Ré-identification de Personnes par Modèle de Noyaux de Graphe

Amal MAHBOUBI[1], Luc BRUN[1], Donatello CONTE[2], Pasquale FOGGIA[2], Mario VENTO[2]

[1]GREYC UMR CNRS 6072
Equipe Image ENSICAEN 6 boulevard Maréchal Juin, 14050 Caen Cedex 04, France

[2]Dipartimento di Inegneria Electtronica e Ingeneria Informatica
Universita di Salerno, Via Ponte Don Melillo, 1 I-84084 Fisciano (SA), Italy

amal.mahboubi@unicaen.fr, luc.brun@ensicaen.fr
dconte@unisa.it, pfoggia@unisa.it, mvento@unisa.it

Résumé – Nous proposons dans cet article, des noyaux sur graphe définis entre le voisinage orienté de chaque noeud à base de noyau Gaussien permettant la modélisation des individus et leur ré-identification. Plusieurs expérimentation sont présentées pour illustrer l'efficacité de l'approche pour la ré-identification.

Abstract – We propose in this paper, a graph kernel devoted to describing a person and its re-identification and based on oriented neighborhoods. Several experiments are presented to illustrate the effectiency and the utilility of the proposed approach for visual surveillance applications.

1 Introduction

La ré-identification des personnes est un domaine actif en reconnaissance des formes. La problématique de la ré-identification est de déterminer si une personne a déjà été observée par un réseau de caméras. De nombreuses études traitant de la ré-identification ont été proposées, elles se répartissent en deux catégories. Une première catégorie de méthodes [1] [2] utilise la signature de la personne sur plusieurs images. Une seconde catégorie [3] [4] utilise une représentation de la silhouette de la personne. Dans cet article on ne s'intéressera qu'à la première catégorie. Le principe de notre approche est de représenter chaque occurrence d'une personne par un graphe et de décrire une personne par un ensemble de graphes calculés sur une fenêtre temporelle. La similarité entre deux personnes est alors calculée par des méthodes à noyaux sur graphes adaptées au type d'image traité. L'architecture des noyaux présentés dans les sections 2 et 3 s'inscrit dans la continuité d'une précédente contribution [5] où nous avons défini un schéma de construction de noyaux sur graphes pour l'indexation d'images. L'originalité de cette étude par rapport à nos précédents travaux est décrite dans les sections 4 et 5. Une série d'expériences est proposée en section 6.

2 Graphes décrivant une personne

La première étape dans un processus de suivi par ré-identification est l'extraction des objets de l'arrière-plan. A cet effet, nous utilisons les travaux de [6] où la détection des objets est effectuée automatiquement par soustraction du fond. Les masques binaires des objets nous sont fournis par [6]. Chaque personne est associée à un masque. Afin de construire le graphe de la personne, nous ne considérons pas l'ensemble des pixels du masque mais seulement ses points d'intérêts. Ce filtrage est réalisé grâce aux descripteurs SIFT [7]. Chaque point d'intérêt est représenté pas ses cordonnées x et y, l'échelle, l'orientation et un vecteur de 128 éléments (qui correspond aux descripteurs SIFT) par composante couleur. Une fois que les descripteurs SIFT d'un masque sont obtenus, on peut construire le graphe $G = (V,E,\mu,e)$ associé à cette personne. Dans notre étude l'ensemble V est constitué par les points SIFT, tandis que l'ensemble $E \subseteq V \times V$ représente les relations spatiales entre les sommets. La fonction μ associe à chaque sommet ses descripteurs SIFT tandis que la fonction e code l'importance d'un sommet codée par son échelle SIFT. Deux sommets sont liés par une arête si et seulement si ils appartiennent au voisinage l'un de l'autre au sens des k plus proches voisins suivant la distance Euclidienne. Ainsi chaque sommet du graphe est affecté d'un degré borné par k. Le voisinage $\mathcal{N}(u)$ d'un noeud u code l'ensemble de ses neuds adjacents. Afin de tenir compte de l'orientation du plan, ce voisinage est orienté dans le sens contraire des aiguilles d'une montre autour du noeud central u. Le premier noeud u_1 d'un voisinage orienté $\mathcal{N}(u) = (u_1,...,u_n)$ est par convention, le noeud le plus à droite dans la liste des noeuds adjacents à u.

3 Noyau sur graphes

Un noyau K sur un ensemble χ est une fonction symétrique $\chi \times \chi \rightarrow \mathscr{R}$ qui modélise le critère de similarité sur χ. Dans le

cas de deux graphes G_1 et G_2, $K(G_1,G_2)$ correspond au produit scalaire $K(G_1,G_2) =< \phi(G_1),\phi(G_2) >$ dans un espace Hilbertien $\mathscr{H}$, avec $\phi : \mathscr{G} \rightarrow \mathscr{H}$ une fonction de projection qui à un graphe fait correspondre un vecteur de $\mathscr{H}$. La fonction ϕ n'a pas besoin d'être explicitement construite pour créer un noyau. Néanmoins, K doit être défini positif.

Le schéma directeur dans la construction de notre noyau consiste à définir un noyau mineur entre le voisinage orienté de deux noeuds préservant ainsi l'information structurelle inhérente aux graphes. Ce noyau défini sur le voisinage orienté de chaque noeud s'appuie sur un noyau Gaussien ce qui garantit sa positivité. Nous regroupons par la suite les résultats obtenus sur le voisinage orienté de chaque noeud en un noyau de plus haut niveau qui définit ainsi un produit scalaire entre les deux graphes.

Noyau mineur. Soient deux noeuds u et v, leurs voisinages orientés notés resp. $\mathcal{N}(u)$ et $\mathcal{N}(v)$, tel que $l_u = |\mathcal{N}(u)|$ et $l_v = |\mathcal{N}(v)|$:

$$K_{seq}(u,v) = \begin{cases} 0 & si\, l_u \neq l_v \\ \prod_{i=1}^{l_u} K_g(u_i,v_i) & sinon \end{cases} \quad (1)$$

Avec $K_g(u,v)$ un noyau gaussien sur les descripteurs SIFT associés à chaque noeud tel que $K_g(u,v) = e^{-\frac{d(\mu(u),\mu(v))}{\sigma}}$, où σ est un paramètre fixé expérimentalement ; $d(.,.)$ est la distance Euclidienne ; $\mu(u)$ et $\mu(v)$ les attributs des noeuds u et v. Le rôle de l'équation 1 est d'apparier deux voisinages de longueur identiques. Le fait est que dans la pratique l'image est généralement bruitée. Afin d'atténuer l'effet des distorsions liées au bruit (suppression ou ajout de noeuds dans le voisinage), on propose d'introduire des règles de réécriture afin de comparer des voisinages de tailles différentes. Étant donné un noeud v, la réécriture de son voisinage orienté $\mathcal{N}(v)$ est défini par la fonction $\kappa(v)$ comme suit : $\kappa(v) = \kappa(v_1,...,\widehat{v_i},...,v_{l_v})$ où $\widehat{v_i}$ est le voisin à supprimer lors de la réécriture ; $\widehat{v_i}$ est le voisin de plus faible poids : $\widehat{v_i} = argmin_{j\in\{1,...,l_v\}} e(v_j)$. Le poids considéré ici est l'échelle du point SIFT. L'itération de cette règle de réécriture fournie une hiérarchie de voisinages orientés: $(\kappa^1(v),\kappa^2(v),...,\kappa^D(v))$ où D est un paramètre fixé expérimentalement. Le procédé de réécriture est accompagné d'un coût. À la fin des réécritures nous obtenons un coût cumulatif qui est la somme des coûts des noeuds supprimés à chaque étape de la réécriture. La formulation de ce coût est la suivante :

$$\begin{array}{l} Ce(v) = 0 \\ Ce(\kappa^i(v)) \quad = \quad e(\widehat{v_i}) + Ce(\kappa^{i-1}(v)) \end{array} \quad (2)$$

Noyau sur voisinage. Ces réécritures induisent la nouvelle formulation de notre noyau en :

$$\begin{array}{l} K_{rewriting}(u,v) = \\ \sum_{i=1}^{D_u} \sum_{j=1}^{D_v} W(\kappa^i(u),\kappa^j(v)) * K_{seq}(\kappa^i(u),\kappa^j(v)) \end{array} \quad (3)$$

Tel que le noyau W permet de pénaliser le coût des suppressions des points importants, il s'exprime ainsi :

$$W(\kappa^i(u),\kappa^j(v)) = e^{-\frac{Ce(\kappa^i(u))+Ce(\kappa^j(v))}{\sigma'}} \quad (4)$$

où σ' est fixée expérimentalement. D est le nombre de réécriture. Le procédé de réglage de cette constante est détaillé dans [5] où nos expérimentations ont aboutis à fixer D à 50% du degré du graphe.

Noyau sur Graphe. Le précédant paragraphe traite de la construction du noyau sur le voisinage c'est une étape nécessaire mais insuffisante. En effet nous ne devons pas omettre la contribution du noeud central dans la construction. Ainsi l'expression finale de notre noyau entre deux noeuds est donnée par :

$$K(u,v) = K_g(u,v)K_{rewriting}(u,v) \quad (5)$$

Une fois la construction du noyau entre deux noeuds définie, nous pouvons définir le noyau entre deux graphes par :

$$K_{graph}(G_1,G_2) = \sum_{u\in V_1} \sum_{v\in V_2} \varphi(u)\varphi(v)K(u,v) \quad (6)$$

où $\varphi(v) = e^{-\frac{1}{\sigma'(1+e(v))}}$ est la fonction qui pondère le poids du noeud central.

4 Description d'un individu

La ré-identification d'une personne à un instant t sur la base d'un unique t-prototype établit à un instant t_0 peut être altérée par exemple par le bruit ou les changements de postures au cours du temps. Concernant le bruit, la réécriture introduite au niveau du noyau mineur entre deux noeuds permet d'atténuer cette instabilité. Quant aux changements de postures, on peut définir une fenêtre temporelle de l'historique du suivi de taille HTW où la représentation d'une personne ne subit que de minimes modifications. Ainsi, afin d'obtenir une description plus robuste de chaque individu présent dans la vidéo, nous proposons de lui associer l'ensemble des graphes lui correspondant calculés sur HTW. Si nous notons S un tel ensemble, celui-ci est mis à jour durant la présence de l'individu dans la scène et est stocké dans un sac d'individus sortis BD_S lorsque celui-ci disparait de la vidéo. Cette description d'un individu par un ensemble de graphes peut toutefois inclure des graphes aberrants (erreur de suivi ou changement soudain de pose). Afin d'augmenter simultanément la compacité et la robustesse de notre description, nous projetons l'ensemble S des graphes décrivant un individu sur la sphère unité de l'espace de Hilbert $\mathcal{H}$ définit par K_{graph} (équation 6). Cette projection revient à normaliser les noyaux. Ensuite, le classifieur v-SVM à une classe est appliqué sur la projection de S. Cette étape permet de supprimer d'éventuels graphes aberrants. L'ensemble S est alors décrit par le vecteur de séparation w et la marge ρ issus du SVM. Ainsi chaque individu est représenté par le triplet (w,ρ,S). Considérant la description de deux personnes $P_A = (w_A,\rho_A,S_A)$ et $P_B = (w_B,\rho_B,S_B)$. La mesure de similarité entre ces deux personnes est la distance géodésique d_{sphere} entre les deux moyennes w_A et w_B sur l'hypersphère [8].

$$d_{sphere}(w_A,w_B) = arccos\left(\frac{w_A^T K(A,B) w_B}{\|w_A\|\|w_B\|}\right) \quad (7)$$

où: $\|w_A\|$ (resp. $\|w_B\|$) est la norme de w_A (resp. w_B) dans $\mathcal{H}$ et $K(A,B)$ est la matrice $|S_A| \times |S_B|$ définie par $K(A,B) = (K_{norm}(t,t'))_{(t,t') \in S_A \times S_B}$, où K_{norm} est notre noyau normalisé. Le noyau gaussien K_{change} entre deux personnes P_A et P_B est alors défini par :

$$K_{change}(P_A,P_B) = e^{\frac{-d^2_{sphere}(w_A,w_B)}{2\sigma^2_{moy}}} e^{\frac{-(\rho_A-\rho_B)^2}{2\sigma^2_{origin}}} \quad (8)$$

où σ_{moy} et σ_{origin} sont fixées expérimentalement.

5 Système de suivi par ré-identification

Notre système de ré-identification illustré dans la figure 1 utilise quatre étiquettes 'new', 'get-out', 'unknown' et 'get-back', leur signification est la suivante :

- new : un objet classifié nouveau.
- get-out : un objet sortie.
- unknown : un objet récemment apparu dans la scène non encore classifié (objet requête).
- get-back : un objet classifié réentrant.

L'ensemble des objets détectés dans la première image sont étiquetés new et affectés d'un numéro d'ordre. A l'instant $t+1$ on utilise un appariement par boîte englobante pour propager les numéros des objets persistants depuis t. Tout objet présent à t et n'ayant pas de correspondance à $t+1$ est étiqueté get-out et son modèle est sauvegardé dans BD_S. Tout objet présent à $t+1$ et n'ayant pas de correspondance à t est étiqueté unknown. Si à l'entrée de cet unknown, aucune sortie dans la vidéo n'a été observée, cet unknown est immédiatement classé new et affecté d'un nouveau numéro. Sinon, on diffère sa classification jusqu'à atteindre la largeur HTW afin de construire son triplet descriptif (w,ρ,S). La classification d'un unknown consiste à comparer le modèle de cet unknown à ceux de BD_S en utilisant l'équation 8. Nous obtenons ainsi autant de mesures de similarité que de get-out représentés par BD_S. Le get-out associé au maximum de ces mesures est déclaré réentrant s'il vérifie le critère SC, new dans le cas contraire. La décision réentrant consiste d'une part à affecter l'objet unknown du numéro du get-out, et d'autre part à supprimer ce get-out de BD_S. La définition du critère SC est la suivante : $\max_{Ker} > th_1$ ET $\sigma_{Ker} > th_2$, où $\max_{Ker}$ est le maximum des valeurs de similarité et σ_{Ker} leur variance. Les seuils th_1 et th_2 sont fixés expérimentalement. Notons que si $|BD_S| < 3$, SC se réduit au seuillage de la similarité maximale.

Parmi les variabilités que peuvent connaitre des individus qui interagissent dans une scène vidéo on peut citer l'occultation. La catégorie d'occultations traitée dans ce travail se cotonne au cas de superposition des boîtes englobantes. Le cas où une occultation à l'instant t entraine une fusion de boîtes englobantes existantes à $t-1$ est assimilé au phénomène de groupement d'individus. Cet aspect est en cours d'investigations et ne relève pas du contenu de cet article. Une occultation est détectée quand le nombre de pixels se recouvrant au niveau des deux boîtes englobantes est supérieur à un seuil déterminé expérimentalement. Quand un individu get-out est déclaré occulté sa description est compromise. Ainsi dans la construction du modèle le long d'une HTW si on détecte une occultation ce modèle est exclu de BD_S. A l'instant de l'identification le modèle de l'individu unknown est comparé à l'ensemble des individus get-out présents dans BD_S.

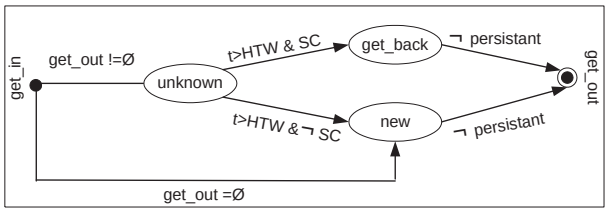

FIG. 1 – *Diagramme du système de ré-identification*

6 Expérimentation

Le jeu de données PETS2009 S2L1 [9] représente une séquence de vidéo surveillance contenant des personnes se déplaçant dans un espace ouvert. La vidéo surveillance est réalisée grâce à 7 caméras (v01, v03, v04, v05, v06, v07 et v08) positionnées à différents points de vue. Afin de quantifier les résultats il est nécessaire de labelliser les vidéos utilisées. Pour l'instant, nous avons labellisé manuellement les vidéo v01, v04, v05 et v06. Affin de quantifier nos résultats de suivi par ré-identification nous utilisons les métriques MODA et MOTA présentés en [10] leur expression est :

- Multiple Object Detection Accuracy

$$MODA = 1 - \frac{\sum_{t=1}^{N_{frames}}(m_t + fp_t)}{\sum_{t=1}^{N_{frames}} N_G^t}$$

- Multiple Object Tracking Accuracy

$$MOTA = 1 - \frac{\sum_{t=1}^{N_{frames}}(m_t + fp_t + c_s)}{\sum_{t=1}^{N_{frames}} N_G^t}$$

où m_t le nombre des mal classés ; fp_t le nombre de Faux Positive ; cs le nombre de commutations (non concordance entre deux frames successives) ; N_G^t le nombre d'objets de la vérité terrain dans la frame t

Bien que la métrique MODA est un score de détection, nous l'utilisons ici à titre indicatif car la qualité du suivi est influencée par la détection.

Afin de mesurer la qualité de la ré-identification nous utilisons les courbes CMC (pour Cumulative Match Characteristic). Cette courbe donne le pourcentage de personnes reconnues en fonction du rang. Ainsi, dans la figure 2 on peut observer que la qualité de la ré-identification de V01 est meilleure que celle de V05, V06 et V04.

Par ailleurs afin de se positionner par rapport à l'état de l'art, nous utilisons la comparaison exhaustive de 13 méthodes réalisée dans [11]. Cette comparaison utilise l'ensemble des métriques décrites dans [10]. Les meilleurs résultats sont obtenus par [12]. Nous reportons dans le tableau 1 la comparaison de

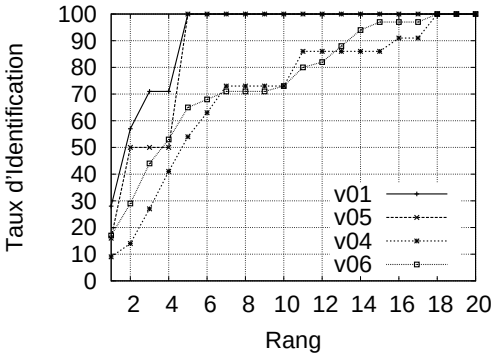

FIG. 2 – *Courbes CMC*

TAB. 1 – *Performances du suivi par ré-identification*

	résultats de [12]	Nos résultats	
Vidéo	MODA	MODA	MOTA
V01	0.67	0.91	0.91
V05	0.72	0.75	0.75
V04	0.61	0.2799	0.2790
V06	0.75	0.506	0.505

nos résultats avec ceux de [12]. La problématique de la ré-identification consiste à dire si l'objet apparu dans la scène est un nouvel objet ou un réentrant. Le tableau 1 résume les résultats obtenus. Ansi, dans le tableau 1 la première colonne présente les résultats de [12], les deux dernières colonnes répertorient nos résultats. On peut constater dans le tableau 1 que pour V04 et V06 nous obtenons de moins bon résultats que ceux de [12] alors que pour V01 et V05 nos performances sont meilleurs. Nous attribuons ceci au fait que les deux vidéos V06 et V04 ont la spécificité d'avoir des groupements tout le long de la séquence ce qui n'est pas traité par notre algorithme, alors que pour V01 et V05 seuls des occultations persistent tout le long de la séquence.

7 Conclusion et perspectives

Dans cet article, nous avons présenté des éléments de solution pour la ré-identification de séquences vidéo en utilisant les noyaux de graphe. La méthode utilisée dans ce travail a l'intérêt d'être simple et robuste. Les premiers résultats sont encourageants. Il nous reste à tester l'approche présentée dans les cas problématiques de groupement. Une analyse dans des réseaux de caméras est prévue. En effet, l'idée d'un modèle d'identité globale, assujetti aux contraintes (colorimétriques, spatio-temporelles, ...) dans le réseaux de caméras serait à considérer.

Références

[1] N. Bird, O. Masoud, N. Papanikolopoulos, A. Isaacs, Detection of loitering individuals in public transportation areas, *IEEE transactions on Intelligent Transportation Systems*, 6-2, pp.167-177, 2005.

[2] T. Le, M. Thonnat, A. Boucher and F. Brémond, Surveillance Video Indexing and Retrieval Using Object Features and Semantic Events *International Journal of Pattern Recognition and Artificial Intelligence*, Special issue on Visual Analysis and Understanding for Surveillance Application 23(7):1439-1476 (2009)

[3] S. Bak, E. Corvee, F. Brmond, M. Thonnat, Person re-identification using haar-based and dcd-based signature, *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp.1-8, 2010.

[4] S. Zhao, F. Precioso and M. Cord, Spatio-Temporal Tube data representation and Kernel design for SVM-based video object retrieval system. *Multimedia Tools Appl.*, (55):105-125 (2011)

[5] A. Mahboubi, L. Brun, F.X. Dupé, Object Classification Based on Graph Kernels, *Proc. IEEE Conf. on High Performance computing and Simulation*, Caen, France, 2010, 385-389.

[6] D. Conte, P. Foggia, G. Percannella, M. Vento, Performance Evaluation of a people tracking system on PETS2009 database, *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 119-126, 2010.

[7] D. Lowe, Distinctive Image Features from Scale-Invariant Keypoints. *International Journal for Computer Vision*, 60(2), 2004, 91-110.

[8] F. Desobry and M. Davy and C. Doncarli, An Online Kernel Change Detection Algorithm, *IEEE Transaction on Signal Processing*,vol 53:2961–2974,August 2005.

[9] PETS2009 database, Corpus établit dans le cadre du workshop Performance Evaluation of Tracking and Surveillance, disponible à: http://www.cvg.rdg.ac.uk/WINTERPETS09/a.html

[10] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova and Jing Zhang. Framework for performance evaluation of face, text, and vehicule detection and tracking in video: Data, metrics, and protocol, *Pattern Analysis and Machine Intelligence*,IEEE Transaction on, 31(2):319-336, Feb. 2009.

[11] A. Ellis, A. Shahrokni, and J. Ferryman, PETS 2009 and Winter PETS 2009 Results, a Combined Evaluation, *Twelfth IEEE Internatioanl Workshop on Performance Evaluation of Tracking and Surveillance*,1-8,2009.

[12] J. Berclaz, F. Fleuret, and P. Fua, Multiple object tracking using flow linear programming, *Twelfth IEEE Internatioanl Workshop on Performance Evaluation of Tracking and Surveillance*,2009.