PERSON RE-IDENTIFICATION BASED ON HIERARCHICAL BIPARTITE GRAPH MATCHING

Yan Huang, Hao Sheng, Zhang Xiong

State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China

ABSTRACT

This work proposes a novel person re-identification method based on Hierarchical Bipartite Graph Matching. Because human eyes observe person appearance roughly first and then goes further into the details gradually, our method abstracts person image from coarse to fine granularity, and finally into a three layer tree structure. Then, three bipartite graph matching methods are proposed for the matching of each layer between the trees. At the bottom layer Non-complete Bipartite Graph matching is proposed to collect matching pairs among small local regions. At the middle layer Semi-complete Bipartite Graph matching is used to deal with the problem of spatial misalignment between two person bodies. Complete Bipartite Graph matching is presented to refine the ranking result at the top layer. The effectiveness of our method is validated on the CAVIAR4REID and VIPeR datasets, and competitive results are achieved on both datasets.

Index Terms— person re-identification, cross views, bipartite graph matching

1. INTRODUCTION

The task of person re-identification under non-overlapping smart surveillance scenario is an important application in the area of computer vision since the smart surveillance systems still cannot comprehensively monitor every corner of the cities due to the cost and limit computational resources [1]. Therefore, person re-identification has become an essential part to track target's trajectory throughout the whole journey.

Normally the notable challenges in person re-identification include finding robust vision features and designing effective distance metric models, thereby integrating the two components together to reach a final decision. Recently, through the study of how human eyes observe person appearance, it is revealed that human eyes can recognize person identities based on salient regions. As a result some researchers converted person re-identification into a salient feature matching problem [2, 3]. Moreover, due to the fact that human eyes can concentrate on differences between small regions, a corresponding patch pairs' feature matching approach is proposed in [4]. Considering human visual ambiguities shared between the first ranks, a ranking optimization approach via discriminant context information analysis is presented in [5].

However, most of current approaches only discuss a single aspect of vision information matching and leave rich information not fully exploited. Therefore, multi information fusion from the perspective of human eyes is important in practice. In this paper, we proposed a hierarchical modeling and matching schema inspired by the fact that human eyes observe person appearance roughly first and then goes further into the details gradually. The hierarchical representation has been employed for person re-identification [6, 7, 8], but most of current approaches only change the scale in different hierarchies with fixed vision information. Different from previous attempts, our hierarchies are used to deal with different vision matching problem with various vision information.

In this research, a three-layer tree structure is firstly modeled on every person's image, which is divided from coarse to fine granularity. After the hierarchical model is obtained, the person re-identification issue is converted into the trees' matching problem between two forests: one forest containing a number of person tree is called probe forest, and the other is called gallery forest. Secondly, in order to solve the issue of trees' matching, a novel Hierarchical Bipartite Graph Matching (HBGM) method is proposed, which contains three bipartite graphs (bi-graphs) to respectively deal with three kinds of vision matching problems with different vision information on each layer between two trees.

The graph based method has gained widespread use in many vision problem [9, 10, 11]. Among them, A bi-graph is a graph in which vertices can be divided into two disjoint sets U and V, such that every edge connects a vertex in U to one in V. At the bottom layer of HBGM, each node represents a small local region. Non-complete Bi-graph matching (NBM) is proposed to collect matching pairs used in bottom layer nodes matching. At the middle layer, each node represents a set which consist of bottom layer nodes, and the spatial positions of bottom layer nodes in image determine the mapping relationship between these two layers. To reduce spatial misalignment between two person bodies, Semi-complete Bi-graph matching (SBM) is used in the middle layer nodes matching. Finally, Complete Bi-graph matching (CBM) is presented to refine the ranking result at the top layer.

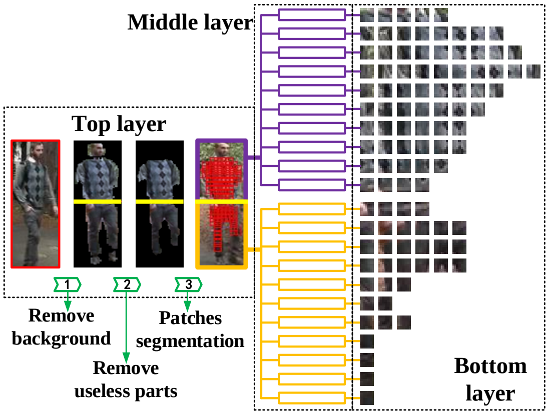


Fig. 1. Tree structure modeling on person image

2. HBGM-ORIENTED MODELING

We first introduce a tree structure to model each person image. Fig.(1) shows that the tree consists of three layers and divides the person image into small local regions layer by layer. The Deep Decompositional Network [12] is used to remove background and get boundary between upper and lower body. Due to less discriminative power, head and foot parts are removed from the foreground. Finally, the upper and lower body are segmented respectively into horizontal stripes with size $8 \times width$ and sample step 4. Local patches are divided on each stripe with size 8×8 and sample step 4. A 672-dimensional dColorSIFT feature which is commonly used as small patch descriptor [2, 13] is extracted on each patch. The nodes of top, middle and bottom layer respectively represent original image, horizontal stripes and local patches.

The top layer nodes in probe and gallery forests are denoted as $\mathbf{T}_{i_1}^P|_{i_1=1}^{n_1}$ and $\mathbf{T}_{i_2}^G|_{i_2=1}^{n_2}$; The middle layer nodes which are the child nodes of the $\mathbf{T}_{i_1}^P$ and $\mathbf{T}_{i_2}^G$ are denoted as $\mathbf{M}_{j_1}^{\mathbf{T}_{i_1}^P}|_{j_1=1}^{m_1}$ and $\mathbf{M}_{j_2}^{\mathbf{T}_{i_2}^G}|_{j_2=1}^{m_2}$; The bottom layer nodes which are the child nodes of the $\mathbf{M}_{j_1}^{\mathbf{T}_{i_1}^P}$ and $\mathbf{M}_{j_2}^{\mathbf{T}_{i_2}^G}$ are denoted as $\mathbf{B}_{k_1}^{\mathbf{T}_{i_1}^P,\mathbf{M}_{j_1}}|_{k_1=1}^{l_1}$ and $\mathbf{B}_{k_2}^{\mathbf{T}_{i_2}^P,\mathbf{M}_{j_2}}|_{k_2=1}^{l_2}$.

The structure of the HBGM method is shown in Fig.(2), where the person re-identification is converted into a tree' matching problem between two forests. Three bi-graphs are built at three layers between two forests and different layers are correlated with each other in the tree structure modeling.

3. TREES' MATCHING USING HBGM METHOD

Given a bi-graph with node sets U and V, the definition of complete, semi-complete and non-complete bi-graph is based on the matching between U and V. In CBM, every node of U is connected to every node of V. The SBM is defined in the way that every node in U is connected to a fixed number of nodes in V. At last, NBM is defined in the way that every node in U is connected to any number of nodes in V.

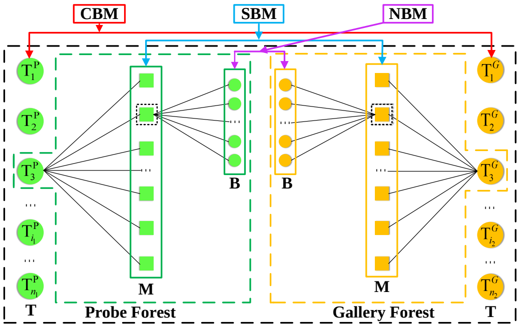


Fig. 2. Tree structure modeling on person image

3.1. The Middle Layer Matching Using SBM

The middle layer nodes matching has direct impact on the matching of the bottom layer nodes. Therefore we introduce the middle layer matching before the bottom layer. Due to the fact that person bodies in different images are not in alignment in the vertical direction, dealing with the problem of spatial misalignment is critical. As a result SBM is designed to handle this problem and a δ slack variable would be used to relax the search range of matching in SBM. Given a middle layer node in probe forest $\mathbf{M}_{j_1}^{\mathbf{T}_{i_1}^P}$, its matching nodes ($\mathbb{N}_{\mathbf{M}}$) in gallery forest are as follows:

$$\begin{aligned}\mathbb{N}_{\mathbf{M}}(\mathbf{M}_{j_1}^{\mathbf{T}_{i_1}^P}) = \left\{\mathbf{M}_{j_2}^{\mathbf{T}_{i_2}^G} | j_2 = j_1 - \delta, ..., j_1 + \delta\right\},\\ s.t. \quad j_2 - \delta \geqslant 0, j_2 + \delta \leqslant m_2\end{aligned} \tag{1}$$

Fig.(3) shows the result of SBM. The relationship between middle and bottom layer node is shown on the right side of figure. The δ slack variable determines the number of matches. If δ is too small, a probe middle layer node may not find correct match due to the person body is misalignment in vertical direction. If δ is too large, redundant matches will cause more mismatch in NBM, which will be discussed in the following section 3.2. At last, $\delta = 2$ is chosen in our experiment.

3.2. The Bottom Layer Matching Using NBM

In order to find the matching patch pairs, we proposed a NBM method with cost matrix M_c, and the result of NBM can be achieved using Hungarian algorithm in M_c. The M_c is used to calculate similarity between any two bottom layer nodes in $\mathbf{M}_{j_1}$ and $\mathbb{N}_{\mathbf{M}}(\mathbf{M}_{j_1})$, which is defined as follows:

$$\begin{aligned} M_c &= [d_{k_1,k_2}^{tst}(f(\mathbf{B}_{k_1}^{\mathbf{T}_{i_1}^P,\mathbf{M}_{j_1}}), f(\mathbf{B}_{k_2}^{\mathbf{T}_{i_2}^G,\mathbb{N}_{\mathbf{M}}(\mathbf{M}_{j_1})}))]_{l_1 \times l_2} \\ &= \begin{bmatrix} d_{1,1}^{tst} & ... & d_{1,l_2}^{tst} \\ ... & ... & ... \\ d_{l_1,1}^{tst} & ... & d_{l_1,l_2}^{tst} \end{bmatrix}_{l_1 \times l_2} \quad i_1 \forall 1, ..., n_1; i_2 \forall 1, ..., n_2 \end{aligned} \tag{2}$$

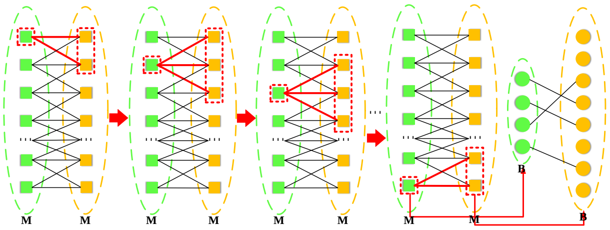


Fig. 3. Middle layer node matching using SBM. One middle layer node contains a number of bottom layer nodes shown in the right side

where d^{tst} represents the cost function used to calculate the similarity between two bottom layer nodes, f represents the dColorSIFT feature. Due to the fact that person images with the same identity are rarely collected in exactly the same environment, which leads to huge intra-person variations including illumination, background, pose, viewpoint, etc. Locally-adaptive decision function (LADF) learning algorithm [14] is used to calculate the cost function d^{tst}.

Given training data (bottom layer node matching pairs collected from any two person images with the same identity label in training set), the LADF trains a metric model for d^{tst}. Since bottom layer nodes represent small local regions and artificial matching pairs collection is impractical which needs heavy manual annotation, collecting training data becomes an important issue. Therefore, a graph degree linkage (GDL) clustering algorithm [15] is used to divide the mixed vision pattern in order to automatically find the training data. We denote all the bottom layer nodes between two images with the same identity as $S_{\mathbf{B}}^{upper}$ and $S_{\mathbf{B}}^{lower}$, where *upper* and *lower* represent different body parts. The GDL algorithm is used to divide the $S_{\mathbf{B}}^{upper}$ into two clusters, and then each sub-cluster is divided into two sub-sub-clusters. Due to the different discriminative power, the $S_{\mathbf{B}}^{lower}$ is divided into three clusters. We also need M_c to calculate the similarity between two bottom layer nodes on training set. Different from Equ.(2), the cost function is determined by our visual clustering and a Gaussian function, which is defined as:

$$d_{k_1,k_2}^{trn}(f(\mathbf{B}_{k_1}^{\alpha}), f(\mathbf{B}_{k_2}^{\beta})) = \begin{cases} exp(-\frac{\left\| f(\mathbf{B}_{k_1}^{\alpha}) - f(\mathbf{B}_{k_2}^{\beta}) \right\|^2}{2\sigma^2}), & \mathbb{C}1 \\ +\infty, & others \end{cases} \tag{3}$$

where $\alpha = \mathbf{T}_{i_1}^{P}, \mathbf{M}_{j_1}$ and $\beta = \mathbf{T}_{i_2}^{G}, \mathbb{N}_{\mathbf{M}}(\mathbf{M}_{j_1})$. $\mathbb{C}1$ represents the two bottom layer nodes $\mathbf{B}_{k_1}^{\alpha}$ and $\mathbf{B}_{k_2}^{\beta}$ must belong to the same cluster, and $i_1 = i_2$ is required to ensure that $\mathbf{B}_{k_1}^{\alpha}$ and $\mathbf{B}_{k_2}^{\beta}$ from two person images with the same identity. At last, the Hungarian algorithm is also used to collect training data in the cost matrix M_c with the cost function d^{trn}, and the training data is used to train the LADF which is used to calculate cost function d^{tst}.

Using SBM+NBM, we can get similarity score between two trees, which is calculated by the sum of d^{tst} on each NBM's matching pair. However, due to the fact that the number of the bottom layer nodes are different between trees, in order to obtain a fair matching result, we would use the minimum number of the NBM's matching pairs between one tree of probe forest and all trees of gallery forest.

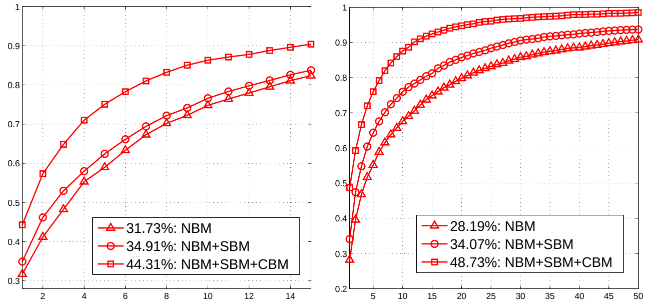
3.3. The Top Layer Matching Using CBM

In top layer node matching, CBM is presented to refine the ranking result. Because the DDN algorithm is imperfect in foreground segmentation, some background noise is retained and some foreground information is lost. Therefore, two baseline models which focus on the whole image (keep all the information including body part and background, denoted as WI) [14] and the entire human body (only keep the foreground without division, denoted as EHB) [16] are used as complementary components in CBM. Based on the complementary result, the highest confident matching can be achieved if SBM+NBM, WI and EHB simultaneously determine that one node in $\mathbf{T}_{i_2}^{G}$ (e.g $\mathbf{T}_{i_*}^{G}$) is the best match of one node in $\mathbf{T}_{i_1}^{P}$. Therefore, other nodes in $\mathbf{T}_{i_1}^{P}$ would be at best becoming the second best match of $\mathbf{T}_{i_*}^{G}$. Moreover, if more than one node in $\mathbf{T}_{i_1}^{P}$ (e.g $\mathbf{T}_{i_*}^{P}|_{i_*=1}^{n_*}$) is the best match of one node in $\mathbf{T}_{i_2}^{G}$ (e.g $\mathbf{T}_{i_*}^{G}$), only the one with the lowest degree of visual ambiguity in $\mathbf{T}_{i_*}^{P}|_{i_*=1}^{n_*}$ will be selected as the best match of $\mathbf{T}_{i_*}^{G}$.

In CBM, the degree of visual ambiguity is determined by the combination of SBM+NBM, WI and EHB, and is calculated by the ratio of similarity between one node in $\mathbf{T}_{i_1}^{P}$ to its best match and its second best match in gallery forest. Intuitively, if one person image (e.g I_a) is the most similar to another person image (e.g $I_{a'}$), and the similarity between I_a and other images is low, the degree of visual ambiguity between I_a and $I_{a'}$ would be considered relatively low.

4. EXPERIMENTAL STUDY

To evaluate the proposed method, experimental study is conducted on two public datasets, e.g., CAVIAR4REID [17] and VIPeR [18]. The CAVIAR4REID dataset is captured in an indoor scenario with 72 persons, and each person has 10 to 20 images under multiple camera views. In the evaluation of CAVIAR4REID, multiple exemplars per individual are available in the probe set and only one exemplar per individual in the gallery set. Different from CAVIAR4REID, VIPeR is captured in an outdoor scenario with complex variations of background and illumination with 632 persons, and each person has only two images under different camera views. In the evaluation of VIPeR, there would have one exemplar per individual in the probe set and the other exemplar in the gallery set. Person re-identification is done by computing the similarity between any two persons in the probe and gallery sets. For a fair comparison, we use the same training and testing protocol in [14], which randomly divide CAVIAR4REID (VIPeR)



(a) CAVIAR4REID (b) VIPeR

Fig. 4. CMC on CAVIAR4REID (a) and VIPeR (b) datasets. The x axis represents rank and y axis represents matching rate. Rank-1 matching rate is marked before approaches.

into two sets, 36 (316) persons for training and the rest for testing. The average cumulative match characteristic (CMC) curves [18] is used over 10 trials to show the ranked matching rates. A rank r matching rate indicates the percentage of the image of probe set with correct matches found in the top r ranks against the number of person in gallery set. Thus, rank-1 matching rate is thus the correct matching rate.

Fig.(4) illustrates the results of NBM, NBM+SBM and NBM+SBM+CBM on CAVIAR4REID and VIPeR datasets. The NBM is calculated by directly matching the bottom layer nodes with $\delta = 0$ in SBM. The experiment results show that NBM can generate weak results, and NBM+SBM proves that the fusion of bottom and middle layer nodes can improve the matching rate by reducing the misalignment between two images. Comparing with the results of NBM+SBM, the result of NBM+SBM+CBM increases the rank-1 matching rate from 34.91% to 44.31% on CAVIAR4REID dataset and 34.07% to 48.73% on VIPeR dataset. The results show that the combination of NBM+SBM+CBM can effectively improve the matching rate of NBM+SBM, which is sensitive to the variation of background. Since VIPeR is captured in an outdoor scenario which suffers more variation of background, the improvement of NBM+SBM+CBM is higher than CAVIAR4REID which is captured in an indoor scenario.

Fig.(5) shows that our HBGM method achieves the best rank-1 performance, 44.31%, and outperforms other state-of-the-art methods on CAVIAR4REID dataset including SDLAF [19], PS [17], GaLF [16], LADF [14], EPKFM [20] and MFA [21], and it also outperforms the best state-of-the-art rank-1 matching rate on this dataset by 4.11%.

As shown in Fig.(6), comparing with ELF [9], SDLAF [19], PatMatch [2], eSDC [2], LADF [14], SalMatch [3], kBiCov [22], DeepArch [23], LOMO+XQDA [24] and mFilter+LADF [13], HBGM improves the matching rate on VIPeR dataset. On this dataset, the combination of LOMO+XQDA [24] and LADF [14] achieves the state-of-the-art rank-1 matching rate, 50.32%. Our method is slightly less than it by 1.59%, but it outperforms the second best rank-1 result mFilter+LADF [13] by 5.34%.

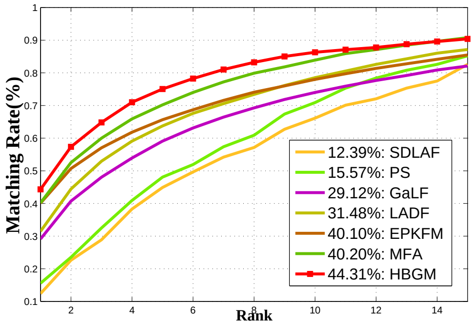


Fig. 5. CMC on CAVIAR4REID dataset. Rank-1 matching rate is marked before the name of each approach.

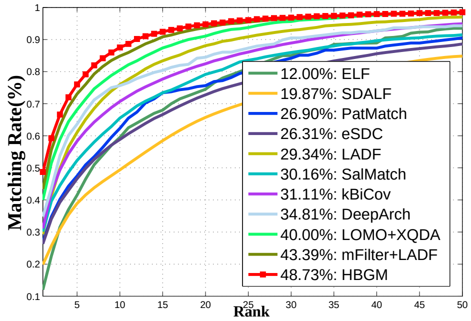


Fig. 6. CMC on VIPeR dataset. Rank-1 matching rate is marked before the name of each approach.

5. CONCLUSION

In this work, we present the HBGM method for person re-identification. Since human eyes observe person appearance roughly first and then goes further into the details gradually, a tree structure modeling method of person image is proposed and HBGM is introduced for the matching of each layer between trees. The NBM, SBM and CBM are further proposed to describe the bottom, middle and top layer nodes matching between trees. The validity of HBGM is demonstrated on two public datasets, and competitive results in terms of quantitative evaluation have proven its potential.

6. ACKNOWLEDGEMENTS

This study was partially supported by the National Natural Science Foundation of China (No. 61472019) and the National High Technology Research and Development Program of China (No. 2013AA01A603). Supported by the Programme of Introducing Talents of Discipline to Universities and the Open Fund of the State Key Laboratory of Software Development Environment under grant #SKLSDE-2015KF-01. Thank you for the support from HAWKEYE Group.

7. REFERENCES

[1] Omar Javed, Khurram Shafique, Zeeshan Rasheed, and Mubarak Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *CVIU*, vol. 109, pp. 146–162, 2008.

[2] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *CVPR*. IEEE, 2013, pp. 3586–3593.

[3] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *ICCV*. IEEE, 2013, pp. 2528–2535.

[4] Hao Sheng, Yan Huang, Yanwei Zheng, Jiahui Chen, and Zhang Xiong, "Person re-identification via learning visual similarity on corresponding patch pairs," in *KSEM*. Springer, 2015, pp. 787–798.

[5] Jorge Garcia, Niki Martinel, Christian Micheloni, and Alfredo Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *ICCV*. IEEE, 2015, pp. 1305–1313.

[6] Yang Hu, Shengcai Liao, Zhen Lei, Dong Yi, and S-tan Z Li, "Exploring structural information and fusing multiple features for person re-identification," in *CVPR Workshops*. IEEE, 2013, pp. 794–799.

[7] Walid Ayedi, Hichem Snoussi, and Mohamed Abid, "A fast multi-scale covariance descriptor for object re-identification," *PR Letters*, vol. 33, pp. 1902–1907, 2012.

[8] Yan Huang, Hao Sheng, Yang Liu, Yanwei Zheng, and Zhang Xiong, "Person re-identification by unsupervised color spatial pyramid matching," in *Knowledge Science, Engineering and Management*. Springer, 2015, pp. 799–810.

[9] Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley, "Person reidentification using spatiotemporal appearance," in *CVPR*. IEEE, 2006, pp. 1528–1535.

[10] Kaspar Riesen and Horst Bunke, "Approximate graph edit distance computation by means of bipartite graph matching," *Image and Vision Computing*, vol. 27, pp. 950–959, 2009.

[11] Liang Lin, Xiaolong Wang, Wei Yang, and Jian-Huang Lai, "Discriminatively trained and-or graph models for object shape detection," *PAMI*, vol. 37, pp. 959–972, 2015.

[12] Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Pedestrian parsing via deep decompositional network," in *ICCV*. IEEE, 2013, pp. 2648–2655.

[13] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," in *CVPR*. IEEE.

[14] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*. IEEE, 2013, pp. 3610–3617.

[15] Wei Zhang, Xiaogang Wang, Deli Zhao, and Xiaoou Tang, "Graph degree linkage: Agglomerative clustering on a directed graph," in *ECCV*. Springer, 2012, pp. 428–441.

[16] Bingpeng Ma, Qian Li, and Hong Chang, "Gaussian descriptor based on local features for person re-identification," in *ACCV*. Springer, 2014, pp. 505–518.

[17] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011, pp. 1–11.

[18] Douglas Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS*, 2007.

[19] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*. IEEE, 2010, pp. 2360–2367.

[20] 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 *CVPR*, 2015.

[21] Fei Xiong, Mengran Gou, Octavia Camps, and Mari-o Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*. Springer, 2014, pp. 1–16.

[22] Bingpeng Ma, Yu Su, and Frederic Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, pp. 379–390, 2014.

[23] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *CVPR*. IEEE, 2015, pp. 3908–3916.

[24] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*. IEEE, 2015, pp. 2197–2206.