Ré-identification de personne avec perte min-triplet

Yiqiang Chen[1] Stefan Duffner[1] Andrei Stoian[2] Jean-Yves Dufour[2] Atilla Baskurt[1]

[1] INSA-Lyon, LIRIS, UMR5205,F-69621, France
[2] Thales Services, ThereSIS, Palaiseau, France

20, Avenue Albert Einstein 69621 Villeurbanne cedex, France
yiqiang.chen@insa-lyon.fr

Résumé

Dans cet article, nous proposons une méthode de ré-identification de personnes basée sur un réseau de neurones en triplet. La ré-identification de personnes consiste à reconnaître un individu parmi un ensemble de personnes acquises dans différentes conditions par une ou plusieurs caméras avec champs de vue non-recouvrant. Pour cela, notre système extrait d'abord les caractéristiques visuelles invariantes aux différents points de vue, et puis effectue l'apprentissage de métrique par un réseau de neurones en triplet. Nous proposons une nouvelle fonction de coût pour le réseau de neurones appelée la perte min-triplet. Nous comparons nos résultats à l'état de l'art sur plusieurs bases de données publics.

Mots Clef

Ré-identification de personne, réseau de neurones, apprentissage de métrique

Abstract

In this paper, we propose a person re-identification method based on a triplet neural network. The person re-identification problem consists to find one identity among a set of people in different places with one or several non-overlapping surveillance cameras. In order to do this, our system extracts first viewpoint invariant features, then uses a triplet neural network to learn a metric. We propose a new loss function for this neural network called min-triplet loss. Finally, we compare our results with the state-of-art on several public person re-identification benchmarks.

Keywords

Person re-identification, neural network, metric learning

1 Introduction

L'analyse automatique de vidéos de surveillance est un domaine de recherche important et concurrentiel afin d'exploiter une grande quantité de données produites par les caméras de surveillance. Dans ce contexte, la ré-identification est l'un des enjeux importants. Son objectif est de retrouver la même personne dans des vidéos enregistrés dans différentes zones du même lieu public durant une période de temps réduite. En outre, la ré-identification est également nécessaire dans des applications comme l'interaction humain-machine, ou l'indexation de contenu de vidéo, etc.

FIGURE 1 – A gauche : l'image d'une personne requête. A droite : la classement des images de galerie en fonction de la similarité. La bonne correspondance qui est dans la rectangle verte, est difficile à retrouver, même pour un humain, à cause de changement de pose et de luminosité.

La ré-identification est une problématique difficile. L'apparence de la personne subit des variations significatives tels que la variation de poses, la variation de point de vue et la variation de l'éclairage (Fig.1). La résolution de la vidéo est généralement basse et des occultations sur les personnes sont fréquentes. Les visages ne sont souvent pas visibles. Donc, les méthodes basées sur la biométrie comme la reconnaissance faciale ne sont pas applicables.

Etant donnée une image de requête, pour trouver la bonne correspondance de cet individu dans un grand ensemble d'images de galerie, il faut prendre en compte deux problèmes. D'abord, il est nécessaire d'avoir une représentation caractéristique des images de requête et des images de galerie. Deuxièmement, une métrique de distance est nécessaire pour déterminer si une image de requête et une image de galerie appartiennent à la même classe d'individus.

Dans cet article, nous nous basons sur les caractéristiques LOMO (Local Maximal Occurrence) proposées par [1] qui montrent une bonne robustesse aux changements de point de vue. Ensuite, comme contribution, nous proposons d'appliquer un réseau de neurones en triplet à la ré-identification pour l'apprentissage de métrique. Le réseau de neurones en triplet prend un triplet d'exemples en en-

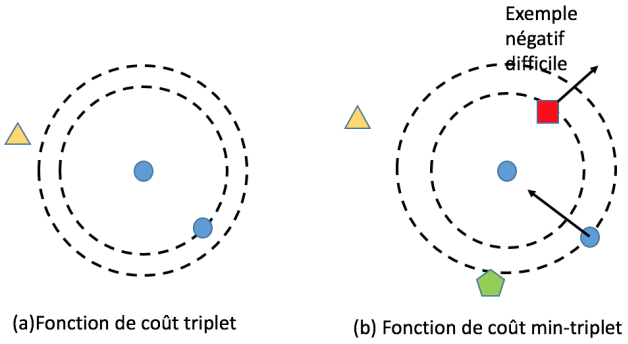

FIGURE 2 – (a) Avec la fonction de coût triplet, il est facile de tomber sur un exemple négatif "facile". Ceci n'apporte rien à la mise à jour des poids du réseau. (b) la fonction de coût min-triplet permet de choisir les exemples négatifs difficiles et d'apprendre plus efficacement.

trée et apprend efficacement une métrique en éloignant les exemples de différentes classes et approchant les exemples de la même classe dans un triplet. Nous proposons également une nouvelle fonction de coût appelée la perte de min-triplet. Notre réseau utilisant cette fonction de coût min-triplet prend un exemple de référence, un exemple positif et plusieurs négatifs en entrée au lieu d'un triplet. La fonction de coût éloigne l'exemple négatif le plus difficile (Figure 2) c'est-à-dire celui le plus proche de la référence. Ceci permet de rendre l'apprentissage plus efficace et plus performant. A la fin, nous montrons que notre approche permet d'obtenir des résultats équivalents à l'état de l'art sur deux bases de données publiques.

2 Travaux antérieurs

Un système de ré-identification est généralement composé de deux parties : une extraction de caractéristiques pour décrire une image de requête et des images de galerie ; une métrique de distance pour comparer le vecteur caractéristique des images. Les travaux antérieurs consistent soit à construire une représentation caractéristique robuste, soit à chercher une meilleure métrique de similarité.

Un grand nombre de travaux ont été proposés pour l'extraction de caractéristiques. Gray et Tao [2] ont proposé d'utiliser Adaboost pour sélectionner les bonnes caractéristiques dans un ensemble de caractéristiques de couleurs et de textures. Farenzena et al. [3] ont proposé la méthode Symmetry-Driven Accumulation of Local Features (SDALF) dans laquelle la symétrie et l'asymétrie de l'image de personne sont prises en compte pour traiter le problème de point de vue. Ma et al. [4] ont encodé les caractéristiques locales comme un vecteur de Fisher pour créer une représentation globale d'une image. Cheng et al. [5] ont appliqué l'approche de "pictorial strcture" pour localiser les parties du corps et extraire les couleurs. Liao et al. [1] ont appliqué une normalisation de couleur basée sur l'algorithme de Retinex pour produire une image avec des couleurs plus cohérentes. Les caractéristiques de couleurs et de textures sont ensuite extraites dans une fenêtre glissante. Pour avoir une invariance horizontale dans l'image, on garde seulement la valeur maximale de chaque dimension du vecteur caractéristique parmi les fenêtres sur la même bande horizontale.

Dans l'étape de mesure de similarité, pour la ré-identification de personne, la distance de Mahalanobis est la plus utilisée. Dans ce cas, nous cherchons à apprendre un sous-espace qui est défini par une matrice de projection W dans lequel les distances entre les exemples (y_i et y_j) reflètent mieux les similarités entre les personnes, c'est-à-dire $||y_i - y_j||_2 = (x_i - x_j)^T A(x_i - x_j)$ où $A = W^T W$. Li et al. [6] ont proposé d'apprendre une fonction Locally-Adaptive Decision (LADF) qui pourrait être considérée comme une fusion d'une distance métrique et d'un seuillage localement adapté. Koestinger et al. [7] ont présenté une métrique appelée KISSME (keep it simple and straightforward metric). La décision "une paire est similaire" est formulée comme un test de ratio de ressemblance. La distance est basée sur la différence entre l'inverse de la matrice de covariance pour les paires similaires et celle des paires dissimilaires. Liao et al. [1] ont proposé d'apprendre un sous-espace dans lequel la variance des paires similaires est minimisée et celle des paires dissimilaires est maximisée, et puis d'appliquer la métrique KISSME dans cet espace réduit.

Suivant le grand succès de l'apprentissage profond dans le domaine de vision par ordinateur, des méthodes basées sur les réseaux de neurones à convolution ont été proposées pour la ré-identification de personnes. Yi et al. [8] ont construit un réseau de neurones siamois à convolution. L'image est découpée en trois parties horizontales. Chaque partie est associée à un réseau de neurones à convolution. A la fin, les trois parties sont fusionnées au niveau de leurs scores. DeepReID [9] ont proposé une nouvelle architecture de réseau Filter Pairing Neural Network (FPNN). Ils ont utilisé une couche de "patch matching" pour modéliser le déplacement des parties du corps. Ahmed et al. [10] ont proposé un réseau qui a une paire d'images en entrée et un score de similarité en sortie. Dans leur réseau, ils ont calculé la différence de cartes caractéristiques pour capturer la relation locale entre deux images d'entrée. Cheng et al. [11] ont proposé un réseau en triplet basé à la fois sur le corps entier et les parties du corps avec une fonction coût améliorée.

Malgré la puissance de l'apprentissage profond, les méthodes de CNN ont des difficultés quand la taille de la base d'apprentissage est réduite. Dans cet article, nous nous sommes inspirés des idées des approches CNN siamois. Mais au lieu d'apprendre les couches convolutions, nous utilisons les caractéristiques de bas niveau. Nous apprenons une métrique par un réseau peu profond et similaire au réseau de neurones en triplet avec une nouvelle fonction de coût. Nous montrons que notre approche obtient de bons résultats sur les bases de données de taille

limitée.

3 Réseau de neurone siamois

Notre approche est une variante du réseau de neurones siamois. Dans cette section, nous revisitons le réseau de neurones siamois.

Le réseau de neurones siamois tire son nom de sa stratégie d'apprentissage, qui met en œuvre plusieurs copies identiques du même réseau en parallèle (Fig.3). Les sous-réseaux partagent les mêmes poids. Dans sa version originale, introduite par Bromley et al. [12] pour la vérification de signatures, et adaptée par Chopra et al. [13] pour la vérification de visage, le réseau encode la similarité entre exemples à l'aide d'une métrique contrôlée dans l'espace de sortie et appliquée à des paires d'échantillons. Une fonction d'erreur est alors définie de telle manière que la similarité intra-classe soit forte, et la similarité inter-classe soit faible.

Deux principales approches sont identifiées : l'approche par paires et l'approche par triplets. L'approche par paires est la stratégie la plus commune : une relation de similarité est encodée à l'aide de deux échantillons, ainsi qu'une étiquette de similarité précisant si ces échantillons sont similaires ou dissimilaires. La perte contrastive est la fonction d'erreur la plus commune.

Avant d'introduire la méthode, nous définissons quelques notations. Sachant respectivement une paire de vecteur **a** et **b** et un label **y** qui est 1 si **a** et **b** sont de la même classe, sinon 0. Les deux sous-réseaux avec les poids partagés donnent les deux sortis f(a) et f(b). La distance euclidienne est utilisée comme la métrique dans l'espace projeté. Donc la perte contrastive pénalise le réseau lorsque la paire positive se trouve à une distance au-delà de la marge ou bien la paire négative se trouve à une distance faible. La fonction se formule comme :

$$E_{contrastive} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\|f(a_i) - f(b_i)\|_2 + (1-y_i)max(m - \|f(a_i) - f(b_i)\|_2), 0)], \quad (1)$$

où m est le marge, et N est le nombre de paire d'exemples. L'approche par triplets, proposée par Lefèbvre et al. [14], étend cette information en introduisant une représentation plus symétrique entre similarité et dissimilarité. Ainsi, un ensemble d'apprentissage est formé de trois échantillons, avec un échantillon de référence **a**, un échantillon dit positif **p**, similaire à la référence, et un échantillon négatif **n**, dissimilaire de la référence. Donc la fonction d'erreur de triplet pénalise le réseau lorsque la distance entre la référence et l'exemple négatif est inférieure à la distance entre la référence et l'exemple positif plus une marge. La fonction se formule comme :

$$E_{triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i) - f(p_i)\|_2^2 - \|f(a_i) - f(n_i)\|_2^2 + m, 0)] \quad (2)$$

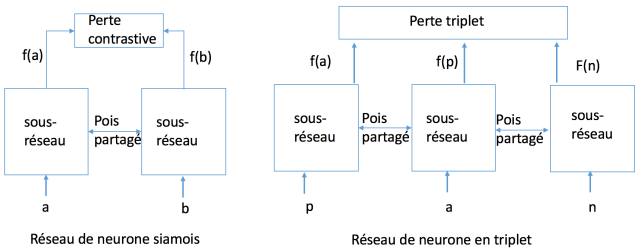

FIGURE 3 – L'architecture du réseau de neurone siamois et en triplet.

4 Méthode proposée : perte de min-triplet

Dans notre approche, au lieu de projeter un triplet dans l'espace caractéristique, on y projete (n+2)-uplet dont un exemple de référence **a**, un exemple positive **p** qui est dans la même classe que l'exemple de référence et n exemples négatifs n^j. Le tuplet est formé aléatoirement. La distance entre l'exemple de référence et l'exemple positif doit être inférieure à la distance entre l'exemple de référence et le plus proche exemple négatif. Nous définissons la contrainte suivante :

$$min(\|f(a) - f(n^j)\|_2^2) - \|f(a) - f(p)\|_2^2 > m \quad (3)$$

L'idée est similaire à la distance de Mahalanobis de [15] et à la sélection de triplet de Facenet [16] utilisée dans la reconnaissance faciale. L'exemple qui a la distance minimale est l'exemple le plus important dans le tuplet, puisque cet exemple a le plus d'information sur la direction qui a le plus de potentiel à être amélioré. Une grande amélioration aurait lieu si l'exemple le plus "faux" est corrigé. De plus, intuitivement, pour classer la bonne correspondance dans l'ensemble de galerie en premier rang, il suffit de la classer devant l'exemple non correspondant mais le plus ressemblant à la requête.

Il serait trop couteûx en calcul de trouver l'exemple négatif le plus difficile sur toute la base de données. Et si on utilise une sélection de triplets aléatoire, une partie des triplets satisfont la contrainte donc la convergence est ralentie. Un bon compromis est donc d'utiliser l'apprentissage sur (n+2)-uplet. Il est plus facile d'avoir un exemple qui ne satisfait pas la contrainte Eq. (3) et le coût supplémentaire de calcul est réduit. Ceci permet également d'assurer une convergence stable et rapide.

Pour rendre la fonction de coût plus robuste, similaire à [11], nous ajoutons un terme dans la perte Eq. (2) qui est la distance entre l'exemple de référence et l'exemple positif dans la fonction de coût. La première partie de perte concerne seulement une comparaison relative. Il donne seulement une règle relative dans l'espace de caractéristiques. Le terme ajouté correspond à une distance absolue dans l'espace. Ceci permet de mieux positionner les exemples dans l'espace caractéristique et de faire rapprocher les ex-

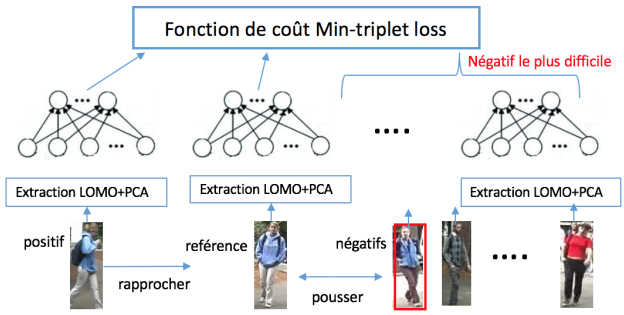

FIGURE 4 – Vue d'ensemble de notre système de ré-identification de personne.

emples de même classe. La fonction de perte finale est définie pour :

$$E_{min-triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i) - f(p_i)\|_2^2 - min(\|f(a_i^j) - f(n_i^j)\|_2^2) + m, 0) + \alpha\|f(a_i) - f(p_i)\|_2] \quad (4)$$

5 Résultats expérimentaux

5.1 Bases de données

Nous évaluons notre approche sur trois benchmarks publics.

VIPeR Gray et al. proposent dans [17] la base d'images publique VIPeR (pour "Viewpoint Invariant Pedestrian Recognition"), le premier jeu de données dédié à la ré-identification de personnes. Cette base est constituée de 632 personnes, vues une fois chacune de deux caméras différentes. Nous suivons le protocole de test utilisé dans [18]. Nous choisissons aléatoirement 100 personnes pour l'apprentissage et le reste pour le test. Dans le test, les images d'une vue sont utilisées comme requêtes et celles de l'autre vue sont utilisées comme galerie.

Nous avons employé le protocole de test avec 100 personnes pour l'apprentissage au lieu de celui proposé par l'auteur [17] (316 pour l'apprentissage, 316 pour le test) pour montrer que notre approche pourait avoir une bonne performance avec peu d'exemples d'apprentissage.

GRID [19] La base de données GRID contient des images captées par 8 caméras disjointes dans une station de métro. Il contient 250 paires d'images et 775 images complémentaire qui n'appartient pas aux 250 personnes. Nous utilisons la partition de l'apprentissage et de test fournie dans la base de données. Nous prenons 125 personnes pour l'apprentissage et le reste pour le test. Les 775 personnes qui ont seulement une image sont utilisées comme les images de galeries.

PRID2011 [20] Cette base de données contient des séquences d'images acquises par deux caméras de surveillance statiques dans une rue. Nous utilisons la version

FIGURE 5 – Exemples d'images dans les base de données VIPeR, GRID, PRID

"single-shot" de cette base. Les vues de caméras A et B contiennet respectivement 385 et 749 personnes dont 200 personnes apparaissant dans les deux vues. Nous sélectionnons aléatoirement 100 personnes pour l'apprentissage et nous prenons la vue de caméra A comme l'ensemble de requêtes. Donc pour le test, nous avons 100 requêtes et 649 images de galerie.

5.2 Caractéristiques visuelles utilisées et paramétrage

Dans notre approche, on utilise les caractéristiques LOMO pour la représentation d'images de personnes. Les caractéristiques LOMO ont une bonne robustesse aux changements de point de vue. Le descripteur LOMO a 26960 dimensions. Nous réduisons la dimensionnalité à 200 par PCA. Nous employons un réseau de neurones avec une couche de 600 dimensions en sortie. Les poids sont initialisés par une distribution Gaussienne. Nous fixons la marge m = 1 et le coefficient α =0.002. Dans les expériences, nous testons le nombre d'exemples négatifs dans le tuplet de n = 1 et de n= 5.

5.3 Mesure d'évaluation

Nous employons les courbes CMC (Cumulative Match Characteristic) qui sont la mesure d'évaluation la plus utilisée dans la littérature. Nous cherchons la correspondance de chaque image de requête dans l'ensemble de galerie. Ensuite, nous classons les images de galerie en fonction de leur distance. Nous regardons les k images les plus proches. Si la bonne correspondance est dans ces k images, le rang de test k est un succès. Nous répétons 10 fois l'expérience avec des partitions différentes de l'apprentissage et de test. La moyenne représente le résultat final de l'évaluation.

5.4 Comparaison et analyse des résultats

Nous comparons le résultat de notre système avec les résultats de l'état de l'art. Nous avons également implémenté deux méthodes de baseline pour la comparaison. Ces deux baselines utilisent respectivement la fonction de coût triplet classique Eq.(2) et la fonction de coût contrastive Eq. (1) indiquée dans la section 3. Pour la fonction de coût triplet, les triplets sont formés aléatoirement. Pour la fonction de coût contrastive, on génère aléatoirement une moitié de paires positives et une moitié de paires négatives.

Sur la base VIPeR (Tab. 1) et PRID (Tab. 3), nous améliorons le résultat de l'état de l'art. Ceci montre la bonne performance de notre système. Notre approche a une fonction de coût similaire à la meilleure approche dans l'état de l'art sur PRID TCP[11] qui est basé sur le réseau de neurones à convolutions et propose d'améliorer la fonction de coût triplet par la distance absolue. La meilleure performance de notre approche par rapport à TCP sur PRID montre l'efficacité de la combinaison des caractéristiques LOMO et du réseau de neurones peu profond sur des bases de données de taille limitée. Notre système a beaucoup moins de paramètres par rapport à TCP. C'est-à-dire notre approche est plus efficace en terme de calcul et a moins de risque de sur-apprentissage. Sur GRID (Tab. 2), nous avons des résultats meilleurs que la plupart des résultats précédents sauf la méthode SCSP[21]. LOMO contient deux types de caractéristiques à trois échelles des régions locales. L'approche SCSP extrait plusieurs types de caractéristiques de couleurs et de textures dans les régions locales et aussi au niveau de l'image globale. Cette représentation riche pourrait être la raison de sa bonne performance. Nous avons une amélioration 1 et 1.5 points en rang 1sur la base GRID et PRID et 1 point en rang 5 sur toutes les 3 bases de données en augmentant le nombre d'exemples négatifs dans le tuplet. Ceci montre l'efficacité de l'apprentissage avec la perte min-triplet. Nous n'avons pas eu une amélioration en rang 1 sur la base VIPeR. Ceci est probablement dû aux grandes variations de pose dans la base VIPeR (Figure 5). Donc dans VIPeR, plus d'exemples négatifs pourraient être considérés "difficiles" dans l'apprentissage. Donc trouver un seul exemple dans un tuple sert moins pour cette base. Par rapport aux baselines, la perte min-triplet obtient un meilleur résultat avec un point ou plus en rang 5. Ceci montre que le fait de combiner la comparaison de distance relative et la comparaison de distance absolue est efficace pour améliorer l'apprentissage.

6 Conclusion

Nous avons proposé un système de ré-identification en utilisant un réseau de neurones avec une nouvelle fonction de coût min-triplet. Dans les expériences, nous avons montré que cette fonction de coût peut améliorer le résultat de ré-identification. En utilisant la caractéristiques LOMO, nous avons montré que notre système est performant sur les bases de données de taille limitées et nous obtenons des résultats équivalents à l'état de l'art sur ces benchmarks.

p=532	rank=1	rank=5	rank =10	rank =20
PCCA[22]	9.3	24.9	37.4	52.9
RPML[23]	10.9	26.7	37.7	51.6
LADR[24]	12.9	30.3	42.7	58.0
FeatureMap[18]	17.4	41.6	55.3	70.8
LOMO+siamois	20.2	43.7	56.9	70.5
LOMO+triplet	17	39.2	51.8	65.9
LOMO+min-triplet(n=1)	**20.7**	43.4	56.2	70.2
LOMO+min-triplet(n=5)	**20.7**	**44.5**	**57.5**	**71.2**

TABLE 1 – Pourcentage de bonnes correspondances sur la base de données VIPeR. La taille de l'ensemble de galerie est 532.

p=900	rank=1	rank=5	rank =10	rank =20
PRDC[25]	9.7	22.0	33.0	44.3
RankSVM[26]	10.2	24.6	33.3	43.7
MRank-PRDC[19]	11.1	26.1	35.8	46.6
MRank-rankSVM[19]	12.2	27.8	36.3	46.6
RQDA[27]	15.2	30.1	39.2	49.3
FeatureMap[18]	16.3	35.8	46.0	57.6
LOMO+XQDA[1]	16.6	33.8	41.8	52.4
SCSP[21]	**24.2**	**44.6**	**54.1**	**65.2**
LOMO+siamois	17.7	34.8	44.3	55.1
LOMO+triplet	13.8	27.6	37.5	48.9
LOMO+min-triplet(n=1)	16.6	34.5	43.7	55.3
LOMO+min-triplet(n=5)	18.1	35.8	44.2	55.0

TABLE 2 – Pourcentage de bonnes correspondances sur la base de données GRID. La taille de l'ensemble de galerie est 900.

p=649	rank=1	rank=5	rank =10	rank =20
LMNN[28]	10	30	42	59
LMNN-R[28]	9	32	43	60
KISSME[29]	15	39	52	68
Mahananobis[30]	16	41	51	64
EIML[31]	16	39	51	68
deepMetric[8]	18	46	55	71
TCP[11]	**22**	47	57	76
LOMO+siamois	21	48	60	79
LOMO+triplet	19	49	61	78
LOMO+min-triplet(n=1)	21	50	62	77
LOMO+min-triplet(n=5)	**22**	**51**	**63**	**79**

TABLE 3 – Pourcentage de bonnes correspondances sur la base de données PRID. La taille de l'ensemble de galerie est 649. Les résutlats de plusieurs travaux antéreurs sont en entier. Nos résultats sont arrondis à l'entier le plus proche pour comparer.

Remerciement

Ce travail est financé par le Groupe Image Mining qui regroupe les chercheurs du laboratoire LIRIS et du groupe THALES dans la vision d'ordinateur et la fouille de données.

Références

[1] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2197–2206.

[2] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized

features," in *Proceedings of the IEEE International Conference on European Conference on Computer Vision (ECCV)*. Springer, 2008, pp. 262–275.

[3] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2360–2367.

[4] Bingpeng Ma, Yu Su, and Frédéric Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proceedings of the IEEE International Conference on European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 413–422.

[5] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino, "Custom pictorial structures for re-identification.," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2011, vol. 2, p. 6.

[6] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3610–3617.

[7] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2288–2295.

[8] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Deep metric learning for person re-identification," in *Proceedings of the IEEE International Conference on International Conference on Pattern Recognition*. IEEE, 2014, pp. 34–39.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid : Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 152–159.

[10] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3908–3916.

[11] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1335–1344.

[12] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah, "Signature verification using a "siamese" time delay neural network," *IJPRAI*, vol. 7, no. 4, pp. 669–688, 1993.

[13] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005, vol. 1, pp. 539–546.

[14] Grégoire Lefebvre and Christophe Garcia, "Learning a bag of features based nonlinear metric for facial similarity," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS)*. IEEE, 2013, pp. 238–243.

[15] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng, "Top-push video-based person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1345–1353.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet : A unified embedding for face recognition and clustering," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[17] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proceedings of International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*. Citeseer, 2007, vol. 3.

[18] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1565–1573.

[19] Chen Change Loy, Chunxiao Liu, and Shaogang Gong, "Person re-identification by manifold ranking," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2013, pp. 3567–3571.

[20] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

[21] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1268–1277.

[22] Alexis Mignon and Frédéric Jurie, "Pcca : A new approach for distance learning from sparse pairwise

constraints," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2666–2672.

[23] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proceedings of the IEEE International Conference on European Conference on Computer Vision (ECCV)*. Springer, 2012, pp. 780–793.

[24] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3594–3601.

[25] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 649–656.

[26] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking.," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2010, vol. 2, p. 6.

[27] Shengcai Liao, Yang Hu, and Stan Z Li, "Joint dimension reduction and metric learning for person re-identification," *arXiv preprint*, 2014.

[28] Kilian Q Weinberger and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[29] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2288–2295.

[30] Peter M Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, pp. 247–267. Springer, 2014.

[31] Martin Hirzer, Peter M Roth, and Horst Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance(AVSS)*. IEEE, 2012, pp. 203–208.