiales et colorimétriques et apte à caractériser des objets déformables. Comme la couleur varie énormément en fonction de l'éclairage ambiant du lieu observé, une étape de normalisation des couleurs a été introduite de manière à parvenir à des signatures pseudo-invariantes.

La comparaison des séquences vidéo. Les deux étapes précédentes amènent à un ensemble de signatures caractérisant chaque passage de personne devant une caméra et il est maintenant nécessaire d'établir une mesure de similarité entre ces ensembles de signatures. Afin de réaliser l'étape de réidentification avec une simple distance euclidienne, sans être confronté aux problèmes des grandes dimensions, une étape de réduction de dimension utilisant le principe de l'analyse spectrale est mise en place.

L'organisation de l'article est la suivante : après cette introduction, nous présentons en section 2 l'approche proposée pour extraire les régions en mouvement. La sélection des résultats d'extraction de mouvement, qui permet de ne conserver que les régions correspondant à la personne, est détaillée en section 3. Dans la section 4, nous introduisons les signatures spatiales/colorimétriques utilisées pour caractériser les silhouettes détectées. L'utilisation de l'analyse spectrale pour réaliser la réidentification des personnes est ensuite explicitée dans la section 5, juste avant les résultats expérimentaux, section 6, et la conclusion, section 7.

2. Extraction des objets mobiles en environnement bruité

L'extraction d'objets en mouvement est un problème largement étudié en vision artificielle et non résolu à ce jour. Les algorithmes présents dans la littérature peuvent être partitionnés en deux grandes classes : (i) ceux qui réalisent la différence d'images successives. Cette approche, très simple, ne permet généralement d'obtenir que les contours des objets en mouvement, mais ne nécessite pas de maintenir à jour une image de référence. Elle s'adapte donc très rapidement aux variations de la scène observée ; (ii) ceux soustrayant l'image courante à une image de référence. Il s'agit de bâtir un modèle de l'arrière-plan de la scène, puis d'appliquer une fonction de décision permettant de classer chaque pixel comme appartenant à l'arrière-plan ou au premier plan. Cette technique conduit à un masque binaire des objets en mouvement de bonne qualité. Par contre, elle requiert l'utilisation d'une image de référence et s'adapte de fait beaucoup moins bien aux brusques changements d'éclairage. Le modèle de l'arrière-plan le plus simple est une moyenne temporelle des images. Cependant, un tel modèle devient vite obsolète, particulièrement en environnement bruité ou évolutif. De nombreuses approches ont été proposées dans la littérature afin de créer un modèle plus robuste. Il est ainsi possible de décrire l'arrière-plan de manière statistique en modélisant la distribution des intensités lumineuses de chaque pixel par une loi gaussienne (Wren *et al.*, 1997), un mélange de gaussiennes (Stauffer *et al.*, 1999) ou une estimation de probabilité non paramétrique (Elgammal *et al.*,

2000). Le modèle de l'arrière-plan peut également être construit en adoptant une technique de quantification/clustering de la distorsion de couleur et de luminosité (Kim *et al.*, 2004).

Même si les méthodes précitées présentent des avantages indéniables, elles amènent à des résultats de faible qualité lorsqu'elles sont utilisées dans des cas extrêmes comme l'intérieur d'un train en mouvement. En effet, des conditions réelles telles que des variations lumineuses locales et globales, lentes et rapides, des changements irréguliers du fond, une forte ressemblance entre les couleurs de vêtements des individus et du fond... rendent la tâche de l'extraction de l'individu du fond extrêmement difficile. De plus, dans cette application, comme les personnes se déplacent essentiellement face à la caméra, une mise à jour systématique du fond amène à inclure les personnes dans le fond et à perdre des parties significatives de leur silhouette. *A contrario*, si le fond n'est pas mis à jour régulièrement, une dégradation très rapide des résultats apparaît due aux forts changements de luminosité présents à l'intérieur du train. Par conséquent, il est nécessaire de développer, à partir des méthodes de détection d'objets en mouvement présentes dans la littérature, une nouvelle approche permettant de s'adapter aux conditions particulières de notre application. Ainsi, nous proposons de mixer les deux approches précitées : celle utilisant une différence d'images successives et celle utilisant une modélisation du fond. Ceci permet d'une part d'obtenir des silhouettes pleines et d'autre part, de s'adapter aux brusques changements de l'arrière-plan.

Le synopsis du processus proposé pour extraire les régions en mouvement est illustré sur la figure 2 où chacune des deux approches est retrouvée. Ainsi, la première étape consiste à extraire un masque binaire des pixels de l'avant-plan par la soustraction de l'arrière-plan et suppression des ombres. La deuxième étape procure une robustesse supplémentaire au processus en supprimant une grande partie du bruit grâce à la détection des régions en mouvement à partir de la différence de trois images successives. Le résultat final de détection est ensuite utilisé pour mettre à jour de manière adaptative les pixels de l'arrière-plan.

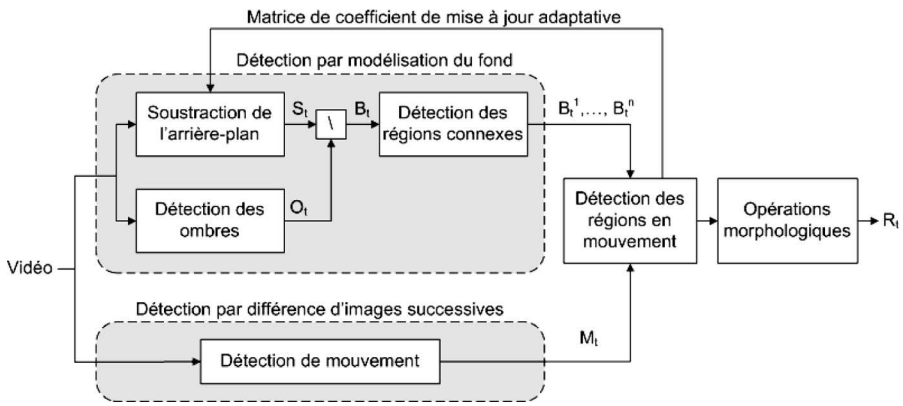


Figure 2. Synopsis du processus d'extraction de silhouettes

2.1. Détection des objets en mouvement avec une modélisation de l'arrière-plan par un mélange de gaussiennes

Cette technique de soustraction du fond (Stauffer *et al.*, 1999), largement utilisée dans la littérature, consiste à modéliser l'historique de chaque pixel de l'arrière-plan par un mélange de gaussiennes. Ainsi, à chaque pixel est associée une somme pondérée de N_g gaussiennes ($2 \leq N_g \leq 5$) où N_g peut être ajusté en fonction de la complexité de la scène observée. A un instant t, la probabilité d'observer la valeur courante du pixel $I^t(p)$ est estimée par :

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

où $\mathbf{I}_p = \left\{I^1(p), I^2(p), ..., I^t(p)\right\}$ est l'historique des valeurs prises par le pixel; w_i^t est le poids de la gaussienne i à l'instant t $\left(\sum_{i=1}^{N_g} w_i^t = 1\right)$; $\vec{u}_i^t$ et Σ_i^t sont la moyenne et la matrice de covariance de la gaussienne $\eta\left(I^t(p)\ ;\ \vec{u}_i^t, \Sigma_i^t\right)$. Supposons que les trois canaux couleur sont statistiquement indépendants, la matrice de covariance Σ_i^t peut être simplifiée par :

$$\Sigma_i^t = \begin{bmatrix} {\sigma_{i,r}^t}^2 & 0 & 0 \\ 0 & {\sigma_{i,v}^t}^2 & 0 \\ 0 & 0 & {\sigma_{i,b}^t}^2 \end{bmatrix} \quad [2]$$

La valeur courante du pixel $I^t(p)$ est associée à une des gaussiennes du mélange si la condition suivante est vérifiée :

$$\left\| I^t(p) - \vec{u}_i^t \right\| < K\sigma_i^t \quad [3]$$

où K est un paramètre empirique de la méthode.

Les paramètres des gaussiennes k qui valident la condition [3] sont mis à jour avec :

$$\begin{aligned} \vec{u}_k^t &\leftarrow \left(1 - \tfrac{\alpha_p}{w_k^t}\right)\vec{u}_k^{t-1} + \tfrac{\alpha_p}{w_k^t} I^t(p) \\ {\sigma_{k,c}^t}^2 &\leftarrow \left(1 - \tfrac{\alpha_p}{w_k^t}\right){\sigma_{k,c}^{t-1}}^2 + \tfrac{\alpha_p}{w_k^t}\left(I_c^t(p) - \mu_{k,c}^t\right)^2 \\ w_k^t &\leftarrow \left(1 - \alpha_p\right) w_k^{t-1} + \alpha_p \end{aligned} \quad [4]$$

où α_p est le coefficient de mise à jour du pixel p. Notons que ce coefficient de mise à jour varie d'un pixel à l'autre et est estimé à la fin de l'étape de détection de régions

en mouvement avec :

$$\alpha_p = \begin{cases} m & \text{si } p \text{ est détecté en mouvement} \\ M & \text{sinon} \end{cases} \quad [5]$$

où M et m ($m << M$) sont deux paramètres empiriques qui règlent le degré de mise à jour de l'image de fond. L'utilisation de deux valeurs (contrairement à ce qui a été proposé initialement par Stauffer *et al.* (1999)) permet aux pixels de l'arrière-plan correspondant à un objet en mouvement, de n'être que très faiblement mis à jour. Ceci permet d'éviter d'intégrer la silhouette de la personne dans l'image de fond lorsque celle-ci se déplace face à la caméra.

Pour les autres gaussiennes qui ne valident pas la condition [3], le poids est ajusté avec :

$$w_k^t \leftarrow \left(1 - \alpha_p\right) w_k^{t-1} \quad [6]$$

Si le pixel ne peut être associé à aucune des gaussiennes, la gaussienne ayant le plus petit poids est réinitialisée par la valeur du pixel actuel.

$$\begin{array}{l} \vec{u}_m^t \leftarrow I^t(p) \\ \Sigma_m^t \leftarrow \Sigma_{initiale} \end{array} \quad [7]$$

Afin de déterminer si le pixel actuel $I^t(p)$ appartient au premier plan, l'ensemble des gaussiennes est ordonné selon un ordre décroissant des valeurs de $\frac{w_k^t}{\sigma_k^t}$. Le modèle du fond est estimé par l'ensemble des gaussiennes qui ont à la fois un fort poids et une variance faible. Les C gaussiennes décrivant le fond sont alors sélectionnées par :

$$C = \text{argmin}_c \left(\sum_{i=1}^{c} w_i^t > S \right) \quad [8]$$

où S est un paramètre empirique.

Pour chaque pixel, il y a 2 possibilités : si le pixel $I^t(p)$ est associé à une des gaussiennes modélisant le fond, le pixel est considéré comme fond, sinon, il est détecté en mouvement.

La figure 3b illustre un exemple de résultat de détection d'objets en mouvement par modélisation du fond avec un mélange de gaussiennes. Pour cette séquence difficile, beaucoup de fausses détections apparaissent à cause d'un brusque changement de luminosité dû au soleil. La détection par modélisation du fond montre ses limites et une combinaison de cette méthode avec une autre technique est nécessaire. Un autre phénomène apparaît : les ombres sont détectées comme des objets en mouvement. Un traitement spécifique va donc être mis en place afin de les supprimer.

2.2. Détection des ombres

La détection des ombres est une étape importante dans le processus d'extraction d'objets en mouvement, car celles-ci amènent à des problèmes importants tels que la déformation de la forme des objets, la fusion de plusieurs objets entre eux, les mauvaises classifications... Plusieurs travaux ont été réalisés afin de détecter les zones d'ombres dans une image. Ils utilisent très souvent l'hypothèse que les pixels d'ombre ont la même couleur que les pixels de l'arrière-plan, avec une plus faible luminance. En effet, ceux-ci ne reçoivent pas directement le flux de lumière à cause d'obstacles.

Dans nos travaux, nous utilisons la méthode de détection de zones d'ombre proposée par Porikli et Tuzel (Porikli *et al.*, 2003). Nous comparons, pour un même pixel, les couleurs de l'image courante $I^t(p)$ et de l'image de référence $I_{ref}(p)$ définie par la gaussienne ayant le plus grand poids. La projection du vecteur couleur $I^t(p)$ sur le vecteur couleur de l'arrière-plan donne le changement de luminance :

$$h = \left\| I^t(p) \right\| \cos \Phi \qquad [9]$$

où Φ est l'angle entre $I^t(p)$ et $I_{ref}(p)$. Le rapport de luminance est alors défini par $r = \left\| I_{ref}(p) \right\| / h$. Un second angle Φ_B entre $I_{ref}(p)$ et l'axe des niveaux de gris est également estimé. Un pixel p, détecté en mouvement, est classé comme ombre s'il vérifie les deux conditions :

$$\Phi < \Phi_B \quad , \quad r_1 < r < r_2 \qquad [10]$$

où Φ_0 est l'angle maximum de séparation, r_1 et r_2 sont des paramètres représentant les éclaircissement et assombrissement maximaux autorisés.

La figure 3c illustre le résultat de détection des ombres. Un étiquetage en composantes connexes (Ballard *et al.*, 1982), réalisé sur la carte binaire de détection, permet de supprimer les petites régions et d'arriver aux régions $B_1^t, B_2^t, ..., B_n^t$ détectées en mouvement (figure 3d).

Les résultats d'extraction obtenus par la détection des objets en mouvement par modélisation du fond par un mélange de gaussiennes et suppression des ombres sont encore bruités. Une fusion de ces résultats avec une détection de mouvement par différence d'images successives va maintenant être mise en place afin de gérer les brusques variations de l'image du fond.

2.3. Détection des régions en mouvement

L'objectif de cette étape consiste à conserver les régions appartenant à la silhouette et à éliminer le bruit et les fausses détections. Une nouvelle détection de mouvement exploitant la différence entre trois images consécutives est donc réalisée. Cette technique présente l'avantage de ne pas nécessiter d'image de référence et s'adapte donc très rapidement aux brusques changements de l'arrière-plan. D'autre part, elle nécessite très peu de ressources, ce qui lui confère un atout supplémentaire.

Le masque binaire de détection de mouvement est défini par :

$$M^t(p) = \left(\frac{\left|I^t(p) - I^{t-1}(p) - \mu_1\right|}{\sigma_1} > S_M\right) \cup \left(\frac{\left|I^{t-1}(p) - I^{t-2}(p) - \mu_2\right|}{\sigma_2} > S_M\right) \quad [11]$$

où μ_1 et σ_1 sont la moyenne et l'écart type de $\left|I^t - I^{t-1}\right|$ et S_M est un paramètre fixant la différence de niveaux de gris maximale autorisée.

La figure 3e illustre le résultat de détection de mouvement où, comme nous l'avions mentionné dans l'introduction de ce chapitre, seuls les contours des objets en mouvement sont détectés.

Ce résultat de détection est alors utilisé pour rechercher, parmi les n régions $B_1^t, B_2^t, ..., B_n^t$ issues de la soustraction de fond, celles qui sont en mouvement et correspondent donc aux objets réellement mobiles de la scène (la silhouette dans notre application). Pour ce faire, le nombre de pixels en mouvement de chaque région B_k^t est estimé avec :

$$Nbm_k^t = \left\{p : p \in B_k^t, M^t(p) = 1\right\} \quad [12]$$

Un seuil est ensuite choisi et toutes les régions ayant un nombre de pixels en mouvement inférieur au seuil sont supprimées. La figure 3f illustre le résultat de cette étape. Notons $R^t = \left\{r_1^t, r_2^t, ..., r_N^t\right\}$ l'ensemble des N régions finalement conservées à l'instant t.

3. Sélection des résultats d'extraction

Avant de débuter le processus de réidentification, il est nécessaire de mettre en forme les résultats de détection et d'obtenir une caractérisation du passage d'une personne. Pour cela, il faut sélectionner, parmi toutes les régions r_i^t issues de la détection, celles qui représentent réellement la silhouette de la personne et constitueront les régions clés utilisées pour caractériser la séquence. Rappelons que dans le cas idéal, les régions clés finales représentent exactement les silhouettes des personnes filmées. Cette sélection des régions pertinentes va être réalisée en utilisant deux critères : un critère spatial dans un premier temps puis un critère colorimétrique.

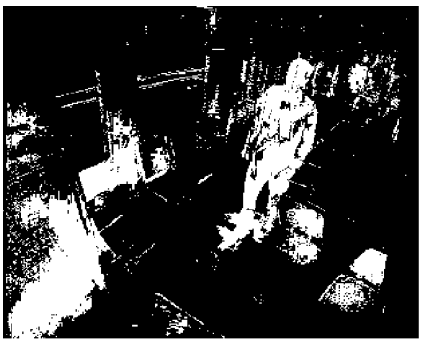
3.1. Regroupement des régions en fonction de leur cohérence spatiale

Cette première étape a pour but de regrouper temporellement les régions appartenant au même objet et utilise pour cela des informations spatiales : on peut raisonnablement penser que deux régions appartenant à des images consécutives, et dont le pourcentage de pixels communs (lorsque les deux régions sont projetées dans la même image) est grand appartiennent au même objet.

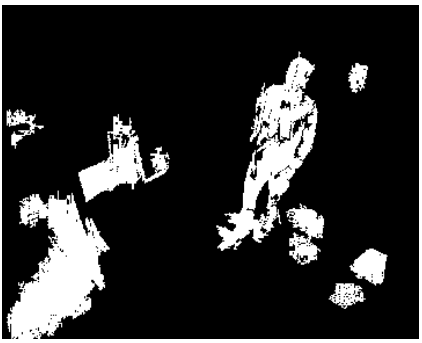
Soit $R = \left\{R^1, R^2, ..., R^T\right\}$, l'ensemble des régions détectées au cours d'une séquence de T images, avec $R^t = \left\{r_1^t, r_2^t, ..., r_N^t\right\}$. Les régions sont regroupées en classes selon l'algorithme suivant :

(a) Image originale

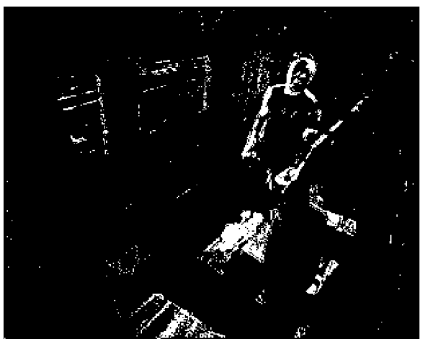
(b) Résultat de détection avec modélisation par mixture de gaussiennes



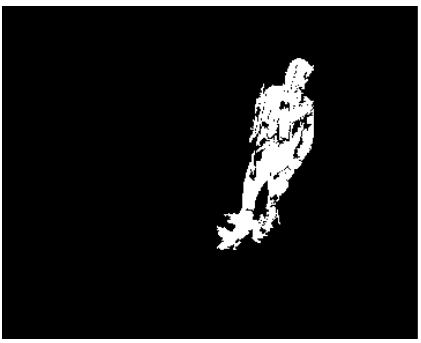
(c) Résultat de détection après la suppression des ombres



(d) Résultat de détection après la suppression de régions de trop petite taille



(e) Résultat de détection par différence entre trois images successives



(f) Résultat final de détection combinant les résultats des figures (d) et (e)

Figure 3. Les différentes étapes de la détection de régions en mouvement

Algorithme 1. Regroupement des régions en fonction de leur cohérence spatiale

Initialisation : $\forall\ j \in [1,N]\ ,\ C\left(r_j^1\right) \leftarrow\ j, t = 2$

tant que $t <= T$ **faire**

  pour *chaque region* r_j^t **faire**

    si $\exists\ i\ :\ \dfrac{\left|r_i^{t-1} \cap r_j^t\right|}{\max\left(\left|r_i^{t-1}\right|, \left|r_j^t\right|\right)} > S$ **alors** $C\left(r_j^t\right) = C\left(r_i^{t-1}\right)$

    sinon $C\left(r_j^t\right) =$ nouvelle classe

    $t = t + 1$

où $C\left(r_j^t\right)$ est la classe de la région r_j^t et $|r|$ est le nombre de pixel de la région r.

A la fin de cette étape, toutes les régions ont été affectées à une classe, et chaque classe correspond à un ensemble de régions spatio-temporellement stables. Comme les problèmes de segmentation ont tendance à rompre ce critère de cohérence spatio-temporelle (voir figure 4), plusieurs classes peuvent être générées au cours du temps pour les silhouettes correspondant à une même personne. Un second regroupement, utilisant un critère colorimétrique, est donc mis en place pour remédier à ce problème.

3.2. Regroupement des régions en fonction de leur caractéristique colorimétrique

Ce nouveau regroupement nécessite de caractériser chaque région r_j^t détectée par un descripteur colorimétrique. Nous avons opté pour un descripteur S_j qui allie des informations colorimétriques et spatiales et qui présente une bonne aptitude à décrire des régions (Truong Cong *et al.*, 2009). L'idée de ce descripteur consiste à découper horizontalement la région d'intérêt en P bandes équidistantes. Chacune de ces bandes est ensuite caractérisée grâce à la moyenne des couleurs des pixels lui appartenant.

L'objectif de ce deuxième regroupement est de fusionner les classes composées de régions ayant les mêmes caractéristiques colorimétriques. L'algorithme leader/suiveur utilisé est le suivant (Hartigan, 1975).

Algorithme 2. Regroupement en fonction de leur caractéristique colorimétrique

Initialisation : $G_1 \leftarrow C_1, n = 1$

pour *chaque* $C_k : j = \text{argmin}_i\ \{dist\ (C_k, G_i)\}$ **faire**

  si $dist\ \left(C_k, G_j\right) < S_C$ **alors** $G_j\ \leftarrow C_k$

  sinon créer un nouveau 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 :

$$dist\,(C_k,G_i) = \min\{dist\,(C_k,C_{ib})\,,b = 1...B\} \quad [13]$$

où C_{ib} est une classe appartenant au groupe G_i et B est le nombre de classes du groupe G_i. La distance entre classes $dist\,(C_k,C_{ib})$ est définie par :

$$dist\left(C_p,C_q\right) = \frac{1}{P \times Q} \sum_{i=1}^{P} \sum_{j=1}^{Q} d\left(S_{pi},S_{qj}\right) \quad [14]$$

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.

Afin d'illustrer le fonctionnement des deux classifications présentées, nous présentons, figure 4, certaines images d'une séquence et les résultats de classification correspondant. Notons que cette séquence est un cas très difficile de notre base de données. Intéressons-nous dans un premier temps aux classes Ci issues du regroupement fonction de la cohérence spatiale. Lors de l'image 1215, une première classe est créée. Elle correspond à la partie haute de la personne. A l'image 1224, un fort changement d'éclairage apparaît. Ceci se traduit par la détection d'une grande région correspondant au fond (classe 3) et par la détection de la personne en entier (classe 2). Comme les régions correspondant à la demi-personne et à la personne en entier ont un taux de recouvrement faible (tel que défini en sous-section 3.1), ces deux régions ne sont pas assignées à la même classe (classes 1 et 2). Le regroupement des silhouettes à la classe C2 continue jusqu'à l'image 1229. A l'image 1230, seul le haut de la personne est de nouveau détecté, il y a donc création d'une nouvelle classe, et ainsi de suite.

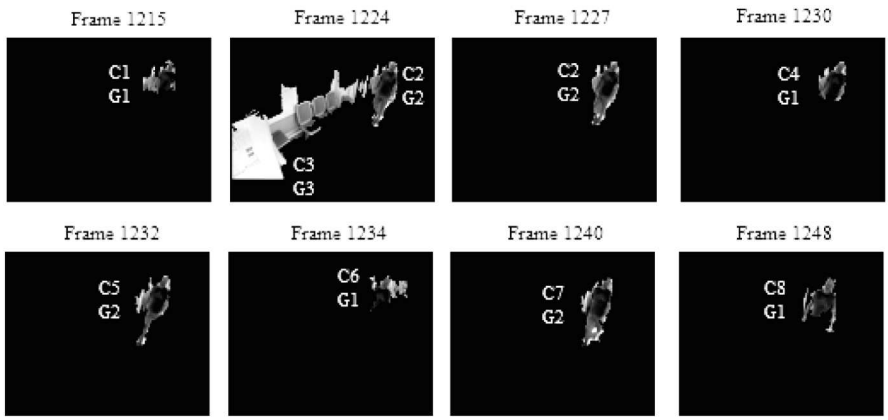


Figure 4. Résultat de classification des régions

Le regroupement des classes en groupe est ensuite basé sur un critère colorimétrique. Le groupe G1 composé des classes (C1, C4, C6, C8) correspond à de mauvais résultats de segmentation dus à une forte ressemblance entre le pantalon de la personne et le fond de l'image. Le groupe G3 résulte d'un fort changement de luminosité, tandis que le groupe G2 épouse relativement bien les silhouettes de la personne.

Nous supposons que l'ensemble le plus stable (composé du plus grand nombre de régions) correspond à la silhouette de la personne en mouvement et conservons les régions correspondantes pour la réidentification. Celles-ci, nommées régions clés, constituent l'entrée du processus de réidentification.

4. Caractérisation des régions clés

La réidentification des personnes passe maintenant par la caractérisation de chaque région clé. Il semble tout à fait logique pour cela d'utiliser des informations spatiales et colorimétriques. Comme l'objectif de notre travail consiste à développer un système de télésurveillance fonctionnant sur des données réelles, de nombreux problèmes apparaissent comme de forts changements d'éclairage entre les différents lieux. Par conséquent, il est indispensable de mettre en place une étape de normalisation permettant de s'affranchir de ces changements et d'obtenir des signatures colorimétriques pseudo-invariantes.

4.1. Invariants colorimétriques

De nombreux invariants couleur ont été proposés dans la littérature (Gevers *et al.*, 2004 ; Finlayson *et al.*, 2005 ; Madden *et al.*, 2007). Nous en avons retenu trois qui ont amené à de bons résultats pour notre problématique. Ils seront testés à tour de rôle dans la section correspondant aux résultats expérimentaux. Nous les détaillons ci-dessous pour la seule composante rouge (I_R). Les mêmes procédures de normalisation sont mises en place pour les plans vert et bleu :

– normalisation de Greyworld (Buchsbaum, 1980) :

$$I_R^* = \frac{I_R}{\texttt{moyenne}\,(I_R)} \qquad [15]$$

– normalisation affine :

$$I_R^* = \frac{I_R - \texttt{moyenne}\,(I_R)}{\texttt{ecart type}\,(I_R)} \qquad [16]$$

– RGB-rang : Finlayson *et al.*, (2005) supposent que les mesures de rang colorimétrique des pixels sont insensibles aux changements d'éclairage. La mesure de rang colorimétrique $I_R(g)$ du niveau g du plan R est par exemple obtenue par l'égalisation de l'histogramme monodimensionnel H_R :

$$I_R(g) = \sum_{u=0}^{g} H_R(u) / \sum_{u=0}^{Nb} H_R(u) \qquad [17]$$

où Nb représente le nombre de niveaux de quantification de la composante couleur.

4.2. Signature des régions

Une fois la normalisation couleur appliquée à l'intérieur de chaque région clé, plusieurs signatures colorimétriques peuvent être estimées pour caractériser ces régions. Le descripteur le plus utilisé dans la littérature à cause de sa simplicité est l'histogramme couleur (Javed *et al.*, 2003). 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 ces couleurs. De nombreuses tentatives ont été faites pour étendre le concept d'histogramme en introduisant des informations spatiales. Parmi celles-ci, nous pouvons citer le spatiogramme (Birchfield *et al.*, 2005), ou le descripteur couleur/path-length (Yoon *et al.*, 2006).

Dans cet article, nous proposons un nouveau descripteur intitulé « couleur-position statistique ». La silhouette extraite est découpée dans un premier temps en P bandes équidistantes. Fixer le nombre de bandes plutôt que leur taille permet d'obtenir une invariance face à l'échelle de la personne. La distribution couleur de chaque bande est ensuite caractérisée grâce à une estimation par noyau (Parzen, 1962) aussi appelée méthode de Parzen-Rozenblatt.

Etant donné Nb pixels $X = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_{Nb}\}$ de la bande n de la région clé, où x_i est un vecteur de dimension 3 correspondant aux 3 composantes couleur, la densité de probabilité de la composante couleur j est estimée par :

$$f_{nj}(x) = \frac{1}{L} \sum_{i=1}^{Nb} \frac{1}{h} K\left(\frac{x - x_{ij}}{h}\right) \qquad [18]$$

où K est la densité gaussienne et h est le paramètre de lissage. Une région clé est donc caractérisée par $P \times d$ fonctions de densité de probabilité f_{nj}.

Pour mesurer la distance entre régions, nous utilisons la divergence de Kullback-Leibler (Kullback *et al.*, 1951), qui est connue comme une mesure de similarité entre distributions de probabilités. La distance entre deux régions clés r et r' est alors définie par :

$$d(r, r') = \sum_{n=1}^{P} \sum_{j=1}^{3} \left| d^*_{KL}\left(f_{nj}, f'_{nj}\right) \right| \qquad [19]$$

où $d^*_{KL}\left(f_{nj}, f'_{nj}\right)$ est la version symétrique de la distance de Kullback-Leibler entre deux distributions de probabilités discrètes f_{nj} et f'_{nj} définie par :

$$d^*_{KL}\left(f_{nj}, f'_{nj}\right) = \left[d_{KL}\left(f_{nj} || f'_{nj}\right) + d_{KL}\left(f'_{nj} || f_{nj}\right) \right] / 2 \qquad [20]$$

où $d_{KL}\left(f_{nj} || f'_{nj}\right) = \sum f_{nj} \log \frac{f_{nj}}{f'_{nj}}$.

Chaque séquence relative au passage d'une personne est maintenant décrite par un ensemble de descripteurs (autant de descripteurs que de régions clés issues du processus de détection) et il est utile de déterminer une mesure entre ces ensembles afin de pouvoir comparer les séquences entre elles.

5. Mesure de distance entre séquences vidéo

Comme les données extraites des séquences vidéo présentent une forte redondance et que le descripteur associé à chaque région est de grande dimension[1], une étape de réduction de la dimension par une méthode non linéaire est mise en place dans un premier temps. La réduction de dimension est un procédé important utilisé dans divers problèmes d'analyse de données. Elle est réalisée en ne gardant que les dimensions les plus importantes, celles qui portent la majorité de l'information.

Les techniques permettant de réduire la dimension d'un problème que l'on trouve dans la littérature se découpent en deux approches principales : l'approche linéaire avec par exemple l'analyse en composantes principales (Hotelling, 1933) et l'analyse discriminate (Fisher, 1936)) et l'approche non linéaire avec les Laplacian eigenmaps (Belkin *et al.*, 2003), diffusion maps (Nadler *et al.*, 2005) et plusieurs variantes de l'analyse spectrale (Ng *et al.*, 2001). Après comparaison des différentes approches de la littérature à travers plusieurs tests (section 6), l'approche non linéaire s'est avérée la plus performante pour notre problématique.

5.1. Formulation mathématique de l'analyse spectrale pour la réduction de dimensions

Considérons un ensemble de m régions clés $\{r_1, r_2, ..., r_m\}$ extraites d'une séquence d'images et décrites par leurs signatures $\{S_1, S_2, ..., S_m\}$. Un graphe $G = (V, E)$ où chaque sommet $v_i \in V$ correspond à une région r_i peut être construit. Tous les sommets v_i et v_j sont reliés deux à deux par une arête dont le poids reflète le degré de similarité entre les 2 régions clés. Ce poids est défini par :

$$W_{ij} = exp\left(-\frac{d\left(S_i, S_j\right)^2}{\sigma^2}\right) \qquad [21]$$

où $d\left(S_i, S_j\right)$ est la distance entre les signatures extraites des deux régions et $\sigma = \text{moyenne}\left[d\left(S_i, S_j\right)\right], \forall i, j = 1, ..., m\ (i \neq j)$.

Le but de cette étape est de passer d'un espace de grande dimension à un espace plus réduit. Pour ce faire, nous cherchons une nouvelle représentation $\{y_1, y_2, ..., y_m\}$ où $y_i \in R^m$ en minimisant le critère de coût $\phi = \sum_{ij} \left\| y_i - y_j \right\|_2 W_{ij}$. En introduisant

1. En utilisant la signature « couleur-position statistique » avec 8 bandes horizontales, chaque région-clé est caractérisée par un vecteur de 6144 dimensions.

la matrice diagonale des degrés des sommets D définie par $D_{ii} = \sum_j W_{ij}$ et la matrice Laplacienne du graphe $L = D - W$, le critère de coût peut être réécrit ainsi :

$$\varphi = \sum_{ij} \left\| y_i - y_j \right\|_2 W_{ij} = 2\text{Trace}\left(Y^T L Y\right) \text{ avec} Y = [y_1, y_2, ..., y_m] \qquad [22]$$

La solution est donc obtenue en recherchant les vecteurs propres de la matrice L : $y_1, y_2, ..., y_m$ et la réduction de dimension est obtenue en ne conservant que les d premiers vecteurs propres où d est très inférieur à la dimension des données d'entrée.

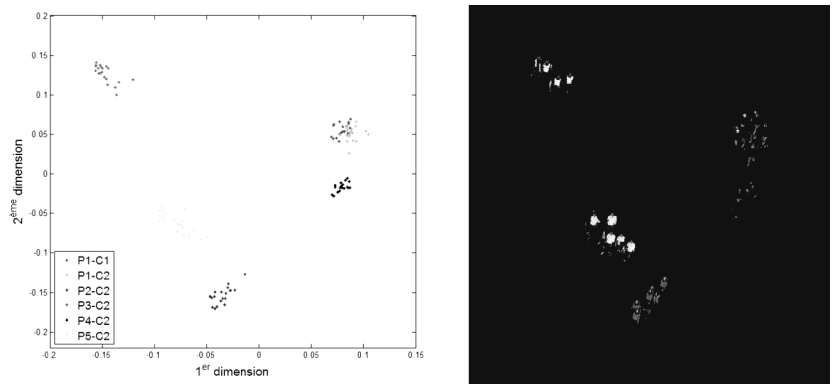


Figure 5. Projection des régions clés sur les deux premières dimensions correspondant aux deux valeurs propres les plus faibles obtenues par l'analyse spectrale

La figure 5 présente un exemple de résultat obtenu par cette méthode de réduction de dimension. Dans cet exemple, nous utilisons seulement les deux premiers vecteurs propres (correspondant aux valeurs propres les plus faibles) pour créer le nouvel espace de représentation. Cet exemple est alimenté par un ensemble de k régions clés appartenant à 6 séquences vidéo. Sur le diagramme de gauche, les régions sont illustrées par des points, tandis que les points sont directement illustrés par les régions clés correspondantes sur le diagramme de droite. Les deux premières séquences qui sont présentées par les points rouges et verts sur le diagramme de gauche appartiennent à la même personne filmée par deux caméras dans différents lieux, avec des champs de vision différents. Les autres points correspondent à des personnes différentes filmées par les deux mêmes caméras. On constate que la projection dans l'espace 2D conserve la quasi-totalité des informations originales et que le descripteur couleur-position statistique est suffisamment riche pour discriminer facilement les personnes entre elles : plus les personnes ont une apparence colorimétrique voisine et plus les points qui leur correspondent dans le nouvel espace réduit sont voisins.

5.2. Mesure de similarité entre séquences vidéo

L'objectif de la réduction de dimension est de pouvoir comparer des séquences d'images afin de ré-identifier une personne qui est passée devant une caméra, et réapparaît devant une autre caméra. Par conséquent, un ensemble de m régions clés appartenant à p passages devant la caméra 1 et le passage requête devant la caméra 2 sont considérés. En appliquant la réduction de dimension non linéaire utilisant l'analyse spectrale, un nouveau système de coordonnées qui tient compte des d premiers vecteurs propres (correspondant aux plus faibles valeurs propres) est obtenu. Comme chaque passage est représenté par plusieurs régions clés, le barycentre des points correspondant à chaque séquence dans le nouvel espace est calculé. La distance entre deux barycentres est considérée comme la dissimilarité entre les deux séquences correspondantes. Un seuil de décision sur cette distance permet ensuite d'opter ou non pour une réidentification.

6. Résultats expérimentaux

Le système complet a été validé sur deux bases de données correspondant aux passages de plusieurs personnes devant des caméras disposées à différents endroits. La première base de données (nommée base INRETS dans la suite de l'article) a été acquise dans les locaux de l'INRETS et contient des séquences vidéo de 40 personnes filmées à deux endroits différents comme illustré en figure 6 : 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 (nommée base train dans la suite de l'article) contient des séquences vidéo de 35 personnes filmées à deux endroits (dans un couloir et dans une voiture) (figure 7). Cette base de données est difficile, car les deux caméras ont des caractéristiques sensiblement différentes ; la qualité des données et le rendu des couleurs ne sont donc pas identiques. De plus, l'acquisition de la vidéo est influencée par de nombreux facteurs dégradants liés à l'application tels que des variations rapides d'éclairage dus au déplacement du train, des reflets sur les vitres, des vibrations.... La figure 8 illustre quelques images d'une séquence correspondant au passage d'une personne dans le couloir du train. On constate que les conditions d'éclairage sont extrêmement différentes même lors d'un seul passage (l'éclairage entrant par les vitres, qui est très variable, influe considérablement sur la scène).

Figure 6. Exemple d'images de la base INRETS

Figure 7. Exemple d'images de la base train

Figure 8. Illustration de changement de l'illumination à cause de mouvement du train

Dans nos expérimentations, nous évaluons les performances du système complet de détection/réidentification de personnes. La signature proposée est comparée avec les autres signatures de la littérature :

– les histogrammes avec 8 cellules par composante couleur. L'intersection d'histogrammes, qui mesure la similarité entre deux histogrammes, est utilisée pour construire le graphe (équation [21]) ;

– les spatiogrammes avec 8 cellules par composante couleur. La similarité entre deux spatiogrammes est estimée en utilisant la distance proposée dans (Birchfield *et al.*, 2005) ;

– la signature couleur/path-length avec 8 cellules par composante couleur et 8 cellules pour le descripteur path-length. La distance entre deux signatures couleur/path-length est estimée par la norme L1.

Pour chacune de ces signatures, nous évaluons l'intérêt de l'utilisation d'une normalisation couleur en testant successivement le calcul des signatures sans normalisation, puis avec chacun des 3 invariants présentés en sous-section 4.1. D'autre part, la réduction de dimension est réalisée en conservant les $d = 4$ premiers vecteurs propres pour créer le nouvel espace de représentation des données. Ce choix est justifié par une étude expérimentale dont les résultats sont présentés en figure 11.

Pour chaque passage requête devant une caméra, les distances entre le passage requête et chacun des passages candidat (passage devant la seconde caméra) sont calculées. Un seuil de décision est choisi. Les distances inférieures au seuil indiquent une réidentification (score = 1) et les distances supérieures au seuil indiquent une distinction (score = 0).

Soit K le nombre de passages devant chaque caméra, $K \times K$ distances sont calculées et comparées avec le seuil de décision. Une matrice de score est ainsi obtenue (idéalement, la matrice identité). Elle résulte des quatre cas suivants :

– vraie réidentification (vrai positif) : le système déclare une réidentification (score = 1 sur la diagonale) lorsque les deux passages appartiennent à la même personne ;

– vraie distinction (vrai négatif) : le système déclare une distinction (score = 0 hors diagonale) lorsque les deux passages appartiennent à deux personnes différentes ;

– fausse réidentification (faux positif) : le système déclare une réidentification (score = 1 hors diagonale) lorsque les deux passages appartiennent à deux personnes différentes ;

– fausse distinction (faux négatif) : le système déclare une distinction (score = 0 sur la diagonale) lorsque les deux passages appartiennent à la même personne.

Les performances du système présenté sont illustrées grâce aux courbes ROC qui mettent en relation les taux de vraie réidentification (TVR) avec les taux de fausse réidentification (TFR). Ces deux taux peuvent être calculés à partir de la matrice de score en utilisant les définitions suivantes :

$$\mathrm{TVR} = \frac{\sum_{k=1}^{K} (score_{kk} = 1)}{K} \qquad [23]$$

$$\mathrm{TFR} = \frac{\sum_{k=1}^{K} \sum_{l=1}^{K} (score_{kl} = 1, k \neq l)}{\mathrm{K}\,(K-1)} \qquad [24]$$

La figure 9, qui est divisée en 4 parties selon les 4 signatures utilisées, présente les courbes obtenues sur la base INRETS. A la lecture de cette figure, nous constatons que les résultats sont très satisfaisants avec un meilleur taux de réidentification de 97,5 % obtenu en utilisant le descripteur « couleur-position statistique » proposé (le taux de réidentification est défini pour chaque méthode à partir du seuil optimal tel que le taux de vrais positifs est égale au taux de vrais négatifs). Remarquons que quel que soit le descripteur utilisé, l'introduction d'une étape de normalisation permettant de devenir quasi invariant aux conditions d'éclairage améliore les résultats de réidentification.

Les résultats obtenus pour la base train sont présentés en figure 10. Le meilleur taux de 88,6 % est obtenu avec le descripteur « couleur-position statistique » malgré la forte difficulté de la base train. Pour les deux bases, les taux obtenus en utilisant l'histogramme couleur comme signature sont toujours les plus faibles. Les autres signatures amènent à de meilleurs résultats, ce qui montre l'intérêt d'introduire des informations spatiales dans la description des régions clés. Notons que sur cette base aussi l'introduction d'invariants colorimétriques améliore considérablement les taux de bonne réidentification.

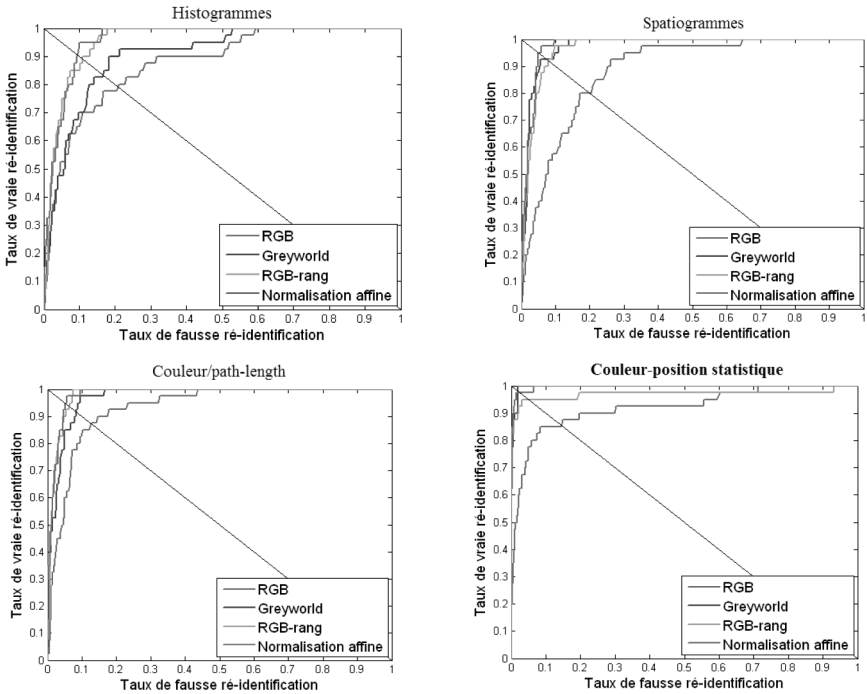


Figure 9. Résultats de réidentification obtenus pour la base INRETS (de gauche à droite, de haut en bas : histogramme, spatiogramme, couleur/path-length, couleur-position statistique)

Ces résultats ont été obtenus en utilisant un espace réduit de dimension 4 lors de l'analyse spectrale. Les tests présentés en figure 11 en faisant varier le nombre de vecteurs propres conservés montrent que les meilleurs résultats sont obtenus avec 4 vecteurs propres pour le descripteur couleur position statistique.

Des expérimentations supplémentaires ont également été menées afin de justifier l'emploi de l'analyse spectrale pour la réduction de dimension. Quatre approches ont ainsi été testées :

(1) Une approche sans réduction de dimension où la moyenne de toutes les signatures estimées pour un passage est calculée. La distance entre deux séquences correspond alors à la distance entre les deux signatures moyennes.

(2) Une approche par vote : pour chaque signature de la séquence requête, on recherche la signature la plus proche dans les autres séquences. La séquence correspondant à cette signature la plus proche reçoit alors un vote. La séquence la plus proche de la séquence requête sera alors celle qui obtient le plus grand nombre de votes.

(3) Une approche de réduction de dimension par analyse en composantes principales (ACP).

(4) Une approche de réduction de dimension par analyse spectrale (AS).

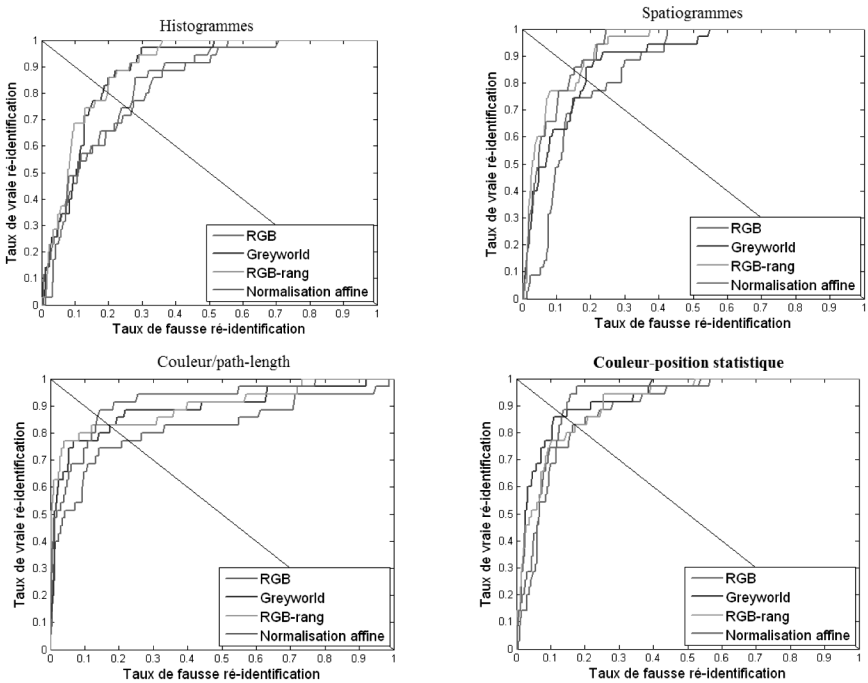


Figure 10. Résultats de réidentification obtenus pour la base train (de gauche à droite, de haut en bas : histogramme, spatiogramme, couleur/path-length, couleur-position statistique)

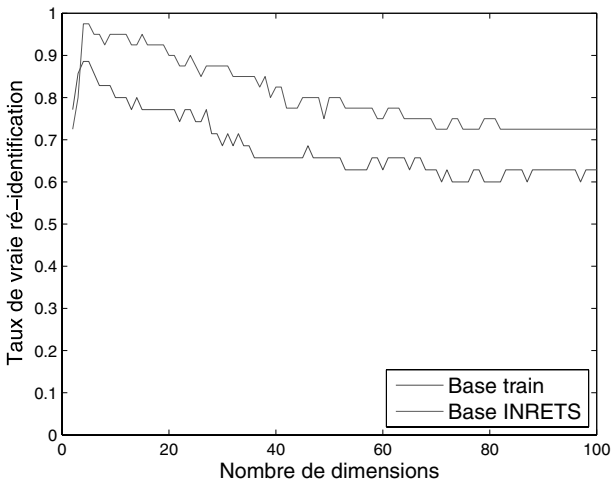


Figure 11. Résultats de réidentification obtenus en faisant varier la dimension de l'espace réduit

Les résultats de ces méthodes sont présentés dans les tableaux 1 et 2, pour chacune des bases testées. Nous constatons que quelque soit l'invariant choisi, l'analyse spectrale amène aux meilleurs taux de réidentification pour les deux bases. Les taux obtenus en utilisant la signature moyenne sont toujours les plus faibles. Ceci montre l'importance qu'il y a à exploiter un ensemble de signatures pour caractériser chaque passage. Ces résultats mettent aussi en avant l'intérêt qu'il y a à utiliser une étape de réduction de dimension : à cause de la grande dimension de la signature proposée, les méthodes n'utilisant pas de réduction de dimension n'amènent pas à de bons résultats. Enfin, pour cette étape de réduction de dimension, l'approche non linéaire utilisant l'analyse spectrale est plus performante qu'une approche classique exploitant les composantes principales.

Tableau 1. Performance des différentes approches de classification pour la réidentification obtenus pour la base INRETS

	RGB	Greyworld	RGB-rang	Affine
Moyenne	45	52,5	62,5	55,5
Vote	50	72,5	70	75
ACP	70	87,5	90	90
AS	85	97,5	95	97,5

Tableau 2. Performance des différentes approches de classification pour la réidentification obtenus pour la base train

	RGB	Greyworld	RGB-rang	Affine
Moyenne	34,29	51,43	54,29	48,57
Vote	37,14	51,43	45,71	51,43
ACP	45,71	54,29	51,43	51,43
AS	82,86	88,57	82,86	88,57

7. Conclusion

Dans cet article, nous avons présenté un système de vidéosurveillance permettant de suivre le déplacement d'un individu lorsque celui-ci passe devant différentes caméras. Nous proposons dans un premier temps un processus automatique d'extraction de silhouette basé sur la fusion de deux méthodes de détection de mouvement complémentaires : une détection par modélisation du fond par un mélange 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. Une étape de sélection des régions obtenues lors de cette détection est ensuite

introduite afin de ne conserver que celles correspondant réellement aux personnes en mouvement (régions clés). Par la suite, un nouveau descripteur de régions intitulé « couleur-position statistique » est proposé afin de caractériser les silhouettes extraites des images couleur avec un double critère utilisant à la fois des informations spatiales et colorimétriques. Celui-ci, qui a été comparé à plusieurs autres descripteurs de la littérature, amène à des résultats très satisfaisants. A cause des forts changements de luminosité forcément présents entre les différents lieux d'acquisition, une étape de normalisation a été introduite sur les images couleur afin d'obtenir une quasi-invariance aux conditions d'éclairage. Celle-ci améliore considérablement les résultats de réidentification. Une fois la signature de chaque région clé estimée, la comparaison des séquences vidéo passe par une étape de réduction de dimension obtenue grâce à l'analyse spectrale. Ceci permet ensuite d'utiliser une simple distance euclidienne pour réaliser la réidentification.

Les performances de cette approche ont été évaluées sur deux bases de données, une acquise dans les locaux de l'INRETS et une autre acquise à l'intérieur d'un train en mouvement. Les résultats expérimentaux obtenus sur ces deux bases montrent la robustesse de l'approche proposée. Les taux de réidentification, de 97,5 % pour la base INRETS et 88,6 % pour la base du train sont très satisfaisants et encourageants compte tenu de la difficulté de cette dernière base.

Les résultats de l'approche proposée sont parmi les meilleurs de l'état de l'art que nous reprenons ici. Gheissari *et al.* (2006) ont utilisé une base de données de 44 personnes et obtenu un taux de réidentification de 60 %. Dans les travaux de Yu *et al.* (2007), le taux obtenu est de 95 % pour une base de 30 personnes. Nakajima *et al.* (2003) ont utilisé une base avec un très faible nombre d'exemples (4 individus) et obtenu un taux de 100 %. Même s'il est vrai que ces résultats ne peuvent pas être directement comparés aux nôtres car les bases de données ne sont pas les mêmes, ils montrent la difficulté du problème traité. En outre, afin de mettre à disposition des chercheurs une base de données commune pour les études comparatives, l'ensemble de données acquises à l'intérieur d'un train en mouvement dans le cadre du projet BOSS est téléchargeable sur le site (BOSS). Nous espérons que ce sera une base de données de référence pour évaluer les travaux similaires de réidentification de personnes.

Plusieurs perspectives sont déjà envisagées pour améliorer ce système. Ainsi, l'introduction de statistiques sur les temps de déplacements entre les caméras devrait pouvoir améliorer le système, en rejetant des faux positifs et des faux négatifs. Des caractéristiques de plus haut niveau et plus gourmandes en temps de calcul, analysant le visage, la démarche ou la stature des personnes peuvent aussi être ajoutées. Enfin, une fusion de cette méthode, modélisant les silhouettes de manière globale, avec une approche très locale, caractérisant des points d'intérêt peut aussi être envisagée. A moyen terme, nous souhaitons également gérer des scénarios plus difficiles correspondant aux passages simultanés de plusieurs personnes devant une caméra.

Bibliographie

http://www.multitel.be/boss

Ballard D., Brown C. (1982). *Computer Vision*, Prentice Hall Professional Technical Reference.

Belkin M., Niyogi P. (2003). « Laplacian eigenmaps for dimensionality reduction and data representation », *Neural Computation*, vol. 15, n° 6, p. 1373-1396.

Birchfield S., Rangarajan S. (2005). « Spatiograms versus Histograms for Region-Based Tracking », *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 1158-1163.

Buchsbaum G. (1980). « A spatial processor model for object color perception », *Journal of the Franklin Institute*, vol. 310, n° 1, p. 1-26.

Elgammal A., Harwood D., Davis L. (2000). « Non-parametric model for background subtraction », *Proceedings of the 6th European Conference on Computer Vision*, p. 751-767.

Finlayson G., Hordley S., Schaefer G., Yun Tian G. (2005). « Illuminant and device invariant colour using histogram equalisation », *Pattern Recognition*, vol. 38, n° 2, p. 179-190.

Fisher R. (1936). « The use of multiple measurements in taxonomic problems », *Annals of Eugenics*, vol. 7, p. 179-188.

Gevers T., Stokman H. (2004). « Robust Histogram Construction from Color Invariants for Object Recognition », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 113-118.

Gheissari N., Sebastian T., Tu P., Rittscher J., Hartley R. (2006). « Person Reidentification Using Spatiotemporal Appearance », *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, p. 1528-1535.

Hartigan J. (1975). *Clustering Algorithms*, John Wiley & Sons, Inc. New York, NY, USA.

Hotelling H. (1933). « Analysis of a complex of statistical variables into principal components », *Journal of Educational Psychology*, vol. 24, p. 417-441.

Javed O., Rasheed Z., Shafique K., Shah M. (2003). « Tracking across multiple cameras with disjoint views », *Ninth IEEE International Conference on Computer Vision*.

Javed O., Shafique K., Shah M. (2005). « Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras », *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 26-33.

Kettnaker V., Zabih R. (1999). « Bayesian multi-camera surveillance », *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2.

Kim K., Chalidabhongse T., Harwood D., Davis L. (2004). « Background modeling and subtraction by codebook construction », *International Conference on Image Processing*, vol. 5.

Kullback S., Leibler R. (1951). « On information and sufficiency », *Annals of Mathematical Statistics*, vol. 22, n° 1, p. 79-86.

Madden C., Piccardi M., Zuffi S. (2007). *Comparison of Techniques for Mitigating the Effects of Illumination Variations on the Appearance of Human Targets*, vol. 4842 of *Lecture Notes in Computer Science*, Springer.

Nadler B., Lafon S., Coifman R., Kevrekidis I. (2005). « Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators », *Advances in Neural Information Processing Systems*, p. 955-962.

Nakajima C., Pontil M., Heisele M., Poggio T. (2003). « Full body person recognition system », *Pattern Recognition*, vol. 36, n° 9, p. 1997-2006.

Ng A., Jordan M., Weiss Y. (2001). « On spectral clustering: Analysis and an algorithm », *Advances in Neural Information Processing Systems*, MIT Press, p. 849-856.

Parzen E. (1962). « On the estimation of a probability density function and mode », *Annals of Mathematical Statistics*, vol. 33, p. 1065-1076.

Porikli F., Tuzel O. (2003). « Human Body Tracking by Adaptive Background Models and Mean-Shift Analysis », *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*.

Stauffer C., Grimson W. (1999). « Adaptive background mixture models for real-time tracking », *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 246-252.

Truong Cong D.-N., Khoudour L., Achard C., Phothisane P. (2009). « People Re-identification by Means of a Camera Network Using a Graph-based Approach », *Proceeding of IAPR Conference on Machine Vision Applications*.

von Luxburg U. (2007). « A Tutorial on Spectral Clustering », *Statistics and Computing*, vol. 17, n° 4, p. 395-416.

Wang X., Doretto G., Sebastian T., Rittscher J., Tu P. (2007). « Shape and Appearance Context Modeling », *EEE 11th International Conference on Computer Vision*, p. 1-8.

Wren C., Azarbayejani A., Darrell T., Pentland A. (1997). « Pfinder: Real-time tracking of the human body », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, n° 7, p. 780-785.

Yoon K., Harwood D., Davis L. (2006). « Appearance-based person recognition using color/path-length profile », *Journal of Visual Communication and Image Representation*, vol. 17, n° 3, p. 605-622.

Yu Y., Harwood D., Yoon K., Davis L. (2007). « Human appearance modeling for matching across video sequences », *Machine Vision and Applications*, vol. 18, n° 3, p. 139-149.

Article reçu le 06/07/2009
Accepté le 03/05/2010

Dung Nghi Truong Cong *est née en 1981. Elle a obtenu le diplôme d'ingénieur en Télécommunications à l'Institut polytechnique de Ho Chi Minh Ville, Vietnam, en 2004, puis le diplôme de Master en Signal, Image, Parole et Télécommunications à l'Institut polytechnique de Grenoble, France, en 2007. Elle poursuit actuellement son doctorat à l'université Pierre et Marie Curie. Ses intérêts de recherche portent sur le traitement des images et de la vidéo, la vision par ordinateur et l'analyse de données.*

Catherine Achard *est née en 1970. Elle a obtenu le doctorat de l'université Blaise Pascal de Clermont-Ferrand en 1996, puis l'habilitation à diriger des recherches en 2007. Depuis 1997, elle est maître de conférences à l'université Pierre et Marie Curie. Ses thématiques de recherche actuelles portent essentiellement sur l'analyse vidéo et plus spécifiquement sur l'étude du comportement humain.*

Louahdi Khoudour *est docteur en automatique et informatique industrielle de l'université de Lille en 1996. De formation de base en statistique, il a obtenu une maîtrise de mathématiques, un DESS de statistiques de l'université Paul Sabatier de Toulouse, puis un DEA en automatique à SUPAERO de Toulouse. Il est chercheur à l'INRETS depuis 1994 dans le domaine du traitement du signal et des images appliqués à la sécurité dans les transports.*