This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

Robust and Efficient Graph Correspondence Transfer for Person Re-identification

Qin Zhou, Heng Fan, Hua Yang, *Member, IEEE,* Hang Su, *Member, IEEE,* Shibao Zheng, *Member, IEEE,* Shuang Wu, and Haibin Ling, *Member, IEEE*

***Abstract*—Spatial misalignment caused by variations in poses and viewpoints is one of the most critical issues that hinders the performance improvement in existing person re-identification (Re-ID) algorithms. To address this problem, in this paper, we present a robust and efficient graph correspondence transfer (REGCT) approach for explicit spatial alignment in Re-ID. Specifically, we propose to establish the patch-wise correspondences of positive training pairs via graph matching. By exploiting both spatial and visual contexts of human appearance in graph matching, meaningful semantic correspondences can be obtained. To circumvent the cumbersome *on-line* graph matching in testing phase, we propose to transfer the *off-line* learned patch-wise correspondences from the positive training pairs to test pairs. In detail, for each test pair, the training pairs with similar pose-pair configurations are selected as references. The matching patterns (i.e., the correspondences) of the selected references are then utilized to calculate the patch-wise feature distances of this test pair. To enhance the robustness of correspondence transfer, we design a novel pose context descriptor to accurately model human body configurations, and present an approach to measure the similarity between a pair of pose context descriptors. Meanwhile, to improve testing efficiency, we propose a correspondence template ensemble method using the voting mechanism, which significantly reduces the amount of patch-wise matchings involved in distance calculation. With aforementioned strategies, the REGCT model can effectively and efficiently handle the spatial misalignment problem in Re-ID. Extensive experiments on five challenging benchmarks, including VIPeR, Road, PRID450S, 3DPES and CUHK01, evidence the superior performance of REGCT over other state-of-the-art approaches.**

***Index Terms*—Person re-identification (Re-ID), graph matching, correspondence transfer, pose context descriptor, correspondence template ensemble.**

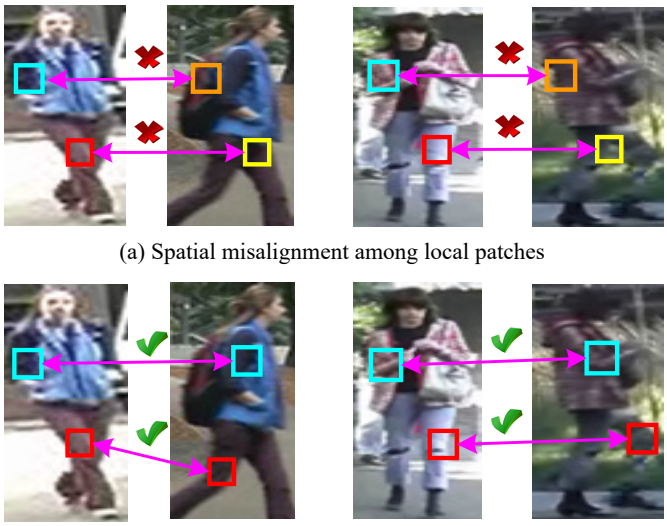


Fig. 1: Illustration of the spatial misalignment problem in Re-ID. Image (a) illustrate the spatial misalignment problem (i.e., spatially corresponding patches do not indicate correct semantic patch-wise matching) caused by pose and viewpoint changes. The proposed REGCT model can capture the correct semantic matching among patches using patch-wise graph matching, as shown in image (b).

I. Introduction

PERSON re-identification (Re-ID), which aims to associate a probe image to images in the gallery set (usually across different non-overlapping camera views), plays a crucial role in various applications including video surveillance, human-computer interaction, etc. Despite great successes in recent years, accurate Re-ID remains challenging due to many factors such as large human appearance changes in different camera views and heavy body occlusions. To deal with these issues, numerous Re-ID approaches have been proposed [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11].

For Re-ID, a major challenge is to deal with the inevitable spatial misalignment problem between image pairs caused by large variations in camera views and human poses, as shown in Fig. 1. Most existing methods [12], [13], [14], nevertheless, focus on addressing the problem of Re-ID by comparing the holistic visual differences between images, which ignore the spatial misalignment problem. To alleviate this issue, there are some attempts to apply part-based approaches to handle misalignment [15], [16], [3], [17]. These methods divide objects into local patches and perform an *on-line* patch-level matching for Re-ID. Though these approaches can handle spatial misalignment to some extent, being in lack of modeling the spatial and visual context information among local patches during correspondence learning, they still fail in presence of visually similar body appearances, or occlusions. Some other algorithms divide the images into fixed stripes, and extract

Qin Zhou, Shibao Zheng and Hua Yang are with Institute of Image Processing and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. Email: {zhou.qin.190, sbzh, hyang}@sjtu.edu.cn.

Heng Fan and Haibin Ling are with Department of Computer & Information Sciences, Temple University, Philadelphia 19122, USA. Email: {hengfan, hbling}@temple.edu.

Hang Su is with Tsinghua University, Beijing 100084, China. Email: suhangss@mail.tsinghua.edu.cn.

Shuang Wu is with YouTu Lab, Tencent, Shanghai 200233, China. Email: calvinwu@tencent.com.

Shibao Zheng and Hua Yang are the corresponding authors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

Fig. 2: Illustration of the fact that *on-line* graph matching is not suitable for generating semantic correspondences between negative pairs.

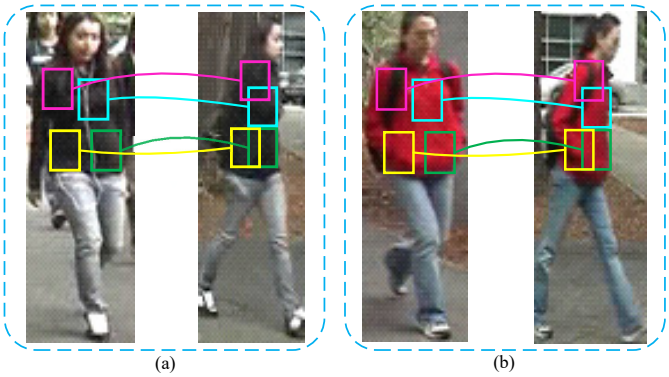


Fig. 3: The observation that two image pairs with similar pose-pair configurations tend to share similar patch-wise correspondences. For example, the left/right images in (a) have similar poses w.r.t. the left/right image in (b). Therefore, the patch-wise matchings denoted by bounding boxes of the same colors in (a) are similar to matchings in (b). Best viewed in color.

visual statistics from each stripe, assuming that the human body is centered and consistently cropped within the bounding boxes [7], [4]. However, it is often not the case that human body is optimally cropped, especially when the bounding boxes are generated by pedestrian detection algorithms.

Intuitively, patch-wise correspondences are spatially and visually compatible to make reasonable semantic matchings. As shown in Fig. 1(b), spatial compatibility means that the semantically corresponding patch is supposed to exist in the local neighborhood of the probe patch in the image plane (This ensures that a patch of the head should not be matched with a patch on the leg). On the other hand, visual compatibility indicates that semantic matchings should be visually similar. Based on these two basic assumptions, we propose to automatically discover the patch-level semantic matching patterns for each positive training pair via graph matching. In our model, both spatial and visual contexts are taken into consideration to establish accurate semantic patch-wise matching results. By using the part-based strategy and implicitly modeling body context information into graph matching, our REGCT algorithm is able to deal with the spatial misalignment problem.

Although graph matching is straightforward for establishing semantic correspondences between positive image pairs, we argue that it is not suitable for negative ones. As shown in Fig. 2, *on-line* graph matching may establish correspondences between the two red patches, since they are spatially and visually compatible within this image pair. However, they are not semantic matchings (one belongs to the torso, while the other belongs to the bag). To address this issue, instead of directly estimating the patch-wise matchings between each test pair, we propose to transfer the *off-line* learned patch-wise matchings from the positive training pairs to the test pairs for performance evaluation.

The concept of correspondence transfer is based on the observation that two image pairs with similar pose-pair configurations tend to share similar patch-level correspondences (as shown in Fig. 3). To better represent the human body configurations, we design a novel pose context descriptor to capture the spatial context of body joints. For each pair of test samples, their pose context descriptors are compared with the pose context descriptors of all the positive training pairs, and the training pairs with the most similar pose-pair configurations are selected as references. Finally, the matching patterns of those referred training pairs are utilized to compute the overall feature distance between this test pair.

In summary, we make the following contributions: (1) We for the first time propose a novel robust and efficient graph correspondence transfer (REGCT) model for Re-ID, which takes into account both spatial and visual contexts to handle the spatial misalignment problem by establishing semantic patch-wise matchings between positive training pairs. (2) We introduce the pose context descriptor to accurately model the body configurations for more robust correspondence transfer. (3) We present a voting based strategy to integrate multiple noisy correspondence templates into a more compact patch-wise matching pattern, which not only reduces the computational load, but also improves the robustness of correspondence transfer. (4) Extensive experiments on five benchmarks demonstrate that our REGCT model performs favorably against state-of-the-art approaches, and in fact even better than many deep learning based solutions.

This paper is an extended version of our preliminary work [18]. The main differences from [18] include: (1) For correspondence transfer, we propose a novel pose context descriptor based on the topology structure of the estimated joint locations [19], which improves the robustness of correspondence transfer and demonstrates superior recognition performance compared with the body orientation based transfer in [18]. (2) We present a voting based approach to integrate multiple noisy correspondence templates into a compact matching pattern, resulting in computation reduction as well as robustness improvement during testing. (3) We conduct more ablative studies to analyze each component in our REGCT model and give insights into the best configurations of different parameter settings. (4) Notable performance gains are obtained with the above-mentioned contributions compared to [18].

II. Related Work

Being extensively studied, Re-ID has drawn extensive attention in the past years. For a comprehensive survey, please refer to [20], [21]. In this section, we briefly review existing Re-ID algorithms from three perspectives: (1) feature representation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

based methods that focus on designing sophisticated features to better represent the human appearance, (2) metric learning based algorithms that pursue discriminative subspaces where features of the same person sit closer than those of different individuals, and (3) deep learning based approaches that aim to learn discriminative representations through end-to-end deep architecture modeling.

1) Feature Designing Algorithms: The early works on person re-identification focus on designing representative features to pursue identity invariance across different cameras. In order to improve the representative and discriminative ability of hand-crafted features, various visual cues are exploited. In [1], three complementary visual cues are used to model the human appearance: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy. Besides, symmetry and asymmetry body structure information are applied to localize the vertical axis of human body, and the local features are reweighted by the distance with respect to the vertical axis such that the effects of pose variations are minimized. The work of [22] proposes a spatio-temporal segmentation algorithm to generate salient edges and combines these salient edges with normalized colors to design invariant signatures. Other feature descriptors, including fisher vector encoded local descriptor [23], HPEsignature [24] and mean riemannian covariance grid [25], are also introduced to better represent the human appearance. Even though carefully designed, hand-crafted features are limited in modeling human appearance in complicated scenes. Therefore, some researchers resort to leveraging learning techniques to model middle or high-level information of human appearance. Representative works include attribute assisted clothes appearance [26], dictionary learning base features [27], [28] etc.

2) Metric Learning Algorithms: Metric learning methods, on the other hand, aim to learn an optimal subspace where the intra-person divergence is minimized and meanwhile the inter-person divergence is maximized. In [5], a simple while effective strategy is proposed to infer the distance from equivalence constraints. In requirement of no iterations for optimization, the algorithm of [5] runs efficiently and can benefit large scale re-identification. PCCA [8] presents a method to learn a low-dimensional discriminative subspace from sparse pairwise similar/disimilar constraints. The authors further introduce the "kernel trick" to generalize PCCA to the nonlinear cases. Local fisher discriminant analysis is introduced in [10] to perform discriminative feature dimension reduction based on which the intra-class instances are pulled together while inter-class ones are pushed apart. To better leverage the advantages of different metric learning methods, the work of [9] seeks metric ensembles by learning to rank techniques. To alleviate the spatial misalignment problem, [4] proposes to learn separate sub-similarity functions for different sub-regions with the help of polynomial feature maps, and complementary strength of local similarities as well as global similarity are combined together for better matching consistency.

3) Deep Learning for Re-ID: Recently deep convolutional features have been demonstrated to significantly boost the performance of various computer vision tasks including object detection [29], [30], tracking [31], [32], object segmentation [33], [34], etc. Inspired by the powerful ability of deep features, many researchers resort to building deep end-to-end architectures to directly learn discriminative high-level features for Re-ID. In [35], the authors propose a novel filter pairing neural network (FPNN) to jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter by designing corresponding layers to take charge of each aspect. Ahmed et al. [36] introduce a novel layer that computes cross-input neighborhood differences to capture local relationships between the two input images, and the cross-input neighborhood differences are aggregated together to form a pairwise cross-view representation for a pair of inputs. The work of [37] introduces a unified deep learning-to-rank framework that learns joint representation and similarities of image pairs directly from image pixels. In [38], Xiao et al. present a novel domain guided dropout algorithm to learn robust feature representations by leveraging information from multiple domains.

The proposed algorithm belongs to the non-deep-learning group. Instead of designing sophisticated features or pursing discriminative distance metrics, we present a novel framework to first establish local patch-wise correspondences, then aggregate the patch-wise feature similarities for recognition. This framework enjoys the flexibility that most of the aforementioned algorithms can be incorporated as part of it to improve the final recognition performance. More specifically, off-the-shelf hand-crafted or deep features can be used as visual cues for graph nodes (image patches in our case) to build the affinity matrix for graph matching. Once the local correspondences are established, existing metric learning algorithms can be adopted to pursue better similarity functions for calculating patch-wise matching scores.

The most relevant work to ours is [15] in which a correspondence structure learning (CSL) method is proposed for Re-ID. However, our REGCT model significantly differs from CSL [15] in two aspects: (i) Instead of learning a holistic correspondence structure for each camera pair in CSL [15], we leverage graph matching to establish accurate patch-wise correspondences for each positive image pair in the training stage, and then transfer the learned matching patterns for distance calculation during testing. (ii) We model the spatial and visual context information in the affinity matrix for graph matching, which is neglected in CSL [15]. Due to the flexible and accurate instance-specific patch-wise correspondence learning and transfer, our algorithm demonstrates superior performance over CSL [15] on all the five benchmarks.

III. Our Approach

In this section, we first give a brief introduction on graph matching, and then elaborate on the proposed REGCT algorithm, which is composed of correspondence learning, reference selection, patch-wise feature distance calculation and aggregation based on correspondence transfer. The overall framework of the proposed REGCT algorithm is illustrated in Fig. 4.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

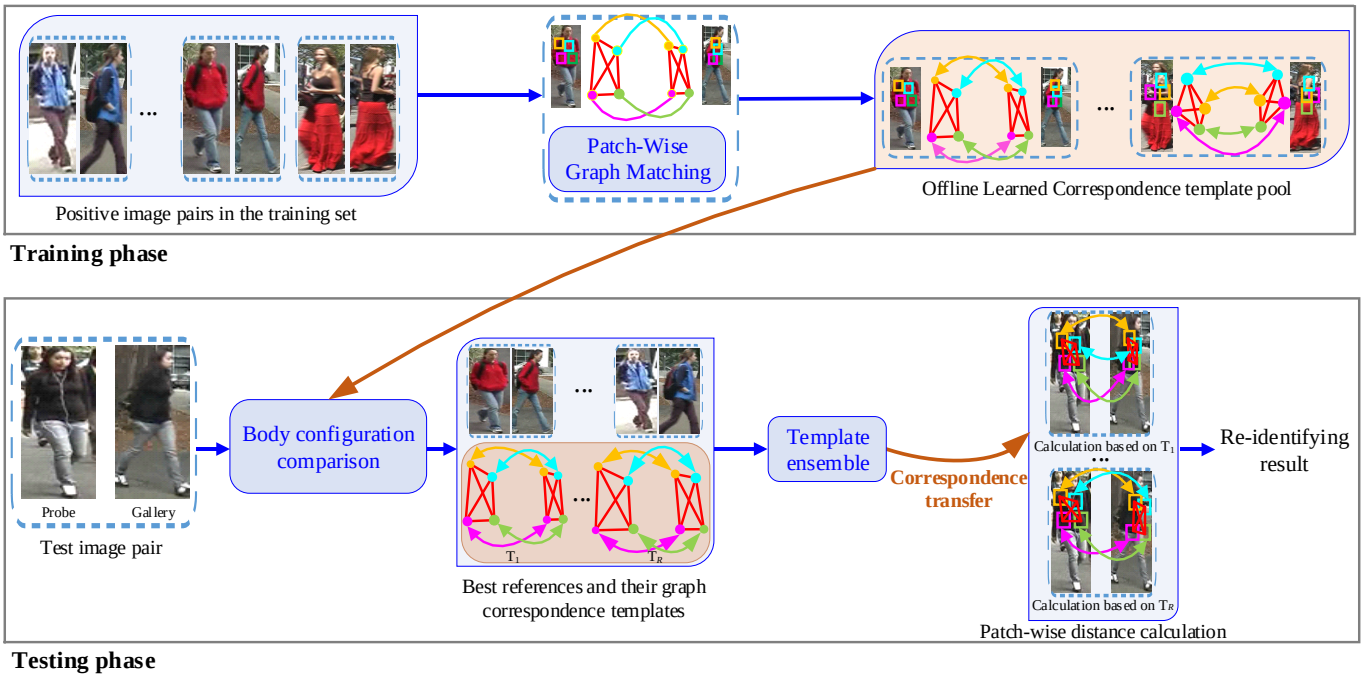


Fig. 4: Illustration of the REGCT model. During training, spatial and visual context information are embedded into graph matching to establish patch-wise matchings for positive training pairs with various pose-pair configurations. During testing, for a pair of test samples, we choose a few positive training pairs with the most similar pose-pair configurations as references, and then transfer the correspondences of these references to this test pair for feature distance calculation. **Different from the preliminary work [18], a novel pose context descriptor is proposed for more accurate body configuration comparison. Besides, the templates ensemble method is introduced to pursue robust and efficient correspondence transfer during testing.**

A. Graph Matching

Graph matching is a fundamental problem closely related to many computer vision tasks including feature registration [39], object recognition [40], visual tracking [41] and so on. In this paper, we omit the detailed literature on graph matching, and only present the commonly adopted formulation to provide some insights into how graph matching is utilized to establish correspondences. For a detailed survey on graph matching, please refer to [42], [43].

Generally, a graph $G = (V, E)$ of size n is defined on a finite set of n vertices $V = \{v_i\}_{i=1}^{n}$ and edges $E \subset V \times V$. For two graphs $G_1 = (V_1, E_1)$ of size n_1 and $G_2 = (V_2, E_2)$ of size n_2, graph matching aims to find an optimal assignment matrix $X \in \{0,1\}^{n_1 \times n_2}$, where $X_{ij} = 1$ indicates an established correspondence between node i in G_1 and node j in G_2. X can be optimized by maximizing the following objective:

$$\arg\max_{\mathbf{x}} \ \mathbf{x}^{\mathrm{T}} K \mathbf{x}, \tag{1}$$

where $\mathbf{x} \in \{0,1\}^{n_1 n_2}$ is the vector form of the assignment matrix X, and $K \in R^{n_1 n_2 \times n_1 n_2}$ represents the affinity matrix that encodes both the node similarity and edge compatibility information between G_1 and G_2.

B. Patch-wise correspondence learning with graph matching

In this paper, we adopt the attributed graph to represent the human body. Specifically, we decompose the images into many overlapping patches, and represent each image with an undirected attributed graph $G = (V, E, A)$, where each vertex v_i in the vertex set $V = \{v_i\}_{i=1}^{n}$ denotes an image patch, each edge encodes the contextual information of the connected vertex pair, and the vertex attributes $A = \{A^P, A^F\}$ represent spatial and visual features of local patches.

During training, given a pair of positive images I_1 and I_2 with identity labels l_1 and l_2, where $l_1 = l_2$ (i.e., I_1 and I_2 belong to the same person), they can be represented with attributed graphs $G_1 = (V_1, E_1, A_1)$ and $G_2 = (V_2, E_2, A_2)$, respectively. The patch-wise correspondence learning aims to establish the vertex correspondences $X \in \{0,1\}^{n_1 \times n_2}$ between V_1 with n_1 vertices and V_2 with n_2 vertices, such that the intra-person matching score (i.e., $l_1 = l_2$) is maximized on the training set.

In Re-ID, $X_{i_1 i_2} = 1$ means the i_1-th patch in I_1 is matched with the i_2-th patch in I_2. We adopt the formulation in Eq. 1 to model our patch-wise correspondence learning problem:

$$\arg\max_{\mathbf{x}} \ \mathbf{x}^{\mathrm{T}} K \mathbf{x}, \quad s.t. \begin{cases} X_{i_1 i_2} \in \{0,1\}, \forall i_1 \in \{1, \cdots, n_1\}, \forall i_2 \in \{1, \cdots, n_2\} \\ \sum_{i_1} X_{i_1 i_2} \leq 1, \forall i_2 \in \{1, \cdots, n_2\}, \\ \sum_{i_2} X_{i_1 i_2} \leq 1, \forall i_1 \in \{1, \cdots, n_1\}, \end{cases} \tag{2}$$

where one-to-one matching constraints are imposed on the assignment matrix X. We follow [44] to optimize Eq. 2.

1) **Affinity matrix design:** Due to the large variations in human body configurations caused by heavy pose and view changes, it is not suitable to directly apply traditional spatial layout based affinity matrix for Re-ID. In addition, taking into consideration the importance of visual appearance in Re-ID, we combine both visual feature and spatial layout of human appearance to develop the affinity matrix.

In specific, the diagonal components $K_{i_1 i_2, i_1 i_2}$ of the affinity matrix K (which capture the node compatibility between vertex $v_{i_1} \in V_1$ and vertex $v_{i_2} \in V_2$) are calculated as follows:

$$K_{i_1 i_2, i_1 i_2} = S^P_{i_1 i_2} \cdot S^F_{i_1 i_2}, \tag{3}$$

where $S^P_{i_1 i_2}$ and $S^F_{i_1 i_2}$ refer to the *spatial proximity* and *visual similarity* between v_{i_1} and v_{i_2} respectively. The $S^P_{i_1 i_2}$ and $S^F_{i_1 i_2}$ can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

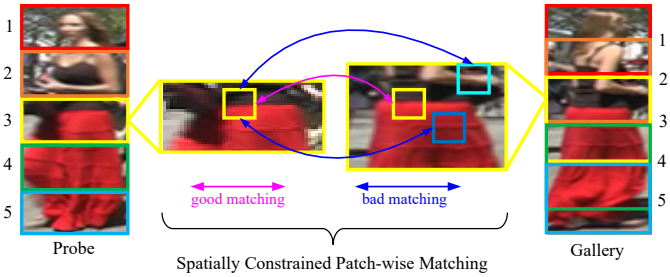


Fig. 5: Illustration of the spatially constrained patch-wise matching. Following the commonly utilized stripe decomposition of the human body [4], we first divide the probe image into a few stripes, then the search space for each probe stripe is spatially constrained to its local neighborhood (e.g., patches within the yellow stripe of the probe image is constrained to be matched with patches in the corresponding yellow stripe of the gallery image). To ensure that the counterpart patch exists in the search space, the corresponding gallery stripe is set to be larger than the probe stripe. Best viewed in color.

Fig. 6: Sample images in eight classes in the TUD dataset [46]. Note that the TUD dataset is only used for training the body orientation classification model, and is different from the benchmark datasets in our experiments.

be mathematically computed as:

$$\begin{aligned} S^P_{i_1 i_2} &= \exp(-\|A^P_{i_1} - A^P_{i_2}\|_2), \\ S^F_{i_1 i_2} &= \exp(-\|A^F_{i_1} - A^F_{i_2}\|_2), \end{aligned} \tag{4}$$

where $A^P_{i_1}$ and $A^P_{i_2}$ denote spatial positions of v_{i_1} and v_{i_2}, and $A^F_{i_1}$ and $A^F_{i_2}$ represent their visual features.

Likewise, for non-diagonal element $K_{i_1 i_2, j_1 j_2}$ in K, which encodes the compatibility between two matched vertex pairs $(v_{i_1} \in V_1, v_{i_2} \in V_2)$ and $(v_{j_1} \in V_1, v_{j_2} \in V_2)$, it can be obtained as the following:

$$K_{i_1 i_2, j_1 j_2} = S^P_{i_1 j_1, i_2 j_2} \cdot S^F_{i_1 j_1, i_2 j_2}, \tag{5}$$

where $S^P_{i_1 j_1, i_2 j_2}$ and $S^F_{i_1 j_1, i_2 j_2}$ represent spatial and visual compatibilities between matched patch pairs (v_{i_1}, v_{i_2}) and (v_{j_1}, v_{j_2}), and they are calculated by

$$\begin{aligned} S^P_{i_1 j_1, i_2 j_2} &= \exp(-\|(A^P_{i_1} - A^P_{j_1}) - (A^P_{i_2} - A^P_{j_2})\|_2), \\ S^F_{i_1 j_1, i_2 j_2} &= \exp(-\|(A^F_{i_1} - A^F_{j_1}) - (A^F_{i_2} - A^F_{j_2})\|_2). \end{aligned} \tag{6}$$

In this way, the calculated affinity matrix K implicitly embeds the spatial and visual contextual information into the graph matching procedure, such that during the optimization, correspondences with larger node similarities and more compatible edges are selected. Therefore, we can obtain a spatially and visually compatible patch-wise matching result for Re-ID.

2) **Spatially constrained matching:** In existing part-based Re-ID methods [15], [3], an image is typically decomposed into hundreds of patches to capture detailed local visual information, leading to intractability in solving Eq. (2). To reduce the search space and inhibit potential matching ambiguity, similar to the commonly utilized spatial constraints [7], [4], we introduce the structure constrained matching. More specifically, a probe image is divided into a few horizontal stripes (non-overlapping), and for each probe stripe, its search space for patch-wise matching is constrained to a corresponding stripe in the gallery image (As shown in Fig. 5). Then patch-wise matchings are established between the probe stripe and the corresponding gallery stripe by optimizing Eq. (2). Fig. 5 illustrates the process of spatially constrained patch-wise matching.

C. Reference selection via pose-pair configuration comparison

We argue that the learned patch-wise correspondence patterns can be favorably transferred to image pairs with similar pose-pair configurations in the testing set, and these transferred correspondences can be directly utilized to compute the distance between probe and gallery images in the test set (as demonstrated in Fig. 3). To this end, we need to find out the best references for each test pair from the training set.

In our preliminary study [18], we adopt body orientations to roughly capture body configuration. However the representative ability of body orientation is limited. In this paper, we propose a novel pose context descriptor to model the relative spatial distributions of the body joints to fully capture the spatial layout of human body. Please note here the locations of the body joints can be easily obtained with off-the-shelf pose estimation algorithms (e.g., [19]). For self-completeness, we present both the orientation based and the pose context descriptor based body configuration comparison strategies and compare their recognition performance in the experimental part.

1) **Comparing body configurations using body orientation:** We propose to utilize a simple yet effective random forest method [45] to compare different body orientations. Specifically, images are classified into eight different clusters including 'left', 'right', 'front', 'back', 'left-front', 'right-front', 'left-back' and 'right-back', according to their body orientations, as shown in Fig. 6. In order to train the random forest model, each image is represented with multi-level HoG features (i.e., cell sizes are set to $8 \times 8, 16 \times 16, 32 \times 32$ respectively, with a block size of 2×2 cells and a block stride of one cell for each direction), and then fed into each decision tree to build the random forest [45]. Once the random forest $\mathcal{M} = \{M_i\}_{i=1}^{|\mathcal{M}|}$ is built, where $|\mathcal{M}|$ denotes the number of trees in $\mathcal{M}$, the body configuration similarity O between two images I_i and I_j can be calculated as:

$$\mathrm{O}(I_i, I_j) = \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} y^m_{ij}, \tag{7}$$

where y^m_{ij} is an indicator, and $y^m_{ij} = 1$ if I_i and I_j arrive at the same terminal node in $M_m \in \mathcal{M}$, otherwise $y^m_{ij} = 0$.

2) **Comparing body configurations using pose context descriptor:** To accurately model the human body configurations, we propose a novel pose context descriptor, which captures the relative spatial distribution across different joint pairs. In detail, fourteen body joints (i.e., head, neck, left/right shoulders, left/right elbows, left/right wrists, left/right coxaes, left/eight knees and left/right ankles) are estimated with the pre-trained model in [19], the spatial locations of which are denoted as $\{J_i, i \in 1, \cdots, 14\}$. Each joint is then associated with a local polar coordinate system, centering at J_i. Then for the other joints with locations $\{J_j, j \in \{1, \cdots, 14\}, j \neq i\}$, we calculate a histogram for each J_j by considering the magnitudes and angles of J_j in the polar system centered at J_i. In this way, two pose context coding matrices $\Psi \in R^{14\times13}, \Phi \in R^{14\times13}$ are formed, where the element $\Psi(i,j)$ in row i column j of Ψ contains the **magnitude bin** that J_j lies in the polar system centered at J_i. Likewise, $\Phi(i,j)$ is the **angle bin** that J_j lies in the polar system centered at J_i. In this paper, the number of bins are set to 8 for both

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

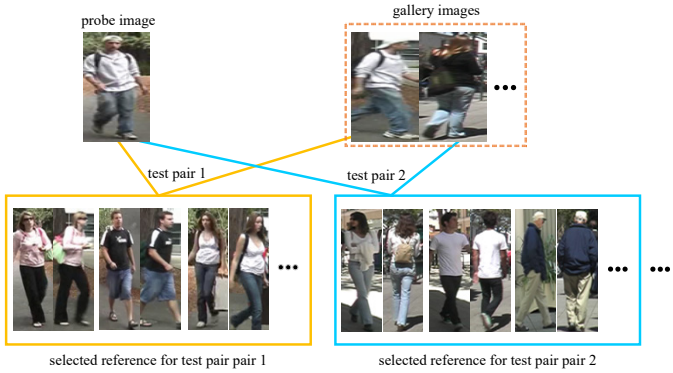


Fig. 7: Demonstration of reference selection results.

the magnitude and the angle. For a pair of images I_1, I_2, their pose configuration similarity is calculated as:

$$\mathrm{O}(I_1, I_2) = S_\Psi(\Psi_1, \Psi_2) \cdot S_\Phi(\Phi_1, \Phi_2), \tag{8}$$

where $S_\Psi(\Psi_1, \Psi_2)$ is the similarity score between the magnitude context matrices Ψ_1 and Ψ_2, and $S_\Phi(\Phi_1, \Phi_2)$ is the similarity score between the angle context matrices Φ_1 and Φ_2, which are calculated as follows:

$$S_\Psi(\Psi_1, \Psi_2) = \mathrm{mean}(\exp(-d(\Psi_1, \Psi_2))), \tag{9}$$

where $d(\cdot)$ is the element-wise Euclidean distance. Considering that the angle bins are cyclic (e.g., 0 degree equals to 360 degree), we design a cyclic Euclidean distance to capture the proximity between angle bins,

$$S_\Phi(\Phi_1, \Phi_2) = \mathrm{mean}(\exp(-d_c(\Phi_1, \Phi_2))), \tag{10}$$

where $d_c(\cdot)$ is the cyclic Euclidean distance. For a pair of counterpart angle bins $(\Phi_1(i,j) \in \{1, \cdots, 8\}, \Phi_2(i,j) \in \{1, \cdots, 8\})$, denote the minimum bins from $\Phi_1(i,j)$ to $\Phi_2(i,j)$ in the angle circle as α_{ij}, then $d_c(\Phi_1(i,j), \Phi_2(i,j))$ is calculated as $d_c(\Phi_1(i,j), \Phi_2(i,j)) = \alpha_{ij}^2$.

Given two image pairs $P = (I_p, I_g)$ and $P' = (I_p^{'}, I_g^{'})$, their pose-pair configuration similarity $S(P, P')$ is computed as

$$S_o(P, P') = \mathrm{O}(I_p, I_p^{'}) \cdot \mathrm{O}(I_g, I_g^{'}), \tag{11}$$

where $\mathrm{O}(I_p, I_p^{'})$, $\mathrm{O}(I_g, I_g^{'})$ can be obtained using Eq. (7) or Eq. (8). With Eq. (11), we can calculate the body configuration similarities between an test image pair and all the positive training image pairs. Afterwards, R training image pairs with highest similarities are selected as references for the test pair. And their learned patch-wise matchings are utilized for distance calculation as described later. Fig. 7 shows some selected references of the sample test pairs.

D. Distance calculation and aggregation with correspondence transfer

Given that image pairs with similar pose-pair configurations tend to share similar patch-level correspondences, for each test pair of images, we propose to transfer the matching results of the selected references (the way to select the references is presented in Section III-C) to calculate the patch-wise feature distances of this test pair. The details of feature distance calculation using the selected references are presented in the following part.

Given a pair of test images $\bar{P} = (\bar{I}_p, \bar{I}_g)$, where $\bar{I}_p$ and $\bar{I}_g$ are the probe and gallery image respectively. Denote their corresponding graphs as $\bar{G}_p = (\bar{V}_p, \bar{E}_p, \bar{A}_p), \bar{G}_g = (\bar{V}_g, \bar{E}_g, \bar{A}_g)$, we can choose R references for $\bar{P}$ as described in Section 3.2. Let $\mathcal{T} = \{T_i\}_{i=1}^R$ represent the correspondence templates set composed by these R references, where each template $T_i = \{c_{ij}\}_{j=1}^n$ contains n patch-wise correspondences (n is the number of graph node in the probe image), and each correspondence $c_{ij} = (w_{ij}^p, w_{ij}^g)$ denotes the indices of the matched patches in the probe and gallery image (i.e., the w_{ij}^p-th node in $\bar{G}_p$ is matched to the w_{ij}^g-th node in $\bar{G}_g$).

For the test pair $\bar{P}$, we can compute the distance $\mathcal{D}$ between $\bar{I}_p$ and $\bar{I}_g$ as the following:

$$\mathcal{D}(\bar{I}_p, \bar{I}_g) = \frac{1}{R \times n} \sum_{i=1}^{R} \sum_{j=1}^{n} \delta(\bar{A}_{p,w_{ij}^p}, \bar{A}_{g,w_{ij}^g}), \tag{12}$$

where $\delta(\cdot,\cdot)$ denotes the distance metric (in this paper, we adopt the KISSME metric [5]), $\bar{A}_{p,w_{ij}^p}$ and $\bar{A}_{g,w_{ij}^g}$ represent the visual attributes of the w_{ij}^p-th patch in the probe image $\bar{I}_p$ and the w_{ij}^g-th patch in the gallery image $\bar{I}_g$ respectively. In this paper, we use Local Maximal Occurrence features [7] as the visual attributes of each node (i.e., local patch).

With Eq. (12), we can calculate the average patch-wise feature distance using all the correspondences (semantically matched patch pairs between the probe and gallery image) of the selected reference templates. For each probe image, the gallery image with the smallest distance is determined to be the re-identifying result.

A more efficient solution for testing: As demonstrated in [18], for larger or more challenging datasets (e.g., CUHK01,VIPeR etc.), the number of selected reference templates R is suggested to be large to accumulate enough correct patch-wise correspondences (e.g., R is set to 20 for CUHK01 and VIPeR datasets). Therefore, according to Eq. (12), for a dataset with N persons (in the single-shot case, for N persons, there are N probe images and N gallery images), the computational complexity involves $R \times n \times N^2$ patch-wise Mahalanobis distance calculations (here n is the number of the graph node in the probe image).

In practical scenarios, it is desirable to have more efficient solutions. Therefore, we present a new evaluation protocol to significantly reduce the computational load with competitive or better performance. Since directly calculating the feature distances using all the reference patch-wise correspondences is time-consuming, we propose to aggregate the selected R reference matching templates into $k, k \ll R$ more compact matching patterns via a voting scheme.

According to Eq. (12), for each probe patch $w_{ij}^p \in \{1, \cdots, n\}$ in $\bar{I}_p$, the i-th selected template matches it with patch $w_{ij}^g \in \{1, \cdots, n\}$ in $\bar{I}_g$. Denote their spatial offset as:

$$\Delta(w_{ij}^p, w_{ij}^g) = \mathcal{L}(w_{ij}^g) - \mathcal{L}(w_{ij}^p), \tag{13}$$

where $\mathcal{L}(w_{ij}^p)$ denotes the center location for the w_{ij}^p-th patch. Then for each probe patch w_{ij}^p, we can obtain a set of suggested matching patches $\Lambda(w_{ij}^p) = \{w_{ij}^g, i \in 1, \cdots, R\}$. Assume these suggested matching patches vote for the hidden semantic matching $w_{ij}^{g*} \in \{1, \cdots, n\}$, then the location of w_{ij}^{g*} can be simply derived as:

$$\mathcal{L}(w_{ij}^{g*}) = \mathcal{L}(w_{ij}^p) + \frac{1}{R} \sum_{i=1}^{R} \Delta(w_{ij}^p, w_{ij}^g), w_{ij}^g \in \Lambda(w_{ij}^p), \tag{14}$$

With the estimated target location, k nearest patches in the target image are then sampled as the compact semantic matching patches of w_{ij}^p, which are calculated as:

$$\{w_{ij}^{g*}\}_k = \mathcal{N}_k(\mathcal{L}(w_{ij}^{g*})), k \ll R, \tag{15}$$

where $\mathcal{N}_k(\cdot)$ returns the indices of the k nearest gallery patches with respect to the inside calculated target location. In this way, the computational load can be reduced to $k \times n \times N^2, k \ll R$, which is significant especially when N is large.

The final feature distance between $\bar{I}_p, \bar{I}_g$ is then derived as,

$$\mathcal{D}(\bar{I}_p, \bar{I}_g) = \frac{1}{k \times n} \sum_{i=1}^{k} \sum_{j=1}^{n} \delta(\bar{A}_{p,w_{ij}^p}, \bar{A}_{g,w_{ij}^{g*}}), w_{ij}^{g*} \in \{w_{ij}^{g*}\}_k. \tag{16}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

Fig. 8: Top ranked images selected by the proposed REGCT algorithm. Images in the first column are four randomly selected probe images, and the following are the top ranked gallery images of each probe by REGCT. Images marked by green/blue/orange bounding boxes in each row are the ground-truth matches of each probe.

IV. Experimental Results

In this section, we present the details of our experimental results. First we briefly clarify the experimental setup of the proposed graph correspondence transfer algorithm, then we elaborate on the ablative study of each component in the REGCT model to explore the importance of each part and give some insights on how to select the best parameter settings. Finally, we present the comparison results with some state-of-the-art algorithms on five challenging benchmarks, including the VIPeR [47] dataset, the Road [15] dataset, the PRID450S [48] dataset, the 3DPES [49] dataset and the CUHK01 [50] dataset.

A. Experimental setup

*1) **Datasets**:* We conduct experiments on three challenging single-shot datasets (VIPeR, Road and PRID450S), and two multi-shot datasets (3DPES and CUHK01). The characteristics of each dataset are detailed as follows:

VIPeR dataset: The VIPeR [47] is a challenging person re-identification dataset consisting of 632 people with two images from two cameras for each person. It bears great variations in poses and illuminations, most of the image pairs contain viewpoint changes larger than 90 degrees.

Road dataset: The Road dataset [15], consisting of 416 image pairs, is captured from a realistic crowd road scene, with serious interferences from occlusions and large pose variations, making it quite challenging.

PRID450S dataset: The PRID 450S [48] dataset contains 450 pairs of images from two camera views. The very similar background scene and many people wearing similar clothes make it very challenging for person re-identification.

3DPES dataset: The 3DPES dataset [49] contains 1011 images of 192 persons captured from 8 disjoint camera views, the images of which bear serious variations in view angles, illuminations, scales and background clutters. The number of images for a specific person ranges from 2 to 26, and the bounding boxes are generated from automatic pedestrian detection.

CUHK01 dataset: The CUHK01 dataset [50] is a medium-sized dataset for Re-id, captured from two disjoint camera views. It consists of 971 individuals, with each person having two images under each camera view. Different from VIPeR, images in CUHK01 are of higher resolutions. On this dataset, we adopt the commonly utilized 485/486 setting for performance evaluation.

Fig. 9: Some visualized patch-wise graph matching results. In each pair of images, the bounding boxes with the same color refer to an established correspondence by our algorithm.

*2) **Parameter setup**:* The proposed algorithm is implemented in Matlab on an Intel(R) Core(TM) i7-5820K CPU of 3.30GHz. The number of trees in the random forest model is 500. The best configurations of (R, k) as well as the impacts of different patch decompositions of human body are discussed in the following part. All the parameters are available in the source code at http://www.dabi.temple.edu/~hbling/code/gct.htm.

*3) **Evaluation**:* We adopt the commonly used half-training and half-testing setting [5], and randomly split the dataset into two equal subsets. The training/testing sets are further divided into the probe and gallery sets according to their view information. On all the datasets, both the training/testing set partition and probe/gallery set partition are performed 10 times and average performance is recorded. The performance is evaluated by cumulative matching characteristic (CMC) curve, which represents the expected probability of finding the correct match for a probe image in the top r matches in the gallery list.

We record the top ranked gallery images of some sample probe images on the VIPeR dataset, which is presented in Fig. 8. As shown in Fig. 8, the proposed REGCT algorithm can successfully rank images visually similar to the probe image ahead of others, which is the key requirement of most existing surveillance systems. Please note the last row of Fig. 8, the top ranked gallery images by REGCT are all with dark coats and blue jeans, and to some extent, the rank 1 image is visually more similar than the correct match marked by orange bounding box w.r.t. the probe image. Therefore, the proposed REGCT algorithm is able to handle the spatial misalignment problem and generate satisfying ranking results for practical applications.

B. Ablation study

*1) **Visualized results of graph matching**:* To validate the effectiveness of graph matching in establishing local semantic correspondences, we present some patch-wise matching results in Fig. 9. As shown, by taking into consideration the spatial and visual context information during the graph matching procedure, the established patch-wise matchings indeed can preserve the semantic correspondences even if there are severe variations in poses and illuminations.

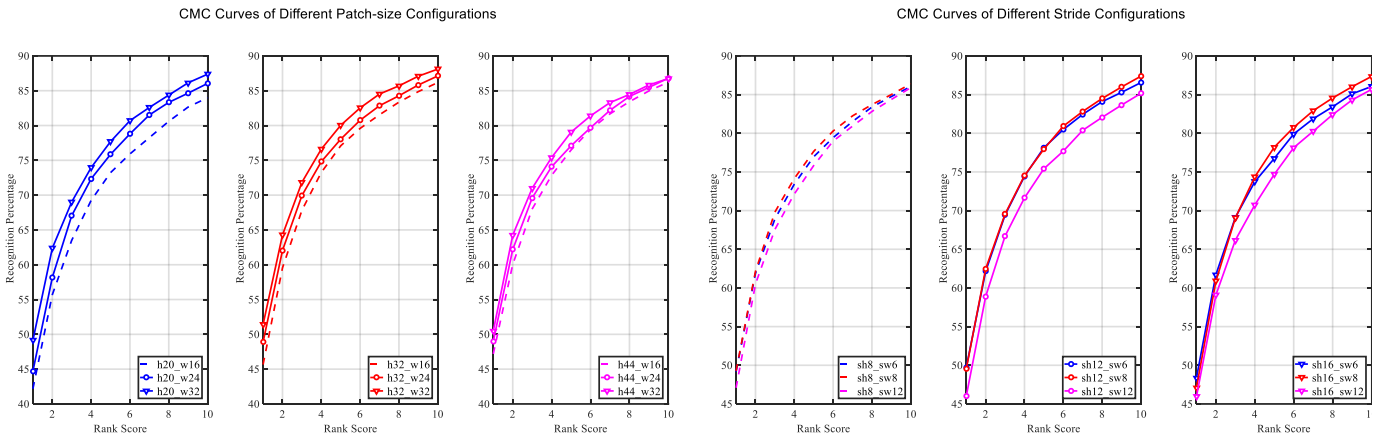
*2) **Influence of different patch decomposition**:* In order to generate some insights into the optimal decomposition of human body into local patches, we study the influence of different patch-wise decompositions on the recognition performance. In specific, we record the recognition performance obtained using different patch sizes and different strides.

To evaluate the influence of patch size, we fix the stride to $stride\_h = 12, stride\_w = 8$ ($stride\_h$ is the stride along the height side, and $stride\_w$ is the stride along the width side). The patch height values and width values are then selected from $\{44, 32, 20\}$ and $\{32, 24, 16\}$ respectively, totally generating 9 combinations. Likewise, we study the impact of different patch strides by fixing the patch size to 32×24, and then varying $stride\_h$ and $stride\_w$ within $\{16, 12, 8\}$ and $\{12, 8, 6\}$ respectively. The detailed comparison results between different patch-wise decomposition settings (on the VIPeR dataset) are illustrated in Table I and Fig. 10.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

TABLE I: Recognition performance with different patch-wise configurations on VIPeR dataset. We study the influence of different configurations of both patch size and stride. Please note, in the patch size config, $h20\_w16$ indicates the patch height and width are set to 20 and 16 respectively. Likewise, in the patch stride config, $sh8\_sw6$ indicates that $stride\_h$ and $stride\_w$ are set to $8, 6$ respectively. The best results are marked in red font.

patch_size config	h20_w16	h20_w24	h20_w32	h32_w16	h32_w24	h32_w32	h44_w16	h44_w24	h44_w32
$rank = 1$	42.2	44.7	49.1	45.7	48.9	**51.4**	47.2	49.0	50.5
$rank = 5$	73.2	75.9	77.7	77.1	**80.8**	80.0	76.5	77.1	79.1
$rank = 10$	84.0	86.0	87.4	86.2	87.1	**88.1**	86.2	86.8	86.7
$rank = 20$	93.7	93.2	93.9	93.6	94.0	**94.6**	93.5	93.3	93.5
stride config	sh8_sw6	sh8_sw8	sh8_sw12	sh12_sw6	sh12_sw8	sh12_sw12	sh16_sw6	sh16_sw8	sh16_sw12
$rank = 1$	49.2	49.3	47.1	49.5	**49.7**	46.0	48.4	47.1	45.9
$rank = 5$	77.0	77.7	75.8	77.9	**78.1**	75.4	76.7	78.2	74.7
$rank = 10$	86.0	86.4	85.9	86.6	**87.4**	85.2	86.0	87.3	85.7
$rank = 20$	93.4	92.9	92.8	93.0	**93.7**	92.1	92.6	**93.7**	93.0



(a) Recognition performance with different patch size configurations.

(b) Recognition performance with different patch stride configurations.

Fig. 10: Analysis on the impacts of different patch decompositions.

As shown in Table I, the optimal patch size is 32×32, achieving the best performance at rank 1, 10, 20 and second best at $rank = 5$. As for the different stride configurations, stride 12×8 performs the best at all the reported ranks among different stride configurations. Meanwhile, from Fig. 10 (a) we can see that, with fixed patch height, the recognition performance increases when patch width gets larger. On the other hand, from Fig. 10 (b), we observe that, best recognition performance is achieved at $stride\_w = 8$, when we fix $stride\_h$ to $8, 12, 16$ respectively. **Based on the above analysis, we set the patch size to 32×32 and strides to $stride\_h = 12, stride\_w = 8$ throughout the following experiments without special clarification.**

3) Analysis on different configurations of (R, k): The number of selected references (R) for calculating the distances between test pairs has an impact on the re-identification performance. With a small R, bad references may have a large impact on the patch-wise distance calculation, deteriorating the recognition performance. By contrast, if the value of R is large, the correspondences transferred from less similar references may introduce inaccurate correspondences, which also degrades the performance. In the original evaluation setting [18], the optimal R for the VIPeR, Road, PRID450S, 3DPES and CUHK01 datasets are 20, 5, 10, 20 and 20, respectively. In this paper, we propose to aggregate the arbitrary optimal number (R) of noisy reference templates into k refined compact templates to improve the robustness of correspondence transfer as well as the testing efficiency. We conduct experiments to find out the best (R, k) combination for all the benchmarks. More specifically, we set R to $R \in \{10, 20, 50, 100\}$, and vary k within $\{1, 3, 5, 10\}$, generating in total 16 configurations of (R, k).

As shown in Table II, generally, the best recognition performance are achieved at small k with arbitrary optimal R. More specifically, $(R = 100, k = 3)$ outperforms other configurations on the VIPeR, Road and PRID450S datasets. While on the 3DPES and CUHK01 datasets, $(R = 50, k = 1)$ generates the best results. Another highly expected phenomenon is that large k (e.g., set k to 10) generally deteriorates the recognition performance regardless of the different settings of R. These two facts validate that the proposed template ensemble approach turns the multiple noisy reference templates into a more robust and accurate matching pattern. And the reason for the performance drop when using large k can be inferred from Eq. 15. In Eq. (15), the final patch-wise matching is selected according to the spatial proximity to the voted target location calculated by Eq. (14). When k is large, more patches that deviate from the estimated target location are included in the distance calculation, thus leading to deterioration in the recognition performance.

As rank-1 performance is a very critical evaluation criterion for person re-identification, we specially study the influence of different (R, k) settings on the rank-1 recognition rates. As shown in Fig. 11, generally, $k = 3$ works best for the VIPeR, Road and PRID450S datasets, $k = 1$ is optimal on the 3DPES and CUHK01 datasets. Compared with k, different configurations of R have relatively small influence on the recognition rates. Since the number of R does not affect the efficiency in the proposed novel evaluation setting, REGCT is guaranteed to obtain favorable performance with better efficiency than the baseline GCT [18]. **In the following part, the best (R, k) configurations are utilized for each dataset without special clarification.**

4) Analysis on the effectiveness of template ensemble: As discussed earlier, the proposed REGCT can reduce the computational load from $R \times n \times N^2$ to $k \times n \times N^2$ ($k \ll R$),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

TABLE II: Recognition performance of different (R, k) configurations. In this table, each column contains the recognition rate at the same rank obtained with different (R, k) configurations. At each rank, the best and second best recognition rates are marked in red and blue respectively. Best viewed in color.

(R,k) \ Datasets		VIPeR				Road				PRID450S				3DPES				CUHK01			
		r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
R=10	k=1	49.9	78.3	87.2	94.2	86.4	95.3	97.5	99.0	69.0	86.3	92.0	96.7	75.3	**92.0**	95.5	99.0	66.2	82.9	89.8	93.7
	k=3	51.6	80.1	88.3	94.5	88.5	96.0	97.7	**99.4**	**69.1**	**88.2**	**93.4**	**97.0**	75.3	91.7	95.8	98.3	65.4	**84.6**	89.5	94.2
	k=5	51.2	80.0	88.3	94.3	88.2	95.5	98.2	99.3	67.8	87.0	92.7	96.8	73.3	89.9	**96.2**	**99.7**	60.3	81.5	88.3	92.9
	k=10	49.4	78.9	87.2	94.4	86.3	95.1	97.3	98.5	60.4	83.6	90.1	95.6	72.9	91.7	95.8	98.4	58.6	78.9	84.7	90.9
R=20	k=1	51.3	79.2	88.3	94.1	87.7	96.1	97.8	99.2	68.4	86.9	92.0	96.4	72.9	90.6	95.5	97.9	**66.7**	83.8	**90.7**	93.9
	k=3	51.8	80.0	87.9	**94.7**	**88.8**	96.3	**98.5**	**99.6**	69.5	87.9	92.9	**96.9**	76.0	91.7	95.1	98.3	64.3	83.5	89.3	**94.5**
	k=5	51.2	79.8	88.0	**94.7**	88.6	96.1	98.1	99.3	67.4	87.5	92.8	96.5	75.7	91.3	**96.5**	99.0	64.0	82.0	87.7	92.7
	k=10	49.1	80.0	87.2	94.2	86.8	95.3	97.3	98.8	63.8	85.5	91.2	96.2	71.7	88.5	94.3	97.7	58.5	78.9	84.8	91.0
R=50	k=1	51.1	79.3	88.0	94.3	87.1	95.0	97.3	98.8	68.9	87.9	92.5	96.2	**77.1**	**92.0**	**96.5**	**99.3**	**67.4**	**86.2**	**90.9**	**94.6**
	k=3	**52.5**	**80.3**	**88.5**	94.5	87.0	95.6	98.1	**99.4**	67.2	86.6	91.8	96.0	**76.7**	91.3	95.5	98.3	63.6	82.1	88.5	93.0
	k=5	48.0	79.4	87.4	94.1	88.5	95.3	97.6	**99.4**	67.7	86.3	92.5	95.9	68.4	89.9	95.1	97.9	61.7	81.6	88.5	92.9
	k=10	49.2	78.4	87.3	93.9	87.8	95.4	97.8	99.3	60.6	84.3	91.1	96.1	75.0	88.0	94.3	97.9	58.6	79.0	85.2	91.0
R=100	k=1	51.2	80.0	88.0	94.4	86.7	95.0	97.2	98.5	69.1	86.7	92.2	96.4	73.3	90.6	95.5	97.6	66.5	84.5	90.0	94.2
	k=3	**52.9**	**80.5**	**88.7**	**94.6**	88.5	**96.5**	**98.6**	**99.4**	**69.7**	**88.3**	**93.3**	96.4	75.3	91.3	95.5	98.3	65.2	83.8	89.9	94.2
	k=5	51.6	80.1	**88.5**	94.6	**89.0**	95.9	98.1	99.2	66.6	86.6	92.4	96.6	71.9	89.9	93.8	97.6	61.6	81.9	88.1	93.4
	k=10	49.5	79.6	87.6	94.0	86.4	95.0	97.0	98.8	62.3	85.4	91.6	96.0	67.7	89.6	93.2	96.4	58.6	78.9	85.2	90.9

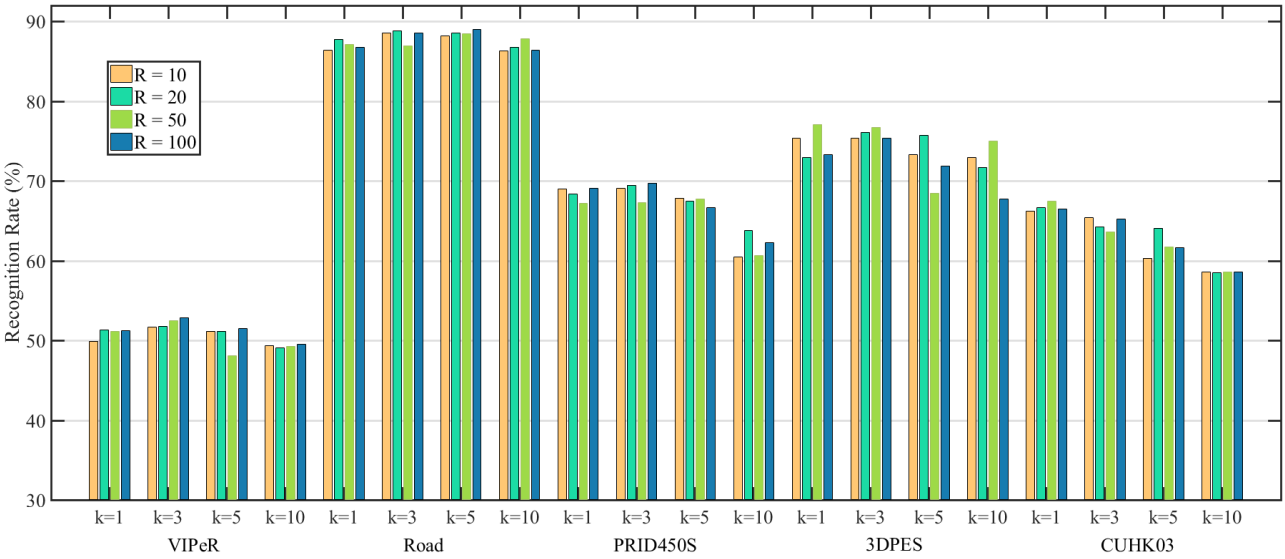


Fig. 11: Influence of different (R, k) settings on the rank-1 recognition performance on five benchmarks. Each group of bars illustrates the rank-1 performance at a fixed k with varying R, and each dataset contains four successive groups of bars.

TABLE III: Comparison results under different experimental settings to demonstrate the effectiveness of templates ensemble.

Settings \ Datasets	VIPeR			Road			PRID450S			3DPES			CUHK03		
	r=1	r=10	r=20	r=1	r=10	r=20	r=1	r=10	r=20	r=1	r=10	r=20	r=1	r=10	r=20
GCT_{best}	49.4	87.2	94.0	**88.8**	98.4	**99.6**	60.0	89.1	94.6	72.5	95.0	97.7	62.2	88.9	93.5
REGCT_{best}	**52.9**	**88.7**	**94.6**	88.5	**98.6**	99.4	**69.7**	**93.3**	**96.4**	**77.1**	**96.5**	**99.3**	**67.4**	**90.9**	**94.6**
GCT_{one}	46.0	85.1	93.1	83.1	96.3	98.8	58.4	84.3	89.8	69.8	95.5	97.2	61.9	87.6	92.8
REGCT_{one}	**51.2**	**88.0**	**94.4**	**87.7**	**97.8**	**99.2**	**69.1**	**92.2**	**96.4**	**77.1**	**96.5**	**99.3**	**67.4**	**90.9**	**94.6**

making REGCT more efficient than GCT when N is large. In this subsection, we demonstrate that template ensemble in the proposed REGCT also brings improvements to the recognition performance. In particular, we record the best recognition rates obtained with optimal parameter settings using both REGCT and the baseline GCT (we denote the experimental settings as REGCT_{best} and GCT_{best} respectively). Besides, to further demonstrate the superiority of the aggregated compact templates over the original noisy templates, we record the performance obtained by using one ensembled template and one original template respectively (denoted as REGCT_{one} and GCT_{one}). Please note for fair comparison, experiments under all the four settings are carried out with $(h, w, stride\_h, stride\_w)$ set to $(32, 32, 12, 8)$ respectively. The detailed comparison results are recorded in Table III.

As shown in Table III, the best performance of REGCT_{best} obtained using the optimal (R, k) settings on each dataset consistently outperforms the corresponding best results recorded in GCT_{best} [18], except slightly lower but competitive rank-1/rank-20 recognition rates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

TABLE IV: Comparison results of utilizing different body configuration comparison strategies for Re-ID. REGCT$_{best}$ records the results obtained by using the body orientation based strategy, while REGCT$_{best(w/p)}$ reports the results of utilizing the pose context based strategy.

Settings \ Datasets	VIPeR			Road			PRID450S			3DPES			CUHK03		
	r=1	r=10	r=20	r=1	r=10	r=20	r=1	r=10	r=20	r=1	r=10	r=20	r=1	r=10	r=20
REGCT$_{best}$	52.9	88.7	94.6	88.5	98.6	99.4	69.7	93.3	96.4	77.1	**96.5**	**99.3**	67.4	90.9	94.6
REGCT$_{best(w/p)}$	**53.5**	**89.1**	**94.7**	**88.9**	**98.9**	**99.6**	**70.9**	**93.5**	**96.5**	**78.0**	**96.5**	99.1	**68.0**	**91.2**	**94.9**

(a) (b) Orientation based results (c) Pose Context based results

Fig. 12: The top-10 ranked reference image pairs for two sample test pairs. (a) Two sample test image pairs. (b) The top-10 selected reference positive training pairs generated by the orientation based method. (c) The top-10 ranked reference image pairs selected by the pose context descriptor based method. As shown, the context descriptor generates reference images with more consistent body configurations.

on the Road dataset. As for the one template setting, compared to GCT$_{one}$, REGCT$_{one}$ significantly boosts the performance at all ranks across all the five benchmarks. This obvious performance gain validate that the ensembled compact templates are more robust and accurate than the original noisy matching templates, therefore benefiting the recognition performance by a large margin. Besides, in REGCT, the results of REGCT$_{one}$ are fairly close to REGCT$_{best}$, which means competitive recognition performance can be obtained with only one reference template (therefore significantly reducing the computational load). Please note that till now, all the experiments are recorded by using the body orientation based reference templates selection. The effectiveness of the proposed pose context descriptor based reference selection is discussed in the following part.

*5) **Evaluation on different body configuration comparison strategies:*** We also conduct experiments to demonstrate the superiority of the proposed novel pose context descriptor in modeling body configurations than the original orientation based method in [18]. We present both the qualitative and the quantitative study as follows:

Qualitative Study: For each test pair, we record the top-10 ranked reference pairs according to their pose-pair configuration similarities calculated by both the orientation based and the pose context based methods. As shown in Fig. 12, the orientation based method generates more noisy reference images (marked by the red bounding boxes) than the pose context based results.

Quantitative Study: We also record the recognition performance based on the two different reference selection strategies. The detailed comparison results on challenging datasets are presented in Table IV. As illustrated in Table IV, the pose context descriptor based reference selection can indeed boost the Re-ID performance. But the relatively slight performance gain indicate that the proposed REGCT algorithm is robust to different templates selection strategies. This indicates that we can save the efforts of trying to build more accurate body configuration descriptors, but pay more attention to improving the matching accuracy of the training templates in the future research.

TABLE V: Comparisons of top r matching rate using CMC (%) on VIPeR dataset. The best result is marked in red font, and the second best in blue.

Methods	r=1	r=5	r=10	r=20
SalMatch [3]	30.2	52.3	65.5	79.2
Semantic [51]	41.6	71.9	86.2	**95.1**
LSSCDL [52]	42.7	–	84.3	91.2
KISSME [5]	27.3	55.3	69.0	82.7
SVMML [6]	30.0	64.7	79.0	91.3
kLFDA [13]	32.4	65.9	79.8	90.8
Polymap [14]	36.8	70.4	83.7	91.7
LMF+LADF [53]	43.4	73.0	84.9	93.7
LOMO+XQDA [7]	40.0	68.1	80.5	91.1
DCSL [54]	44.6	73.4	82.6	–
TMA [11]	48.2	–	87.7	**95.5**
TCP [55]	47.8	74.7	84.8	91.1
DGD [56]	35.4	62.3	69.3	–
Spindle-Net [57]	**53.8**	74.1	83.2	92.1
DML^{2}V [58]	50.4	**80.5**	**88.7**	95.0
CSL [15]	34.8	68.7	82.3	91.8
REGCT$_{best(w/p)}$	**53.5**	**81.3**	**89.1**	94.7

C. *Comparison with state-of-the-arts*

To demonstrate the effectiveness of the proposed approach, we compare the proposed REGCT with some state-of-the-art approach-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

es on five challenging datasets. Please note results generated by REGCT$_{best(w/p)}$ are utilized in the following comparisons. And the detailed comparison results are presented as follows.

On the VIPeR dataset, we compare the REGCT with other sixteen algorithms, including SalMatch [3], Semantic [51], LSS-CDL [52], KISSME [5], SVMML [6], kLFDA [13], Polymap [14], LMF+LADF [53], LOMO+XQDA [7], DCSL [54], TMA [11], TCP [55], DGD [56], Spindle Net [57], DML^{2}V [58] and CSL [15]. The comparison results are presented in Table V. As illustrated in Table V, the proposed REGCT algorithm achieves the best recognition rate at rank 5, 10, and competitive performances at rank 1, 20. Please note DCSL [54], TCP [55], DGD [56], and Spindle Net [57] are deep feature based end-to-end framework, the favorable performance of REGCT against these algorithms demonstrates the effectiveness of our system for Re-ID. Besides, compared with CSL [15] and DCSL [54], which also aim to establish local semantic correspondences, our REGCT evidences notable performance gain. This validates the superiority of exploiting contextual information via graph matching to address the spatial misalignment problem.

TABLE VI: Comparison of top r matching rate using CMC (%) on Road dataset. The best result is marked in red font, and the second best in blue.

Methods	r=1	r=5	r=10	r=20
eSDC-knn [3]	52.4	74.5	83.7	89.9
CSL [15]	**61.5**	**91.8**	**95.2**	**98.6**
REGCT$_{best(w/p)}$	**88.9**	**96.8**	**98.9**	**99.6**

The Road dataset is proposed in CSL [15]. For comprehensive comparison, we also report the result on this dataset, and compare it with eSDC-knn [3] and CSL [15]. As shown in Table VI, compared to CSL [15], our algorithm obtains significant improvements of 27.4%, 5.0%, 3.7%, 1.0% at rank 1, 5, 10, 20 respectively. Owing to the sample-specific patch-wise matching adopted in our algorithm, significant performance gain is achieved compared with the camera-specific global matching structure adopted in CSL [15].

On the PRID450S dataset, we compare with KISSME [5], SC-NCDFinal [12], Semantic [51], TMA [11], NSFT [59], DML^{2}V [58] and CSL [15]. As shown in Table VII, our algorithm achieves the best recognition performance at all ranks. This obvious performance gain can be attributed to the accurate patch-wise correspondence via graph matching as well as the robust correspondence transfer via template ensemble.

TABLE VII: Comparison of top r matching rate using CMC (%) on the PRID450S dataset. The best result is marked in red font, and the second best in blue.

Methods	r=1	r=5	r=10	r=20
KISSME [5]	33	–	71	79
SCNCDFinal [12]	41.6	68.9	79.4	87.8
Semantic [51]	44.9	71.7	77.5	86.7
TMA [11]	54.2	73.8	83.1	90.2
NSFT [59]	40.9	64.7	73.2	81.0
DML^{2}V [58]	**64.5**	**85.7**	**92.1**	**96.0**
CSL [15]	44.4	71.6	82.2	89.8
REGCT$_{best(w/p)}$	**70.9**	**89.1**	**93.5**	**96.5**

On the 3DPES dataset, we compare the REGCT method with state-of-the-arts including LFDA [10], ME [9], kLFDA [13], PCCA [8], rPCCA [13], SCSP [4], WARCA [60], DGD [56], Spindle Net [57] and CSL [15]. As shown in Table VIII, the proposed algorithm significantly outperforms the state-of-the-art algorithms, and even deep learning based algorithm [57]. Note that the images in this dataset are automatic detection results from videos captured under eight cameras, bringing serious pose variations, illumination changes and scale variations. With the help of the learned correspondences templates, our REGCT model is robust against these issues.

TABLE VIII: Comparison of top r matching rate using CMC (%) on 3DPES dataset. The best result is marked in red font, and the second best in blue.

Methods	r=1	r=5	r=10	r=20
LFDA [10]	45.5	69.2	–	86.1
ME [9]	53.3	76.8	–	92.8
kLFDA [13]	54.0	77.7	85.9	92.4
PCCA [8]	41.6	70.5	81.3	90.4
rPCCA [13]	47.3	75.0	84.5	91.9
SCSP [4]	57.3	79.0	–	91.5
WARCA [60]	51.9	75.6	–	–
DGD [56]	56.0	–	–	–
Spindle-Net [57]	**62.1**	**83.4**	**90.5**	**95.7**
CSL [15]	57.9	81.1	89.5	93.7
REGCT$_{best(w/p)}$	**78.0**	**92.4**	**96.5**	**99.1**

TABLE IX: Comparison of top r matching rate using CMC (%) on CUHK01 dataset. The best result is marked in red font, and the second best in blue.

Methods	r=1	r=5	r=10	r=20
Semantic [51]	32.7	51.2	–	76.3
kLFDA [13]	32.8	59.0	69.6	–
IDLA [36]	47.5	71.5	80.0	–
DeepRanking [37]	50.4	75.9	84.1	–
ME [9]	53.4	76.3	84.4	–
GOG [61]	57.8	79.1	86.2	–
SalMatch [3]	28.5	46.0	–	67.3
CSBT [62]	51.2	76.3	–	91.8
TCP [55]	53.7	84.3	91.0	**96.3**
DML^{2}V [58]	**65.0**	**85.6**	**91.1**	**95.1**
REGCT$_{best(w/p)}$	**68.0**	**86.9**	**91.2**	94.9

On the CUHK01 dataset, we compare with Semantic [51], kLFDA [13], IDLA [36], DeepRanking [37], ME [9], GOG [61], SalMatch [3], CSBT [62], TCP [55] and DML^{2}V [58]. The detailed comparison results are presented in Table IX. As shown in Table IX, the proposed graph matching and correspondence transfer framework can achieve favorable results on this medium-sized dataset. More specifically, the proposed REGCT algorithm obtains best rank 1 recognition rate (a 14.3% performance gain over TCP [55], a part based deep learning algorithm). And the better performance compared with other algorithms also demonstrate the superiority of the REGCT model.

D. Analysis on Typical Failure Cases

We record the typical failure cases to explore the limitations of the proposed REGCT algorithm. As shown in Fig. 13, when severe self-occlusion occurs, the appearances of the same person may be dramatically different across different camera views, rendering it difficult for REGCT to establish enough local correspondences between matched image pairs. Even though, the proposed REGCT algorithm can rank visually similar images with the probe ahead of others, which is valuable for further manual verification.

V. Conclusion

This paper proposes a robust and efficient graph correspondence transfer (REGCT) approach to explicitly address the spatial mis-

Fig. 13: Typical failure cases. The first image in each row is the probe image, the second one is the correctly matched gallery image, followed by the ranking list obtained by REGCT.

alignment issue in Re-ID. The framework of *off-line* patch-wise correspondence learning and *on-line* correspondence transfer helps to flexibly establish robust and accurate patch-level matching patterns for each test pair. The proposed template ensemble strategy is demonstrated to improve the efficiency as well as notably boost the recognition performance compared to the baseline GCT. The proposed pose context descriptor further benefits the REGCT model via more accurate templates selection, leading to more robust patch-wise correspondence transfer in the testing phase. Extensive experiments on five challenging datasets demonstrate the effectiveness of the REGCT model.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 61671289, 61221001, 61528204, 61771303, 61521062 and 61571261), STCSM (18DZ2270700) and by US National Science Foundation (NSF) (Grant No.1618398 and 1350521).

References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010.
[2] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015.
[3] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.
[4] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016.
[5] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.
[6] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *CVPR*, 2013.
[7] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.
[8] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.
[9] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, 2015.
[10] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013.
[11] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *ECCV*, 2016.
[12] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014.
[13] F. Xiong, M. Gou, O. I. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014.
[14] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *CVPR*, 2015.
[15] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *ICCV*, 2015.
[16] O. Oreifej, R. Mehran, and M. Shah, "Human identity recognition in aerial images," in *CVPR*, 2010.
[17] Y. Yang, L. Wen, S. Lyu, and S. Z. Li, "Unsupervised learning of multi-level descriptors for person re-identification," in *AAAI*, 2017.
[18] Q. Zhou, H. Fan, S. Zheng, H. Su, X. Li, S. Wu, and H. Ling, "Graph correspondence transfer for person re-identification," in *AAAI*, 2018.
[19] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.
[20] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv*, 2016.
[21] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.
[22] N. Gheissari, T. B. Sebastian, and R. I. Hartley, "Person reidentification using spatiotemporal appearance," in *CVPR*, 2006.
[23] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV*, 2012.
[24] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by HPE signature," in *ICPR*, 2010.
[25] S. Bak, E. Corvée, F. Brémond, and M. Thonnat, "Multiple-shot human re-identification by mean riemannian covariance grid," in *AVSS*, 2011.
[26] A. Li, L. Liu, and S. Yan, "Person re-identification by attribute-assisted clothes appearance," in *Person Re-Identification*, 2014, pp. 119–138.
[27] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *CVPR*, 2014.
[28] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[30] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[31] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," in *ICCV*, 2017.
[32] ——, "Sanet: structure-aware network for visual tracking," in *CVPRW*, 2017.
[33] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *ICCV*, 2015.
[34] H. Fan, X. Mei, D. Prokhorov, and H. Ling, "Multi-level contextual rnns with attention model for scene labeling," *IEEE Transactions on Intelligent Transportation Systems*, 2018.
[35] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.
[36] E. Ahmed, M. J. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.
[37] S. Chen, C. Guo, and J. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.
[38] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016.
[39] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: models and global optimization," in *ECCV*, 2008.
[40] O. Duchenne, F. R. Bach, I. Kweon, and J. Ponce, "A tensor-aased algorithm for high-order graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2383–2395, 2011.
[41] T. Wang, H. Ling, C. Lang, and S. Feng, "Graph matching with adaptive and branching path following," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
[42] F. Zhou and F. D. la Torre, "Factorized graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1774–1789, 2016.
[43] T. Wang, H. Ling, C. Lang, and S. Feng, "Graph matching with adaptive and branching path following," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
[44] M. Cho, J. Lee, and K. M. Lee, "Reweighted random walks for graph matching," in *ECCV*, 2010, pp. 492–505.
[45] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.
[46] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *CVPR*, 2010.
[47] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS Workshop*, 2007.
[48] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2914575, IEEE Transactions on Image Processing

[49] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *ACM Workshop on MA3HO*, 2011.
[50] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.
[51] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *CVPR*, 2015.
[52] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *CVPR*, 2016.
[53] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014.
[54] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification," in *IJCAI*, 2016.
[55] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.
[56] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016.
[57] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wng, and X. Tang, "Spindle net: person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017.
[58] C. Sun, D. Wang, and H. Lu, "Person re-identification via distance metric learning with latent variables," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 23–34, 2017.
[59] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.
[60] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *ECCV*, 2016.
[61] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016.
[62] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao, "Fast person re-identification via cross-camera semantic binary transformation," in *CVPR*, 2017.

Qin Zhou received her B.S. degree in information engineering from Xi'an Jiao Tong University, Xi'an, China, in 2013. She is currently a Ph.D. student at the Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include computer vision, machine learning, convex optimization and person re-identification. She is now a visiting student in Professor Haibin Ling's lab at Temple University.

Heng Fan received his B.E. degree in College of Science, Huazhong Agricultural University (HZAU), Wuhan, China, in 2013. He is currently a Ph.D. student in the Department of Computer and Information Science, Temple University, Philadelphia, USA. His research interests include computer vision, pattern recognition and machine learning.

Hua Yang received her Ph.D. degree in communication and information from Shanghai Jiaotong University, in 2004, and both the B.S. and M.S. degrees in communication and information from Haerbin Engineering University, China in 1998 and 2001, respectively. She is currently an associate professor in the Department of Electronic Engineering, Shanghai Jiaotong University, China. Her current research interests include video coding and networking, computer vison, and smart video surveillance.

Hang Su , received his Ph. D Degree in Electronic Engineering from Shanghai Jiaotong University in 2014. He is currently working as a PostDoc in department of Computer Science, Tsinghua University. His current research interests include computer vision and large-scale machine learning. Mr. Hang Su has also served as TPC memember in several conferences including IJCAI, AAAI and UAI, and contributed as a reviewer for TPAMI, CVPR, NIPS, ICML, etc.

Shibao Zheng received his B.S. degree in communication engineering from Xidian University, Xi'an and M.S.degree in the signal and information processing from the 54th institute of CETC, Shijiazhuang, China, in 1983 and 1986, respectively. He is currently a professor of electronic engineering department and vice director of Elderly Health Information and Technology Institute, Shanghai Jiao Tong University (SJTU), Shanghai, China. And he is also a professor committee member of Shanghai Key Laboratory of Digital Media Processing and Transmission, and a Consultant Expert of ministry of public security in video surveillance field. His current research interests include urban video surveillance system, intelligent video analysis, and elderly health technology, etc.

Shuang Wu received the BS degree in electronic information engineering from the Southeast University (SEU), Nanjing, China, in 2010. He is currently a Ph.D. student at the Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include computer vision, machine learning and crowd behavior analysis. He is now a researcher at Youtu Lab, Tencent.

Haibin Ling received the BS and MS degrees from Peking University, China, in 1997 and 2000, respectively, and the PhD degree from the University of Maryland College Park in 2006. From 2000 to 2001, he was an assistant researcher at Microsoft Research Asia. From 2006 to 2007, he worked as a postdoctoral scientist at the University of California Los Angeles. After that, he joined Siemens Corporate Research as a research scientist. Since fall 2008, he has been with Temple University where he is now an Associate Professor. Ling's research interests include computer vision, augmented reality, medical image analysis, and human computer interaction. He received the Best Student Paper Award at the ACM Symposium on User Interface Software and Technology (UIST) in 2003, and the NSF CAREER Award in 2014. He serves as associate editors for *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Pattern Recognition*, and *Computer Vision and Image Understanding*, and has served as area chairs for CVPR 2014 and CVPR 2016.