Approche structurelle pour la ré-identification de personnes

Amal MAHBOUBI[1], Luc BRUN[1], Donatello CONTE[2]

[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC,
F-14050 Caen, France

[2]Université de Tours, LI EA 6300
F-37200, Tours, France
amal.mahboubi@unicaen.fr, luc.brun@ensicaen.fr, donatello.conte@univ-tours.fr

Résumé – Nous proposons dans cet article, une approche non supervisée pour la ré-identification de personnes fondée sur le représentation structurelle des personnes. Plusieurs expérimentations sont présentées sur deux jeux de données publics (ETHZ_REID et CAVIAR4REID) pour illustrer l'efficacité de l'approche pour la ré-identification multi-échantillons.

Abstract – This paper proposes an approach for pedestrian re-identification based on structural representation of people. The experimental evaluation is carried out on two public data sets (ETHZ_REID and CAVIAR4REID) and they show promising results compared to others state-of-the-art approaches in multiple-shot scenario.

1 Introduction

Le fort déploiement des systèmes de vidéosurveillance depuis le début des années 2000 s'accompagnât d'une effervescence de l'activité de recherche dans ce domaine. La ré identification des personnes (ré-id) consiste à suivre une personne dans un réseau de caméras à champs de vue disjoints. Le problème de ré-id peut être traité par deux types d'approches : les approches biométriques et les approches fondées sur l'apparence. Les approches biométriques requièrent la coopération des personnes, alors que la vidéosurveillance fournit des images de faible résolution et qui sont acquises sans soumettre les personnes à des contraintes spécifiques. Par opposition, les approches fondées sur une apparence globale de la personne ont moins de contraintes techniques mais font l'hypothèse que la personne conserve les mêmes vêtements le long de la ré-id.

Nous proposons ici deux approches structurelles où on décrit la personne par des caractéristiques d'apparence extraites dans le premier cas grâce à la moyenne pondérée de la norme du gradient afin de construire une courbe RGB modélisant l'apparence de la personne et dans le second cas grâce à la segmentation du blob enfermant une personne et l'encodage de cette segmentation au moyen d'un graphe d'adjacence de régions GAR. Pour chaque description structurelle, nous proposons un noyau approprié qui est utilisé pour la tâche de ré-identification 'non-supervisée'.

Un état de l'art de la ré-id est présenté dans la section 2. Les sections 3 et 4 détaillent les approches structurelles proposées. Nous présentons une série d'expériences et analysons les résultats dans la section 5 avant la conclusion dans la section 6.

2 La ré-identification

On appelle galerie l'ensemble des personnes connues. On le note C. On appelle personne requête une personne dont on souhaite retrouver l'identité. L'ensemble des personnes requête est noté T. Étant donné une ou plusieurs images d'une personne requête T et un ensemble galerie C constitué d'un certain nombre de personnes connues (où pour chaque personne on a une ou plusieurs images), le but de la ré-id est de définir pour la personne requête une liste classée des identités présentes dans la galerie en fonction de leur similitude. On distingue deux catégories d'approches fondées sur l'apparence : les méthodes supervisées et les méthodes non-supervisées dites directes. La première catégorie 'supervisée' nécessite une phase d'apprentissage initial qui doit être ajustée fréquemment. Nous pouvons citer les travaux de [9] qui utilisent une analyse du type moindres carrées partielles pour l'apprentissage d'une signature (couleur, texture, arêtes) individuelle, l'appariement quant à lui est réalisé avec la distance euclidienne. Dans la deuxième catégorie 'directes' un ensemble de descripteurs définissant la signature de la personne est extrait par la suite un appariement entre les signatures de la galerie et celle de la requête est réalisé. Ainsi [6] détecte dans un premier temps la personne en utilisant les points d'intérêts de type SURF. L'appariement des SURFs est réalisé grâce à une règle de minimisation des résiduels des SURFs. Finalement une règle de vote majoritaire minimisant l'erreur de reconstruction permet de connaître l'identité de la requête. Dans [1] le descripteur SDALF est utilisé comme signature, l'appariement est réalisé par une estimation de log-vraisemblance. Afin de résoudre le problème de ré-id, trois aspects sont à considérer : (1) détecter l'objet d'intérêt qu'est la personne, (2) le choix de la représentation ie. décrire

de manière appropriée un individu pour être en mesure de le reconnaître, (3) l'appariement d'une représentation galerie avec une représentation requête.

Détection des personnes à ré-identifier :
La détection de personnes est une tâche à accomplir en amont de la ré-id. Le but est de localiser la personne sur une image. Différentes stratégies de segmentation peuvent être considérées pour la ré-id comme la soustraction du fond ou les détecteurs de piétons fondés sur l'apparence. Dans la littérature, il est généralement admis que la boite englobant la personne soit disponible. Par conséquent, dans cet article, nous supposons que les masques binaires des personnes sont déjà extraits.

Signature visuelle d'une personne :
Dans un scénario de vidéosurveillance la galerie contient une petite quantité de données par personnes. De ce point de vu, les méthodes directes offrent une alternative aux méthodes supervisées qui nécessitent un corpus d'apprentissage conséquent. Dans la littérature, l'apparence d'une personne est décrite par des primitives de couleurs, de textures, de formes ou combinaison de ces primitives. Il n'y a aucun type de primitives systématiquement plus performant que les autres : chacun a des avantages et des limitations. La difficulté des méthodes directes est de définir des caractéristiques robustes et discriminantes formant une signature qui permet la ré-id. En effet un inconvénient des approches directes est la définition d'une signature par un sac de descripteurs ne permettant pas de capturer correctement la cohérence spatiale 2D de ces descripteurs. L'idée derrière notre travail est d'ajouter de la structure à ces sacs en utilisant deux approches différentes. Par ailleurs comme l'expression de notre signature repose grandement sur l'aspect structurel cela nous permet d'utiliser des descripteurs simples (comme la couleur, région, etc.). Nous proposons deux représentations : une chaîne RGB et le graphe d'adjacence de régions (GAR) d'une personne couplée à la distance d'édition. Ces deux représentations seront respectivement présentées dans la section 3 et 4.

Appariement de signatures :
Une fois que les descripteurs de la requête et de l'ensemble de la galerie ont été extraits, l'appariement consiste à identifier la requête dans la galerie. L'appariement peut être réalisé en utilisant un classifieur pouvant être un SVM, le plus proche voisin, etc. en utilisant la galerie comme jeu d'apprentissage. Ainsi, nous proposons la méthode suivante : à partir des descripteurs, nous construisons une distance qui renvoie l'identité de la requête en utilisant l'un des scénarios usuels en ré-id : mono-échantillon ou multi-échantillons. Donc dès que les descripteurs visuels sont extraits pour la requête et la galerie l'appariement peut être enclenché. Pour les chaînes RGB, l'appariement entre la requête et tous les individus de la galerie est obtenu en maximisant la similitude entre chaque paire possible de personnes contenues dans les deux ensembles. Ce procédé sera détaillé dans la section 3. Le même principe est utilisée pour la représentation région en remplaçant le noyau de la chaîne RGB par le noyau de la distance d'édition sur le GAR.

3 Noyau de chaine RGB

Notre noyau de chaine RGB s'inscrit dans la continuité d'une précédente contribution [7]. Cette approche s'appuie sur la description de chaque personne par un descripteur appelé chaîne RGB (pour 'Red-Green-Blue') qui consiste en une courbe modélisant l'apparence d'une personne. Considérons la boite englobante $W \times H$ d'un objet obj_a dont les coordonnées du point supérieur gauche sont notés (tl_x,tl_y). Pour chaque valeur $h \in \{0, \ldots, H-1\}$, nous considérons le segment de ligne horizontale définit par l'intersection de la boite englobante de obj_a et la ligne $y_h = h + tl_y$. La coordonnée x du point central de obj_a à la hauteur y_h est noté $\overline{x}_h$, il est définit comme étant la moyenne pondérée par la norme du gradient de tous les points le long du segment de droite.

La chaine RGB modélisant l'apparence d'une personne peut être altérée par la malencontreuse fusion d'un bout de fond avec la personne ou une différence d'échantillonnage due aux variations de la distance entre la personne et la caméra. Afin d'y remédier, nous considérons chaque courbe comme une chaîne et nous calculons la similitude entre deux chaînes en utilisant le noyau d'alignement global [2] définit par :

$$K_{GA}(s_1, s_2) = \sum_{\pi \in A(n,m)} e^{-D_{s_1,s_2}(\pi)}, \qquad (1)$$

où n et m, est respectivement la longueur de la première et seconde chaînes s_1 et s_2. Le symbole D est la distance de distorsion dynamique. Il mesure l'écart entre deux chaînes s_1 et s_2 selon un alignement π.

$\surd$ **Appariement mono-échantillon :** Étant donnée une personne requête x, l'ensemble requête T_x est un singleton. Nous détaillerons le choix des singletons dans la section 5. Afin de comparer la similarité entre deux individus, nous normalisons le noyau entre deux chaînes en utilisant la formule suivante :

$$\tilde{k}(s, s') = \frac{K_{GA}(s, s')}{\sqrt{K_{GA}(s, s) K_{GA}(s', s')}} \qquad (2)$$

Ensuite, l'appariement entre la requête et chaque sous-ensemble de la galerie est mesuré en utilisant la distance définie par l'équation :

$$d^2(s_T, s_{C_i}) = 1 - \tilde{k}(s_T, s_{C_i}) \; i = 1..H \qquad (3)$$

où s_T est la chaîne de l'ensemble requête, s_{C_i} est la chaîne de l'individu i et H le nombre d'individus.

Enfin, l'identité de la requête $l(T)$ est celle du plus petit élément de la liste classée des identités obtenues en utilisant l'équation 3. Ainsi $l(T) = \arg\min_{i=1}^{N} d^2(s_T, s_{C_i})$ où N est le nombre des sous ensembles dans C.

$\surd$ **Appariement multi-échantillons :** Un ensemble de N images est nécessaire pour établir la signature d'une personne, tel que chaque personne est décrite par une chaîne RGB par image. Comme pour le cas mono-échantillon, nous détaillerons le choix des images constituant les ensembles T et C dans la section 5.

Soient S_A et S_B l'ensemble des signatures de deux personnes A et B. La similarité entre A et B est obtenue en comparant chaque paire possible de chaines RGB contenus dans les

deux ensembles S_A et S_B afin de ne conserver que la paire la plus proche. En utilisant l'équation 2, la similarité entre A et B est définit comme suit :

$$SIM_{MvsM}(A,B) = \max_{i=1,j=1}^{N} \tilde{k}(s_{A_i}, s_{B_j}) \quad (4)$$

où N est le nombre d'images pour chaque personne.

La ré-id consiste à utiliser l'équation 4 pour calculer pour chaque requête T d'une personne inconue x une liste classée d'identités de la galerie. Ainsi, étant donné une requête T et un ensemble galerie C, la ré-id est réalisée par :

$$l(T) = \max_{i=1}^{N} (SIM_{MvsM}(T, C_i)) \quad (5)$$

où N est le nombre de sous-ensembles dans C.

4 Noyau de graphe et distance d'édition

La partie de l'image correspondant au masque de la personne détectée est segmentée en utilisant l'algorithme de fusion statistique de régions SRM [8]. Cette segmentation sert à construire un graphe d'adjacence des régions GAR. Le noyau utilisé pour la comparaison entre les graphes afin de dire si deux images représentent la même personne ou pas, est le noyau fondé sur la distance d'édition entre graphes décrit ci-dessous.

$\surd$ **Attributs :**

Chaque sommet du graphe possède les attributs suivants : la moyenne de couleur RGB de la région, la taille S (en pixels) de la région et la proportion η de la région par rapport à l'objet d'intérêt. Deux régions sont similaires si leur taille et couleur sont similaires. Les arrêtes représentent l'adjacence entre régions et ne possèdent pas d'attributs. Ce qui permet d'avoir une stabilité aux rotations de l'objet.

$\surd$ **La définition du noyau :**

Soit un graphe $G = (V, E)$ où V est l'ensemble des sommets et $E \subset V \times V$ l'ensemble des arêtes. Un sac de chemin P associé à G est défini comme un ensemble de chemins de G dont la cardinalité est noté $|P|$. Étant donné K_{path} un noyau de chemin, deux graphes G_1 et G_2 et deux chemins $h_1 \in P_1$ et $h_2 \in P_2$ respectivement de G_1 et G_2, $K_{path}(h_1, h_2)$ peut être considéré comme une mesure de similarité entre h_1 et h_2. Le but d'un sac de noyau de chemin est d'agréger les mesures locales entre les paires de chemins en une mesure de similarité globale entre deux graphes.

Le noyau entre deux chemins $h_1 = (v_1^1, \ldots, v_n^1)$ et $h_2 = (v_1^2, \ldots, v_p^2)$ vaut 0 si les deux chemins n'ont pas la même longueur, autrement il est défini comme suit :

$$K_{classic}(h_1, h_2) = \prod_{i=1}^{|h|} K_v(v_i^1, v_i^2) \quad (6)$$

Le terme K_v désigne le noyau des attributs du sommet vv. Nous définissons pour un noeud, la fonction (dans [8]) :

$$b(v) = 256\sqrt{\frac{1}{2Q\,|v|} \ln \frac{\left|\mathcal{R}_{|v|}\right|}{\delta}} \quad (7)$$

où $\delta = 1/(6\,|I|^2)$ ($|I|$ est la taille de l'image), $|v|$ est la taille du noeud, Q est un paramètre et $\mathcal{R}_{|v|}$ est l'ensemble des noeuds avec la même taille que v. Par conséquent, le coût de la contraction des arêtes est défini comme suit :

$$w_e(v_1, v_2) = \frac{\max_{k=r,g,b} \left|\overline{R}_k(v_1) - \overline{R}_k(v_2)\right|}{\sqrt{b^2(v_1) + b^2(v_2)}} \quad (8)$$

où v_1 et v_2 sont les noeuds d'extrémité de la contraction de l'arête et la fonction $\overline{R}_k(v)$ les états de la moyenne couleur du canal k dans la région représentée par le noeud v. Une valeur élevée de ce coût signifie qu'il est peu probable que les deux régions fusionnent l'une avec l'autre.

De plus, nous supposons que cet attribut est additif : le poids de deux arêtes consécutives le long d'un chemin est la somme des deux poids.

On posant κ la fonction qui applique la contraction la moins couteuse d'une arrête sur un chemin et D le nombre maximal de réductions. Les applications successives de la fonction κ associe à chaque chemin h une séquence réduite de chemins $(h, \kappa(h), \ldots, \kappa^D(h))$. Chaque $\kappa^k(h)$ est associé à un coût : $cost_k(h)$ défini comme la somme des coûts de κ opérations produisant $\kappa^k(h)$ dans h. En utilisant $K_{classic}$ pour la comparaison des chemins, nous introduisons le noyau K_{edit} comme étant la somme des noyaux entre les chemins réduits.

Étant donné deux chemins h_1 et h_2, le noyau $K_{edit}(h_1, h_2)$ est défini comme suit :

$$\frac{1}{2D} \sum_{k=0}^{D} \sum_{l=0}^{D} e^{-\frac{cost_k(h_1) + cost_l(h_2)}{2\sigma_{cost}^2}} K_{classic}(\kappa^k(h_1), \kappa^l(h_2)) \quad (9)$$

où σ_{cost} est un paramètre réglé expérimentalement.

Le noyau $K_{classic}$ est un noyau de produit tensoriel défini positif puisqu'il est un produit de noyaux définis positifs. Comme le noyau du coût de l'édition est construit à partir d'un produit scalaire il est défini positif. Ces deux derniers noyaux forment un noyau de produit tensoriel. Finalement K_{edit} est proportionnel (par un facteur $2D$) à un noyau R-convolution [5], et donc défini positif.

5 Expérimentation

Nous avons testé notre approche sur deux bases d'images publiques en ré-id (ETHZ_REID et CAVIAR4REID). La base ETHZ_REID [9] est filmée par une caméra mobile dans la rue. Cette base est divisée en trois séquences composés de 83, 35 et 28 identités respectivement. La base CAVIAR4REID [3] contient 1220 images de 72 individus : 50 d'entre eux sont filmés par deux caméras (10 images pour chaque camera par individu) et 22 individus sont filmés avec une seule caméra (10 images par individu). Les masques binaires des individus nous sont fournis pas [4]. Nous avons suivi le protocole d'évaluation utilisé par [4] et [1]

$\surd$ **Mono-échantillon :** Nous avons sélectionné aléatoirement une image par individu afin de construire l'ensemble requête, le restant des images forment l'ensemble galerie $|T_x| =$

1, $|C_x| = 1$. Pour chaque image requête la position de l'appariement correct est obtenue. Cette procédure est répétée 10 fois.

$\surd$ **Multi-échantillons :** Nous avons sélectionné aléatoirement un sous ensemble de N images pour chaque individu afin de construire l'ensemble galerie et l'ensemble requête ($|C_x| = N$, $|T_x| = N$). Pour chaque N, nous répétons l'expérience k fois ($k = 100$) afin d'obtenir des statistiques fiables. Pour la base ETHZ_REID, $N = \{2, 5, 10\}$ alors que pour CAVIAR4REID, $N = 5$.

Résultats : Dans un premier temps nous avons comparé les performances de ré-id des deux représentations chaîne RGB et GAR par distance d'édition. La chaîne RGB a montrée de meilleures performances. La courbe CMC (pour "Cumulative Match Characteristic") est utilisée pour mesurer les performances de la ré-id. Une courbe CMC représente la probabilité de trouver un appariement correct parmi les r meilleurs appariements (axe des X), pour $r = 1...k$. La variable r est appelée le rang de ré-identification. Une courbe CMC atteignant les 100% au rang 1 représente les performances maximales atteignables par un système de ré-id. Nous comparons maintenant nos performances à celles d'autres méthodes dans la littérature : SDALF [1] et SURF [6]. Ces résultats sont donnés dans le Tableau 1 où les valeurs de CMC sont fournies pour les rangs 1, 5, 10 et 15. Nous observons que nous sommes moins performant dans le cas mono-échantillon. Cela est dû au fait que la chaîne RGB est corrompue pour les fortes occultations. Cet inconvénient est atténué par l'utilisation de plusieurs images dans le cas multi-échantillons. En effet, dans le cas multi-échantillons nous observons des performances moyennes pour seq1 et seq2 alors que pour seq3 et CAVIAR4REID nos performances sont meilleures. Ceci peut être expliqué par le fait que dans l'absolu il n'y a pas de méthode parfaite. D'une manière générale, plus une méthode est robuste, moins elle est précise. Notre modèle véhicule une bonne description spatiale de la personne. L'appariement déployé cf. table 2 donne de bons résultats en présence d'occultations modérées (seq3 et CAVIAR4REID), malheureusement il n'est pas robuste aux fortes occultations et grands changements d'illumination dans seq1 et seq2. D'autres investigations sur les occultations et l'ajout d'une étape de normalisation de la couleur sont prévues dans nos futurs travaux.

TABLE 1 – Comparaison avec l'état de l'art

La base ETHZ_REID						
	seq1		seq2		seq3	
Méthode	r=1	r=5	r=1	r=5	r=1	r=5
GAR mono-échantillon	37.72	43.71	42.46	56.16	52.83	64.15
RGB mono-échantillon	36.14	56.62	**77.14**	**90**	39.28	78.57
[1] mono-échantillon	**65**	**81**	64	85	**76**	**90**
RGB multi-échantillons	62.65	85.54	77.14	**98**	85.71	**100**
[1] multi-échantillons	**90**	**94**	**90.5**	**98**	**94**	97

La base CAVIAR4REID				
Méthode	r=1	r=5	r=10	r=15
GAR mono-échantillon	31.37	42.15	51.96	53.92
RGB mono-échantillon	**16.66**	**40.27**	**51.38**	56
[1] mono-échantillon	10	25.8	45	**60**
RGB multi-échantillons	**23.16**	**56.94**	**77.77**	**87.5**
[1] multi-échantillons	18	50	68	80
[6] multi-échantillons	20	-	-	-

TABLE 2 – Exemples de mauvaises ré-identifications

seq1		seq2	
requête	résultat r=1	requête	résultat r=1
			

6 Conclusion

Cet article présente une approche de ré-id de personnes dans un système de vidéosurveillance. Nous avons proposées deux descriptions structurelles pour fournir une signature discriminante pour chaque personne. Pour chaque signature, nous proposons un noyau approprié qui est utilisé pour la tâche de ré-id. Les résultats sont prometteurs. Pour les travaux futurs, nous prévoyons d'utiliser une représentation structurelle plus complexe afin de garder plus d'informations sur le modèle d'apparence et ainsi avoir de meilleurs résultats.

Références

[1] Loris BAZZANI, Marco CRISTANI et Vittorio MURINO. "Symmetry-driven accumulation of local features for human characterization and re-identification". In : *Comput. Vis. Image Underst.* 117.2 (fév. 2013), p. 130-144.

[2] Marco CUTURI. "Fast Global Alignment Kernels". In : *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Sous la dir. de Lise GETOOR et Tobias SCHEFFER. ICML '11. Bellevue, Washington, USA : ACM, juin 2011, p. 929-936.

[3] Cheng DONG SEON et al. "Custom Pictorial Structures for Re-identification". In : *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2011, 68.1—68.11.

[4] M. FARENZENA et al. "Person re-identification by symmetry-driven accumulation of local features". In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Juin 2010, p. 2360 -2367.

[5] David HAUSSLER. *Convolution Kernels on Discrete Structures*. Technical Report UCS-CRL-99-10. Department of Computer Science, University of California at Santa Cruz, 1999.

[6] Mohamed IBN KHEDHER, Mounim EL YACOUBI et Bernadette DORIZZI. "Fusion of appearance and motion-based sparse representations for multi-shot person re-identification". In : *Neurocomputing* 248 (juil. 2017), p. 94 -104.

[7] Amal MAHBOUBI et al. "Tracking System with Re-identification Using a RGB String Kernel". In : *Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, August 20-22, 2014. Proceedings*. 2014, p. 333-342.

[8] Richard NOCK et Frank NIELSEN. "Statistical Region Merging". In : *IEEE Trans. Pattern Anal. Mach. Intell.* 26.11 (nov. 2004), p. 1452-1458.

[9] W.R. SCHWARTZ et L.S. DAVIS. "Learning Discriminative Appearance-Based Models Using Partial Least Squares." In : *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*. IEEE Computer Society, 2009, p. 322-329.