Approche spectrale et descripteur “couleur-position statistique” pour la ré-identification de personnes à travers un réseau de caméras

Dung-Nghi TRUONG CONG[1], Louahdi KHOUDOUR[1], Catherine ACHARD[2]

[1]Institut National de Recherche sur les Transports et leur Sécurité (INRETS)
20 rue Elisée Reclus, 59650 Villeneuve d’Ascq, France.

[2]UPMC Univ Paris 06, Institut des Systèmes Intelligents et de Robotique (ISIR)
Case Courrier 252, 3 rue Galilée, 94200 IVRY SUR SEINE, France.
truong@inrets.fr, louahdi.khoudour@inrets.fr, catherine.achard@upmc.fr

Résumé – La problématique décrite dans cet article consiste à mettre en place un système robuste permettant de ré-identifier des personnes évoluant dans différentes sites surveillés par des caméras de surveillance. Nous extrayons dans un premier temps des signatures colorimétriques qui sont à la fois discriminantes et invariantes aux différentes conditions d’éclairage. Dans un second temps, afin de tenir compte de toutes les informations utiles caractérisant le passage d’une personne devant une caméra, nous proposons une approche basée sur l’analyse spectrale pour la comparaison des séquences vidéo. Cette approche est capable de caractériser au mieux l’information en se basant sur les graphes de similarités et l’exploitation de leurs propriétés spectrales. Deux bases de données représentant le passage de plusieurs dizaines de personnes devant des caméras ont été utilisées pour l’évaluation des algorithmes. Les résultats expérimentaux permettent de valider les approches en cours de développement.

Abstract – The problem described in this paper consists in re-identifying moving people in different sites with non-overlapping views. Firstly, we propose a new feature called “color-position statistics” signature combined with several illuminant invariant methods in order to characterize extracted silhouettes in static images. Then, a graph-based approach which is capable of learning the global structure of the manifold and preserving the properties of the original data in a lower dimensional representation is introduced to reduce the effective working space and to carry out the comparison of the video sequences. The global system is tested on two real and difficult data sets recorded in very different environments. The experimental results show that the combination of color-based feature, invariant normalization procedures and the graph-based approach leads to very satisfactory results.

1 Introduction

Depuis quelques années, la vidéosurveillance a connu un essor considérable lié à son rôle primordial dans la sécurité. Des efforts de recherche importants ont été consacrés à la détection automatique et rapide d’événements amenant à une atteinte à la sécurité tels que les agressions envers les personnes, le vandalisme contre les biens, les actes de terrorisme, les accidents,.... Une des tâches les plus importantes et difficile à réaliser consiste à suivre le déplacement d’un individu à travers les différentes caméras. Dans la plupart des cas, la résolution de cette tâche passe par une modélisation colorimétrique et spatiale de l’apparence des personnes. Ceci implique de gérer plusieurs problèmes tels que les conditions d’éclairage variables, les points de vue différents pour chaque caméra, les variations de la pose des personnes,....

Plusieurs approches ont été proposées dans la littérature pour la ré-identification de personnes basée sur l’apparence globale. Kettnaker et Zabih [1] exploitent la similarité des vues de personnes, ainsi que la “plausibilité” du temps de parcours d’une caméra à l’autre. Dans les travaux de Nakajima et al. [2], des descripteurs colorimétriques sont utilisés comme signature puis introduits dans un algorithme de reconnaissance à base de SVMs (Support Vector Machines) multi-classes. Gheissari et al. [3] proposent un système utilisant la mise en correspondance de points d’intérêt. Récemment, Yu et al. [4] introduisent un nouveau descripteur couleur-spatial et un algorithme pour sélectionner des images-clés et comparer des séquences vidéo.

Dans cet article, nous proposons un système de détection/ré-identification de personnes se déplaçant à travers un réseau de caméra, se décomposant en deux étapes principales : 1) l’extraction de l’individu du fond et l’estimation de sa signature colorimétrique invariante. 2) La comparaison des séquences vidéo afin de suivre les personnes lors de leur déplacement au travers les champs des différentes caméras. Nous proposons dans un premier temps une signature colorimétrique spécifique, robuste, capable de s’adapter aux variations brutales d’illumination et aux changements de vue de la personne. Par la suite, afin de tenir compte de toutes les informations utiles caractérisant le passage d’une personne devant une caméra, nous proposons une approche basée sur l’analyse spectrale pour la comparaison des séquences vidéo. Cette approche caractérise au mieux l’information en se basant sur les graphes de similarités et l’exploitation de leurs propriétés spectrales. La performance de notre système est évaluée dans un contexte très difficile sur deux bases de données représentant le passage de plusieurs dizaines de personnes devant des caméras.

L’organisation de l’article est la suivante : après cette introduction, nous présentons en section 2, les signatures colorimétriques utilisées pour caractériser les silhouettes détectées.

En section 3, nous expliquons comment nous avons adapté l'analyse spectrale à notre problématique. Des résultats globaux sur la performance de notre système sur deux bases de données sont présentés en section 4, avant les conclusions et perspectives à court terme de ces travaux.

2 Caractérisation de silhouette

La première étape de notre système consiste à extraire des signatures colorimétriques caractérisant chaque individu. Pour ce faire, les silhouettes des personnes sont extraites dans un premier temps avec l'algorithme codebook [5] combiné à des opérations de morphologie mathématique (dilatation et érosion). Cet algorithme a l'avantage d'être très robuste aux environnements bruités. Ainsi, nous arrivons à extraire des silhouettes tout à fait exploitables dans un environnement embarqué difficile. Supposons à présent que la silhouette de chaque personne est localisée dans toutes les images d'une séquence vidéo.

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 [6] ont étendu le concept d'histogramme en introduisant des informations spatiales. 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. [7] 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 intitulé "couleur-position statistique". L'idée de ce descripteur consiste à effectuer un découpage horizontal de la silhouette en N bandes équidistantes. Fixer le nombre de bandes plutôt que leur taille permet d'obtenir une invariance face à l'échelle de la personne. Chacune de ces bandes est ensuite caractérisée par sa distribution de couleur grâce à une estimation par noyau [8] (aussi appelée méthode de Parzen-Rozenblatt).

Soit $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$ l'ensemble des pixels de la bande n où $\mathbf{x}_i$ est un vecteur couleur de dimension k. La densité de probabilité de chaque composante couleur est estimée par :

$$f_{nj}(x) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{h}K\left(\frac{x - x_{ij}}{h}\right) \quad (1)$$

où K est le noyau gaussien, h est le paramètre de lissage et j représente la composante couleur étudiée. Une silhouette est donc caractérisée par $N \times k$ fonctions de densité de probabilité f_{nj}.

Pour mesurer la ressemblance des silhouettes, nous introduisons une distance définie à partir de la divergence de Kullback-Leibler [9] qui mesure la similarité entre deux distributions de probabilités. La distance entre deux silhouettes S et $S^{'}$ est alors définie par :

$$d\left(S, S^{'}\right) = \sum_{n=1}^{N}\sum_{j=1}^{k}\left|d^{*}_{KL}\left(f_{nj}, f^{'}_{nj}\right)\right| \quad (2)$$

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 :

$$d^{*}_{KL}\left(f_{nj}, f^{'}_{nj}\right) = \frac{d_{KL}\left(f_{nj}||f^{'}_{nj}\right) + d_{KL}\left(f^{'}_{nj}||f_{nj}\right)}{2} \quad (3)$$

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

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. De nombreux invariants couleur ont été proposés dans la littérature [10, 11, 12]. Après de multiples tests,, nous présentons dans cet article, les trois invariants qui ont amené aux meilleurs résultats :

- Normalisation de Greyworld [13] :

$$I^{*}_{R,V,B} = \frac{I_{R,V,B}}{\texttt{moyenne}\,(I_{R,V,B})} \quad (4)$$

- Normalisation affine :

$$I^{*}_{R,V,B} = \frac{I_{R,V,B} - \texttt{moyenne}(I_{R,V,B})}{\texttt{écart type}(I_{R,V,B})} \quad (5)$$

- RGB-rang : Finlayson et al. [11] supposent que les mesures de rang colorimétrique des pixels sont insensibles aux changements d'éclairage. La mesure de rang colorimétrique $M_{R,V,B}(g)$ du niveau g présent dans l'image de composante $I_{R,V,B}$ est obtenue par l'égalisation de l'histogramme monodimensionnel $H_{R,V,B}$:

$$M_{R,V,B}(g) = \sum_{u=0}^{g} H_{R,V,B}(u) / \sum_{u=0}^{Nb} H_{R,V,B}(u) \quad (6)$$

où Nb indique le nombre de niveaux de quantification des composantes couleurs.

3 Réduction de dimensions pour la mesure de similarité entre séquences

La réduction de dimensions est un procédé important utilisé dans divers problèmes d'analyse de données. Elle est réalisée en ne gardant que les dimensions les plus importantes qui portent la majorité de l'information. Une représentation des données dans un espace de dimension inférieure est avantageuse dans bien des cas - classification, visualisation, compression de données, etc - et permet ne de pas être confronté aux problèmes liés aux grandes dimensions.

Au cours des dernières années, un grand nombre de techniques non-linéaires pour la réduction de dimension ont été proposées, telles que Laplacian Eigenmaps [14], Diffusion

Maps [15] et plusieurs variantes de l'analyse spectrale [16]. Ces techniques peuvent traiter avec des données non-linéaires complexes en préservant les propriétés globales et/ou locales des données originales dans la nouvelle représentation. Dans cet article, nous nous concentrons sur le principe de l'analyse spectrale pour la réduction de dimensions et son application spécifique à notre problématique.

Soit un ensemble de m images $\{I_1, \ldots, I_m\}$, nous considérons un graphe $G = (V, E)$ où chaque image I_i correspond à un sommet $v_i \in V$. Deux sommets v_i et v_j correspondant à deux images I_i et I_j sont reliés par une arête dont le poids reflète le degré de similarité entre les 2 images. Ce poids est défini par $W_{ij} = exp\left(-\frac{d(S_i,S_j)^2}{\sigma^2}\right)$, où $d(S_i, S_j)$ est la distance entre les signatures extraites des deux images (eq. 2) et $\sigma = moyenne\left[d(S_i, S_j)\right],\ \forall i, j = 1, \ldots, m\ \ (i \neq j)$.

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 $\{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}\}$ où $\mathbf{y}_i \in R^m$ en minimisant le critère de coût $\phi = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2 W_{ij}$.

En introduisant 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 :

$$\phi = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2 W_{ij} = 2\mathtt{Tr}\left(\mathbf{Y}^T L \mathbf{Y}\right) \qquad (7)$$

avec $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}]$.

Finalement, la réduction de dimension est obtenue en résolvant $Ly = \lambda Dy$ avec les valeurs propres généralisées.

L'objectif de notre recherche consiste à ré-identifier une personne qui est passée devant une caméra, et réapparaît devant une autre caméra. Par conséquent, un ensemble de m images appartenant à p passages devant la caméra 1 et le passage requête devant la caméra 2 est considéré (notons que chaque passage est caractérisé par la concaténation de n images, donc $m = n * (p + 1)$). En appliquant la méthode basée sur l'analyse spectrale, on obtient un nouvel espace de représentation qui tient compte des d plus faibles vecteurs propres. Dans notre application, nous exploitons les 10 plus faibles vecteurs propres pour générer le nouvel espace réduit. La réduction de dimensionnalité est définie par : $h : S_i \rightarrow u_i = [y_1(i), \ldots, y_{10}(i)]$, où $y_k(i)$ est la $i^{\text{ème}}$ coordonnée de vecteur propre $\mathbf{y}_k$. Comme chaque passage est représenté par n images, le barycentre des n points dans le nouvel espace est calculé. La distance entre deux barycentres est considérée comme la dissimilarité entre les deux séquences correspondantes. Pour chaque passage requête devant une caméra, les distances entre le passage requête et chacun des passages candidat sont calculées. Un seuil de décision est choisi. Les distances inférieures au seuil indiquent une ré-identification.

4 Résultats expérimentaux

Afin d'évaluer la performance de notre approche, nous avons utilisé deux bases de données correspondant au passage de plusieurs personnes devant des caméras disposées à différents endroits. La première base de données acquise dans les locaux de l'INRETS contient des séquences vidéo de 40 personnes à deux endroits différents (à l'intérieur 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 à deux endroits (dans un couloir et dans une voiture). Cette base de données est plus difficile, car l'acquisition de la vidéo est influencée par de nombreux facteurs, tels que des variations rapides d'éclairage, des reflets, des vibrations,....

FIGURE 1 – Exemple d'images de deux bases de données.

Dans nos expérimentations, nous comparons dans un premier temps notre signature proposée avec les trois autres signatures : les histogrammes avec 8 bins par composante couleur, les spatiogrammes avec 8 bins par composante couleur, la signature couleur/path-length avec 8 bins par composante couleur et 8 bins pour le descripteur path-length. Pour chaque passage d'un individu, une seule image-clé est extraite dans un premier temps et les différentes signatures sont estimées sur cette image. Les taux de bonne ré-identification sont ensuite calculés grâce à l'algorithme du plus proche voisin (Table 1).

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

	RGB	Greyworld	RGB-rang	Affine
Histogrammes	37.1	42.9	51.4	40
Spatiogrammes	48.6	65.7	57.1	54.3
Couleur /Path-length	48.6	62.9	62.9	45.7
Couleur-position statistique	54.3	**74.3**	68.6	65.7

Nous constatons que le meilleur taux de 74.3% est obtenu en utilisant le descripteur "couleur-position statistique" proposé et la normalisation Greyworld. 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 silhouettes. Notons aussi que l'introduction d'invariants améliore considérablement les taux de bonne ré-identification.

Afin d'améliorer les résultats de ré-identification, un ensemble de signatures caractérisant chaque passage de personne devant une caméra est utilisé. Par conséquence, l'algorithme de réduction de la dimension utilisant le principe de l'analyse spectrale est mis en place. Les performances obtenues par cette approche sont illustrées en utilisant les courbes ROC qui mettent en relation les taux de fausse ré-identification et les taux de vraie ré-identification. La figure 2 présente ces courbes obtenues sur les deux bases de données de test.

A la lecture de la figure 2, nous constatons que notre approche de comparaison des séquences vidéo basée sur la réduction de dimension obtenue grâce à l'analyse spectrale conduit à des résultats très satisfaisants. Celle-ci améliore

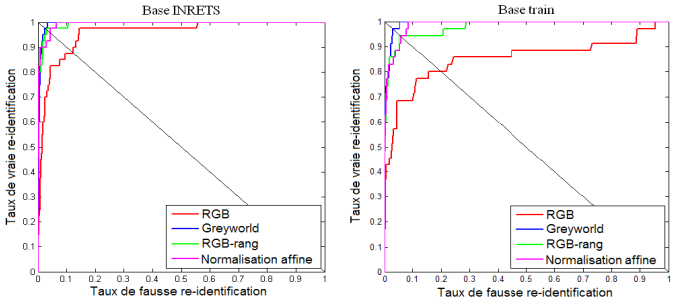

FIGURE 2 – Résultats expérimentaux obtenus pour les deux bases de données.

considérablement les taux de ré-identification par rapport à l'utilisation d'une seule signature pour chaque passage de personne (Les meilleurs taux obtenus sont de 97.5% pour la première base et 97.1% pour la deuxième base). Ces résultats montrent aussi l'avantage de l'utilisation des invariants : l'introduction d'une étape de normalisation permet de devenir quasi-invariant aux conditions d'éclairage et d'améliorer les résultats de ré-identification.

Les résultats de l'approche proposée sont parmi les meilleurs de l'état de l'art. Gheissari et al. [3] ont utilisé une base de données de 44 personnes et obtenu un taux de ré-identification de 60%. Dans les travaux de Yu et al. [4], le taux obtenu est de 95% pour une base de 30 personnes. Nakajima et al. [2] ont utilisé une base avec un très faible nombre d'exemples (4 individus) et obtenu un taux de 100%.

5 Conclusion

La recherche présentée dans cet article s'inscrit dans le cadre du projet européen BOSS dans lequel nous avons en charge le développement d'un système de vidéosurveillance permettant de suivre le déplacement d'un individu lorsque celui-ci passe devant un réseau de caméras. Un nouveau descripteur intitulé "couleur-position statistique" est proposé afin de caractériser la silhouette extraite à partir des images couleur. Celui-ci, qui a été comparé à plusieurs autres descripteurs de la littérature, amène à de très bons résultats. Une approche basée sur l'analyse spectrale pour la réduction de dimension est ensuite introduite dans l'étape de comparaison des séquences vidéo.

Les performances de notre système 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 ré-identification, de 97.5% pour la base INRETS et 97.1% pour la base du train sont très satisfaisants et encourageants compte tenu de la difficulté de la base.

Références

[1] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 1999.

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

[3] 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. IEEE Computer Society.

[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] K. Kim, TH Chalidabhongse, D. Harwood, and L. Davis. Background modeling and subtraction by codebook construction. In *International Conference on Image Processing, ICIP'04.*, volume 5, 2004.

[6] 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.

[7] K. Yoon, D. Harwood, and L.S. Davis. Appearance-based person recognition using color/path-length profile. *Journal of Visual Communication and Image Representation*, 17(3) :605–622, 2006.

[8] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[9] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1) :79–86, 1951.

[10] T. Gevers and H. Stokman. Robust histogram construction from color invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 113–118, 2004.

[11] 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.

[12] C. Madden, M. Piccardi, and S. Zuffi. *Comparison of Techniques for Mitigating the Effects of Illumination Variations on the Appearance of Human Targets*, volume 4842 of *Lecture Notes in Computer Science*. Springer, 2007.

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

[14] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6) :1373–1396, 2003.

[15] Boaz Nadler, Stephane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *Advances in Neural Information Processing Systems*, pages 955–962, 2005.

[16] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4) :395–416, 2007.