Person Re-identification with Correspondence Structure Learning

Yang Shen[1], Weiyao Lin[1], Junchi Yan[1], Mingliang Xu[2], Jianxin Wu[3] and Jingdong Wang[4]

[1]Shanghai Jiao Tong University, China [2]Zhengzhou University, China
[3]Nanjing University, China [4]Microsoft Research Asia

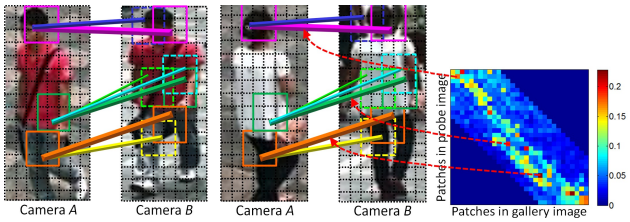
Abstract

This paper addresses the problem of handling spatial misalignments due to camera-view changes or human-pose variations in person re-identification. We first introduce a boosting-based approach to learn a correspondence structure which indicates the patch-wise matching probabilities between images from a target camera pair. The learned correspondence structure can not only capture the spatial correspondence pattern between cameras but also handle the viewpoint or human-pose variation in individual images. We further introduce a global-based matching process. It integrates a global matching constraint over the learned correspondence structure to exclude cross-view misalignments during the image patch matching process, hence achieving a more reliable matching score between images. Experimental results on various datasets demonstrate the effectiveness of our approach.

1. Introduction

Person re-identification (Re-ID) is of increasing importance in visual surveillance. The goal of person Re-ID is to identify a specific person indicated by a probe image from a set of gallery images captured from cross-view cameras (i.e., cameras that are non-overlapping and different from the probe image's camera).[1] It remains challenging due to the large appearance changes in different camera views and the interferences from background or object occlusion.

One major challenge for person Re-ID is the uncontrolled spatial misalignment between images due to camera-view changes or human-pose variations. For example, in Fig. 1a, the green patch located in the lower part in camera A's image corresponds to patches from the upper part in camera B's image. However, most existing works [25, 11, 12, 7, 8, 9, 22, 19] focus on handling the overall appearance variations between images, while the spatial misalignment among images' local patches is not addressed. Although some patch-based methods [17, 15, 27] address the spatial misalignment problem by decomposing images into patches and performing an online patch-level matching, their performances are often restrained by the online matching process which is easily affected by the mismatched patches due to similar appearance or occlusion.



(a) (b) (c)

Figure 1. (a) and (b): Two examples of using a correspondence structure to handle spatial misalignments between images from a camera pair. Images are obtained from the same camera pair: A and B. The colored squares represent sample patches in each image while the lines between images indicate the matching probability between patches (line width is proportional to the probability values). (c): The correspondence structure matrix including all patch matching probabilities between A and B (the matrix is down-sampled for a clearer illustration). (Best viewed in color)

In this paper, we argue that due to the stable setting of most cameras (e.g., fixed camera angle or location), each camera has a stable constraint on the spatial configuration of its captured images. For example, images in Figures 1a and 1b are obtained from the same camera pair: A and B. Due to the constraint from camera angle difference, body parts in camera A's images are located at lower places than those in camera B, implying a lower-to-upper correspondence pattern between them. Meanwhile, constraints from camera locations can also be observed. Camera A (which monitors an exit region) includes more side-view images, while camera B (monitoring a road) shows more front or back-view images. This further results in a high probability of side-to-front/back correspondence pattern.

Based on this intuition, we propose to learn a correspondence structure (i.e., a matrix including *all* patch-wise matching probabilities between a camera pair, as Fig. 1c) to encode the spatial correspondence pattern constrained by a

[1]In this paper, an image refers to the pixel region of one person which is cropped from a larger image of a camera view (cf. Fig. 1) [6].

camera pair, and utilize it to guide the patch matching and matching score calculation processes between images. With this correspondence structure, spatial misalignments can be suitably handled and patch matching results are less interfered by the confusion from appearance or occlusion. In order for the correspondence structure to model human-pose variations or local viewpoint changes inside a camera view, the correspondence structure for each patch is described by a one-to-many graph whose weights indicate the matching probabilities between patches, as in Fig. 1. Besides, a global constraint is also integrated during the patch matching process, so as to achieve a more reliable matching score between images. Note that our approach is not limited to person re-identification with fixed camera settings. Instead, it can also be applied to capture the camera-and-person configuration and cross-view correspondence for unfixed cameras, as demonstrated in the experimental results.

In summary, our contributions to person Re-ID are three folds. First, we introduce a correspondence structure to encode cross-view correspondence pattern between cameras, and develop a global-based matching process by combining a global constraint with the correspondence structure to exclude spatial misalignments between images. These two components in fact establish a novel framework for addressing the person Re-ID problem. Second, under this framework, we propose a boosting-based approach to learn a suitable correspondence structure between a camera pair. The learned correspondence structure can not only capture the spatial correspondence pattern between cameras but also handle the viewpoint or human-pose variation in individual images. Third, this paper releases a new and challenging benchmark ROAD DATASET for person Re-ID.

The rest of this paper is organized as follows. Sec. 2 reviews related works. Sec. 3 describes the framework of the proposed approach. Sections 4 to 5 describe the details of our proposed global-based matching process and boosting-based learning approach, respectively. Sec. 6 shows the experimental results and Sec. 7 concludes the paper.

2. Related Works

Many person re-identification methods have been proposed. Most of them focus on developing suitable feature representations about humans' appearance [25, 11, 12, 7, 14], or finding proper metrics to measure the cross-view appearance similarity between images [8, 9, 22, 19]. Since these works do not effectively model the spatial misalignment among local patches inside images, their performances are often limited due to the interferences from viewpoint changes and human-pose variations.

In order to address the spatial misalignment problem, some patch-based methods are proposed [23, 17, 3, 15, 27, 26, 5, 20] which decompose images into patches and perform an online patch-level matching to exclude patch-wise misalignments. In [23, 3], a human body in an image is first parsed into semantic parts (e.g., head and torso). And then, similarity matching is performed between the corresponding semantic parts. Since these methods are highly dependent on the accuracy of body parser, they have limitations in scenarios where the body parser does not work reliably.

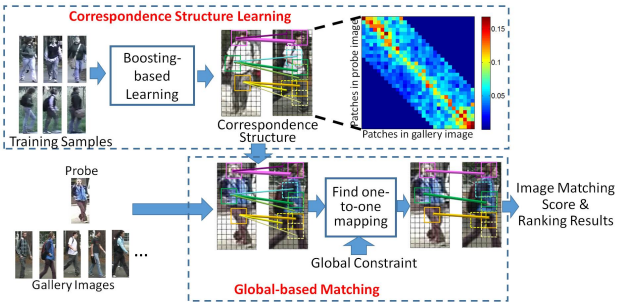


Figure 2. Framework of the proposed approach.

In [17], Oreifej et al. divide images into patches according to appearance consistencies and utilize the Earth Movers Distance (EMD) to measure the overall similarity among the extracted patches. However, since the spatial correlation among patches are ignored during similarity calculation, the method is easily affected by the mismatched patches with similar appearance. Although Ma et al. [15] introduce a body prior constraint to avoid mismatching between distant patches, the problem is still not well addressed, especially for the mismatching between closely located patches.

To reduce the effect of patch-wise mismatching, some saliency-based approaches [27, 26] are recently proposed, which estimate the saliency distribution relationship between images and utilize it to control the patch-wise matching process. Although these methods consider the correspondence constraint between patches, our approach differs from them in: (1) our approach focuses on constructing a correspondence structure where patch-wise matching parameters are jointly decided by both matched patches. Comparatively, the matching weights in the saliency-based approach [26] is only controlled by patches in the probe-image (probe patch). (2) Our approach models patch-wise correspondence by a one-to-many graph such that each probe patch will trigger multiple matches during the patch matching process. In contrast, the saliency-based approaches only select one best-matched patch for each probe patch. (3) Our approach introduces a global constraint to control the patch-wise matching result while the patch matching result in saliency-based approaches is locally decided by choosing the best-matched one within a neighborhood set.

3. Overview

The framework of our approach is shown in Fig. 2. During the training process, which is detailed in Section 5, we

present a boosting-based process to learn the correspondence structure between the target camera pair. During the prediction stage, which is detailed in Section 4 given a probe image and a set of gallery images, we use the correspondence structure to evaluate the patch correlations between the probe image and each gallery image, and find the optimal one-to-one mapping between patches, and accordingly the matching score. The Re-ID result is achieved by ranking gallery images according to their matching scores.

4. Person Re-Identification with Correspondence Structure

In this section, we introduce the concept of correspondence structure, show the scheme of computing the patch correlation using the correspondence structure, and finally present the patch-wise mapping method to compute the matching score between the probe image and the gallery image.

4.1. Correspondence structure

The correspondence structure, $\mathbf{\Theta}_{A,B}$, encodes the spatial correspondence distribution between a pair of cameras, A and B. In our problem, we adopt a discrete distribution, which is a set of patch-wise matching probabilities, $\mathbf{\Theta}_{A,B} = \{P(x_i^A, B)\}_{i=1}^{N_A}$, where N_A is the number of patches of an image in camera A. $P(x_i^A, B) = \{P(x_i^A, x_1^B), P(x_i^A, x_2^B), \ldots, P(x_i^A, x_{N_B}^B)\}$ describes the correspondence distribution in an image from camera B for the ith patch x_i^A of an image captured from camera A, where N_B is the number of patches of an image in B. An illustration of the correspondence distribution is shown on the top-right of Fig. 1c.

The definition of the matching probabilities in the correspondence structure only depends on a camera pair and are independent to the specific images. In the correspondence structure, it is possible that one patch in camera A is highly correlated to multiple patches in camera B, so as to handle human-pose variations and local viewpoint changes in a camera view.

4.2. Patch correlation

Given a probe image U in camera A and a gallery image V in camera B, the patch-wise correlation between U and V, $C(x_i^U, x_j^V)$, is computed from both the correspondence structure between cameras A and B and the visual features and written as:

$$C(x_i^U, x_j^V) = \lambda_{T_c}(P(x_i^U, x_j^V)) \cdot \log \Phi(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V}; x_i^U, x_j^V). \quad (1)$$

Here x_i^U and x_j^V are ith and jth patch in images U and V; $\mathbf{f}_{x_i^U}$ and $\mathbf{f}_{x_j^V}$ are the feature vectors for x_i^U and x_j^V. $P(x_i^U, x_j^V) = P(x_i^A, x_j^B)$ is the correspondence structure of cameras A and B. $\lambda_{T_c}(P(x_i^U, x_j^V) = 1$ $P(x_i^U, x_j^V) > T_c$, and 0 otherwise, and $T_c = 0.05$ is a threshold. $\Phi(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V}; x_i^U, x_j^V)$ is the correspondence-structure-controlled similarity between x_i^U and x_j^V,

$$\Phi(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V}; x_i^U, x_j^V) = \Phi_z(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V}) P(x_i^U, x_j^V), \quad (2)$$

where $\Phi_z(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V})$ is the appearance similarity between x_i^U and x_j^V.

The correspondence structure $P(x_i^U, x_j^V)$ in Equations 1 and 2, is used to adjust the appearance similarity $\Phi_z(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V})$ such that a more reliable patch-wise correlation strength can be achieved. The thresholding term $\lambda_{T_c}(P(x_i^U, x_j^V))$ is introduced to exclude the patch-wise correlation with a low correspondence probability, which effectively reduces the interferences from mismatched patches with similar appearance.

The patch-wise appearance similarity $\Phi_z(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V})$ in Eq. 2 can be achieved by many off-the-shelf methods [27, 26, 2]. In this paper, we extract Dense SIFT and Dense Color Histogram [27] from each patch and utilize the KISSME distance metric [9] to compute $\Phi_z(\Phi_z(\mathbf{f}_{x_i^U}, \mathbf{f}_{x_j^V}))$ (note that we train different KISSME metrics for patch-pairs at different locations).

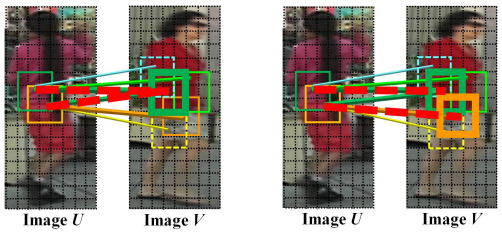
4.3. Patch-wise mapping

With $C(x_i^U, x_j^V)$, the alignment-enhanced correlation strength, we can find a best-matched patch in image V for each patch in U and herein calculate the final image matching score. However, locally finding the largest $C(x_i^U, x_j^V)$ may still create mismatches among patch pairs with high matching probabilities. For example, Fig. 3a shows an image pair U and V containing different people. When locally searching for the largest $C(x_i^U, x_j^V)$, the yellow patch in U will be mismatched to the bold-green patch in V since they have both large appearance similarity and high matching probability. This mismatch unsuitably increases the matching score between U and V.

To address this problem, we introduce a global one-to-one mapping constraint and solve the resulting linear assignment task [10] to find the best matching:

$$\mathbf{\Omega}_{U,V}^* = \arg\max_{\mathbf{\Omega}_{U,V}} \sum_{\{x_i^U, x_j^V\} \in \mathbf{\Omega}_{U,V}} C(x_i^U, x_j^V) \quad (3)$$
$$\text{s.t. } x_i^U \neq x_s^U, x_j^V \neq x_t^V \quad \forall \{x_i^U, x_j^V\}, \{x_s^U, x_t^V\} \in \mathbf{\Omega}_{U,V}$$

where $\mathbf{\Omega}_{U,V}^*$ is the set of the best patch matching result between images U and V. $\{x_i^U, x_j^V\}$ and $\{x_s^U, x_t^V\}$ are two matched patch pairs in $\mathbf{\Omega}$. According to Eq. 3, we want to find the best patch matching result $\mathbf{\Omega}_{U,V}^*$ that maximizes the



(a) (b)

Figure 3. Patch matching result (a) by locally finding the largest correlation strength $C(x_i^U, x_j^V)$ for each patch and (b) by using a global constraint. The red dashed lines indicate the final patch matching results and the colored solid lines are the matching probabilities in the correspondence structure. (Best viewed in color)

total image matching score

$$\psi_{U,V} = \sum_{\{x_i^U, x_j^V\} \in \mathbf{\Omega}_{U,V}} C(x_i^U, x_j^V), \tag{4}$$

given that each patch in U can only be matched to one patch in V and vice versa.

Eq. 3 can be solved by the Hungary method [10]. Fig. 3b shows an example of the patch matching result by Eq. 3. From Fig. 3b, it is clear that by the inclusion of a global constraint, local mismatches can be effectively reduced and a more reliable image matching score can be achieved. Based on the above process, we can calculate the image matching scores ψ between a probe image and all gallery images in a cross-view camera, and rank the gallery images accordingly to achieve the final Re-ID result [15].

5. Correspondence Structure Learning

5.1. Objective function

Given a set of probe images $\{U_\alpha\}$ from camera A and their corresponding cross-view images $\{V_\beta\}$ from camera B in the training set, we learn the optimal correspondence structure $\mathbf{\Theta}^*_{A,B}$ between cameras A and B so that the correct match image is ranked before the incorrect match images in terms of the matching scores. The formulation is give as below,

$$\min_{\mathbf{\Theta}_{A,B}} \sum_{U_\alpha} R(V_{\alpha'}; \psi_{U_\alpha, V_{\alpha'}}(\mathbf{\Theta}_{A,B}), \mathbf{\Psi}_{U_\alpha, V_{\beta \neq \alpha'}}(\mathbf{\Theta}_{A,B})), \tag{5}$$

where $V_{\alpha'}$ is the correct match gallery image of the probe image U_α. $\psi_{U_\alpha, V_{\alpha'}}(\mathbf{\Theta}_{A,B})$ (as computed from Eq. 4) is the matching score between U_α and $V_{\alpha'}$ and $\mathbf{\Psi}_{U_\alpha, V_{\beta \neq \alpha'}}(\mathbf{\Theta}_{A,B})$ is the set of matching scores of all incorrect match images. $R(V_{\alpha'}; \psi_{U_\alpha, V_{\alpha'}}(\mathbf{\Theta}_{A,B}), \mathbf{\Psi}_{U_\alpha, V_{\beta \neq \alpha'}}(\mathbf{\Theta}_{A,B}))$ is the rank of $V_{\alpha'}$ among all the gallery images according to the matching scores. Intuitively, the penalty is the smallest if the rank is 1, i.e., the matching score of $V_{\alpha'}$ is the greatest. The optimization is not easy as the matching score calculation (Eq. 4) is complicated. We present an approximate solution, a boosting-based process, to solve this problem.

5.2. Boosting-based learning

The boosting-based approach utilizes a progressive way to find the best correspondence structure with the help of *binary mapping structures*. A binary mapping structure is similar to the correspondence structure except that it simply utilizes 0 or 1 instead of matching probabilities to indicate the connectivity or linkage between patches, cf. Fig. 4a. It can be viewed as a simplified version of the correspondence structure which includes rough information about the cross-view correspondence pattern.

Since binary mapping structures only include simple connectivity information among patches, their optimal solutions are tractable for individual probe images. Therefore, by searching for the optimal binary mapping structures for different probe images and utilizing them to progressively update the correspondence structure, suitable cross-view correspondence patterns can be achieved.

The entire boosting-based learning process can be describe by the following steps as well as Algorithm 1.

Finding the optimal binary mapping structure. For each training probe image U_α, we first create multiple candidate binary mapping structures under different search ranges by adjacency-constrained search [27], and then find the optimal binary mapping structure $\mathbf{M}_\alpha$ such that the rank order of U_α's correct match image $V_{\alpha'}$ is minimized under $\mathbf{M}_\alpha$. Note that we find one optimal binary mapping structure for each probe image such that the obtained binary mapping structures can include local cross-view correspondence information in different training samples.

Correspondence Structure Initialization. In this paper, patch-wise matching probabilities $P(x_i^U, x_j^V)$ in the correspondence structure are initialized by:

$$P^0(x_i^U, x_j^V) \propto \begin{cases} 0, & \text{if } d(x_i^V, x_j^V) \geq T_d \\ \dfrac{1}{d(x_i^V, x_j^V) + 1}, & \text{otherwise} \end{cases}, \tag{6}$$

where x_i^V is x_i^U's co-located patch in camera B. $d(x_i^V, x_j^V)$ is the distance between patches x_i^V and x_j^V. It is defined as the number of strides to move from x_i^V to x_j^V in the zig-zag order. T_d is a threshold which is set to be 32 in this paper. According to Eq. 6, $P^0(x_i^U, x_j^V)$ is inversely proportional to the co-located distance between x_i^V and x_j^V and will equal to 0 if the distance is larger than a threshold.

Binary mapping structure selection. During each iteration k in the learning process, we first apply correspondence structure $\mathbf{\Theta}^{k-1}_{A,B} = \{P^{k-1}(x_i^U, x_j^V)\}$ from the previous iteration to calculate the rank orders of all correct match

Algorithm 1 Boosting-based Learning Process

Input: A set of training probe images $\{U_\alpha\}$ from camera A and their corresponding cross-view images $\{V_\beta\}$ from camera B
Output: $\mathbf{\Theta}_{A,B} = \{P(x_i^U, y_j^V)\}$, the correspondence structure between $\{U_\alpha\}$ and $\{V_\beta\}$

1: Find an optimal binary mapping structure $\mathbf{M}_\alpha$ for each probe image U_α, as described in the 4-th paragraph in Sec 5.2
2: Set $k = 1$. Initialize $P^0(x_i^U, y_j^V)$ by Eq. 6.
3: Use the current correspondence structure $\{P^{k-1}(x_i^U, x_j^V)\}$ to perform Re-ID on $\{U_\alpha\}$ and $\{V_\beta\}$, and select 20 binary mapping structures $\mathbf{M}_\alpha$ based on the Re-ID result, as described in the 6-th paragraph in Sec 5.2
4: Compute updated match probability $\hat{P}^k(x_i^U, x_j^V)$ by Eq. 7
5: Update the matching probabilities $P^k(x_i^U, x_j^V)$ by Eq. 12
6: Set $k = k + 1$ and go back to step 3 if not converged or not reaching the maximum iteration number
7: Output $\{P(x_i^U, y_j^V)\}$

images $V_{\alpha'}$ in the training set. Then, we randomly select 20 $V_{\alpha'}$ where half of them are ranked among top 50% (implying better Re-ID results) and another half are ranked among the last 50% (implying worse Re-ID results). Finally, we extract binary mapping structures corresponding to these selected images and utilize them to update and boost the correspondence structure.

Note that we select binary mapping structures for both high- and low-ranked images in order to include a variety of local patch-wise correspondence patterns. In this way, the final obtained correspondence structure can suitably handle the variations in human-pose or local viewpoints.

Calculating the updated matching probability. With the introduction of the binary mapping structure $\mathbf{M}_\alpha$, we can model the updated matching probability in the correspondence structure by:

$$\hat{P}^k(x_i^U, x_j^V) = \sum_{\mathbf{M}_\alpha \in \mathbf{\Gamma}^k} \hat{P}(x_i^U, x_j^V | \mathbf{M}_\alpha) \cdot P(\mathbf{M}_\alpha), \quad (7)$$

where $\hat{P}^k(x_i^U, x_j^V)$ is the updated matching probability between patches x_i^U and x_j^V in the k-th iteration. $\mathbf{\Gamma}^k$ is the set of binary mapping structures selected in the k-th iteration. $P(\mathbf{M}_\alpha) = \frac{\tilde{\mathcal{R}}_n(\mathbf{M}_\alpha)}{\sum_{\mathbf{M}_\gamma \in \mathbf{\Gamma}^k} \tilde{\mathcal{R}}_n(\mathbf{M}_\gamma)}$ is the prior probability for binary mapping structure $\mathbf{M}_\alpha$, where $\tilde{\mathcal{R}}_n(\mathbf{M}_\alpha)$ is the CMC score at rank n [21] when using $\mathbf{M}_\alpha$ as the correspondence structure to perform person Re-ID over the training images. n is set to be 5 in our experiments.

$\hat{P}(x_i^U, x_j^V | \mathbf{M}_\alpha)$ is the updated matching probability between x_i^U and x_j^V when including the local correspondence pattern information of $\mathbf{M}_\alpha$. It can be calculated by:

$$\hat{P}(x_i^U, x_j^V | \mathbf{M}_\alpha) = \hat{P}(x_j^V | x_i^U, \mathbf{M}_\alpha) \cdot \hat{P}(x_i^U | \mathbf{M}_\alpha), \quad (8)$$

$\hat{P}(x_j^V | x_i^U, \mathbf{M}_\alpha)$ is the updated probability to correspond from x_i^U to x_j^V when including $\mathbf{M}_\alpha$, calculated as

$$\hat{P}(x_j^V | x_i^U, \mathbf{M}_\alpha) \propto \begin{cases} 1, & \text{if } m_{\{x_i^U, x_j^V\}} \in \mathbf{M}_\alpha \\ \tilde{\mathcal{A}}_{x_j^V | x_i^U, \mathbf{M}_\alpha}, & \text{otherwise} \end{cases}, \quad (9)$$

where $m_{\{x_i^U, x_j^V\}}$ is a patch-wise link connecting x_i^U and x_j^V. $\tilde{\mathcal{A}}_{x_j^V | x_i^U, \mathbf{M}_\alpha} = \frac{\overline{\Phi}_z(x_i^U, x_j^V)}{\sum_{x_t^V, m_{\{x_i^U, x_t^V\}} \in \mathbf{M}_\alpha} \overline{\Phi}_z(x_i^U, x_t^V)}$, where $\overline{\Phi}_z(x_i^U, x_j^V)$ is the average appearance similarity [27, 9] between patches x_i^U and x_j^V over all correct match image pairs in the training set. x_t^V is a patch that is connected to x_i^U in the binary mapping structure $\mathbf{M}_\alpha$. From Eq. 9, $\hat{P}(x_j^V | x_i^U, \mathbf{M}_\alpha)$ will equal to 1 if $\mathbf{M}_\alpha$ includes a link between x_i^U and x_j^V. Otherwise, $\hat{P}(x_j^V | x_i^U, \mathbf{M}_\alpha)$ will be decided by the relative appearance similarity strength between patch pair $\{x_i^U, x_j^V\}$ and all patch pairs which are connected to x_i^U in the binary mapping structure $\mathbf{M}_\alpha$.

Furthermore, $\hat{P}(x_i^U | \mathbf{M}_\alpha)$ in Eq. 8 is the updated importance probability of x_i^U after including $\mathbf{M}_\alpha$. It can be calculated by integrating the importance probability of each individual link in $\mathbf{M}_\alpha$:

$$\hat{P}(x_i^U | \mathbf{M}_\alpha) = \sum_{m_{\{x_s^U, x_t^V\}} \in \mathbf{M}_\alpha} \hat{P}(x_i^U | m_{\{x_s^U, x_t^V\}}, \mathbf{M}_\alpha) \cdot \hat{P}(m_{\{x_s^U, x_t^V\}} | \mathbf{M}_\alpha), \quad (10)$$

where $m_{\{x_s^U, x_t^V\}}$ is a patch-wise link in $\mathbf{M}_\alpha$, as the red lines in Fig. 4a. $\hat{P}(m_{\{x_s^U, x_t^V\}} | \mathbf{M}_\alpha)$ is the importance probability of link $m_{\{x_s^U, x_t^V\}}$ which is defined similar to $P(\mathbf{M}_\alpha)$:

$$\hat{P}(m_{\{x_s^U, x_t^V\}} | \mathbf{M}_\alpha) = \frac{\tilde{\mathcal{R}}_n(m_{\{x_s^U, x_t^V\}})}{\sum_{m_{\{x_h^U, x_g^V\}} \in \mathbf{M}_\alpha} \tilde{\mathcal{R}}_n(m_{\{x_h^U, x_g^V\}})}, \quad (11)$$

where $\tilde{\mathcal{R}}_n(m_{\{x_s^U, x_t^V\}})$ is the rank-n CMC score [21] when only using a single link $m_{\{x_s^U, x_t^V\}}$ as the correspondence structure to perform Re-ID.

$\hat{P}(x_i^U | m_{\{x_s^U, x_t^V\}}, \mathbf{M}_\alpha)$ in Eq. 10 is the impact probability from link $m_{\{x_s^U, x_t^V\}}$ to patch x_i^U, defined as:

$$\hat{P}(x_i^U | m_{\{x_s^U, x_t^V\}}, \mathbf{M}_\alpha) \propto \begin{cases} 0, & \text{if } d(x_i^U, x_s^U) \geq T_d \\ \dfrac{1}{d(x_i^U, x_s^U) + 1}, & \text{otherwise} \end{cases}$$

where x_s^U is link $m_{\{x_s^U, x_t^V\}}$'s end patch in camera A. $d(\cdot)$ and T_d are the same as Eq. 6.

Correspondence structure update. With the updated matching probability $\hat{P}^k(x_i^U, x_j^V)$ in Eq. 7, the matching probabilities in the k-th iteration can be finally updated by:

$$P^k(x_i^U, x_j^V) = (1 - \varepsilon) P^{k-1}(x_i^U, x_j^V) + \varepsilon \hat{P}^k(x_i^U, x_j^V), \quad (12)$$

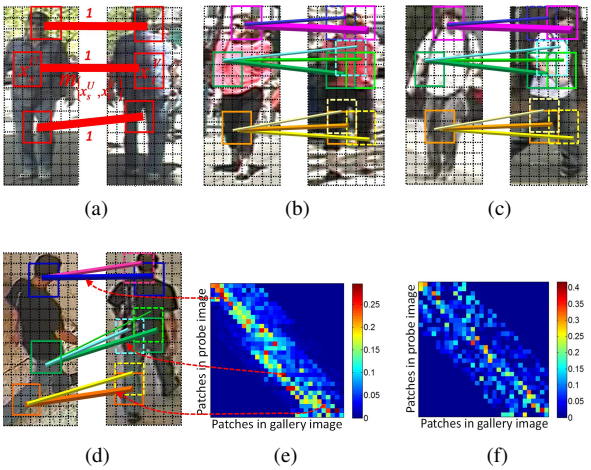


Figure 4. (a): An example of binary mapping structure (the red lines with weight 1 indicate that the corresponding patches are connected). (b)-(d): Examples of the correspondence structures learned by our approach where (b)-(c) and (d) are the correspondence structures for the VIPeR [6] and 3DPeS [1] datasets, respectively. The line widths in (b)-(d) are proportional to the patch-wise probability values. (e): The complete correspondence structure matrix of (d) learned by our approach. (f): The correspondence structure matrix of (d)'s dataset obtained by the simple-average method. (Patches in (e) and (f) are organized by a zig-zag scanning order. Matrices in (e) and (f) are down-sampled for a clearer illustration of the correspondence pattern). (Best viewed)

where $P^{k-1}(x_i^U, x_j^V)$ is the matching probability in iteration $k-1$. ε is the update rate which is set 0.2 in our paper.

From Equations 7–12, our update process integrates multiple variables (i.e., binary mapping structure, individual links, patch-link correlation) into a unified probability framework. In this way, various information cues such as appearances, ranking results, and patch-wise correspondence patterns can be effectively included during the model updating process. Besides, although the exact convergence of our learning process is difficult to analyze due to the inclusion of rank score calculation, our experiments show that most correspondence structures become stable within 300 iterations, which implies the reliability of our approach.

Figures 1 and 4 show some examples of the correspondence structures learned from different cross-view datasets. From Figures 1 and 4, we can see that the correspondence structures learned by our approach can suitably indicate the matching correspondence between spatial misaligned patches. For example, in Figures 1 and 4d-4e, the large lower-to-upper misalignments between cameras are effectively captured. Besides, the matching probability values in the correspondence structure also suitably reflects the correlation strength between different patch locations, as displayed by the colored points in Figures 1c and 4e.

Furthermore, comparing Figures 1a and 1b, we can see that the human-pose variation is also suitably handled by the learned correspondence structure. More specifically, although images in Fig 1 have different human poses, patches of camera A in both figures can correctly find their corresponding patches in camera B since the one-to-many matching probability graphs in the correspondence structure suitably embed the local correspondence variation between cameras. Similar observations can also be obtained from Figures 4b and 4c. It should be noted that images in the dataset of Figures 4b and 4c are taken by unfixed cameras (i.e., cameras with unfixed locations). However, the correspondence structure learned by our approach can still effectively encode the camera-person configuration and capture the cross-view correspondence pattern accordingly.

6. Experimental Results

We perform experiments on the following four datasets:

VIPeR. The VIPeR dataset [6] is a commonly used dataset which contains 632 image pairs for 632 pedestrians, as in Figures 4a-4c and 5d. It is one of the most challenging datasets which includes large differences in viewpoint, pose, and illumination between two camera views. Images from camera A are mainly captured from 0 to 90 degree while camera B mainly from 90 to 180 degree.

PRID 450S. The PRID 450S dataset [19] consists of 450 person image pairs from two non-overlapping camera views. It is also challenging due to low image qualities and viewpoint changes.

3DPeS. The 3DPeS dataset [1] is comprised of 1012 images from 193 pedestrians captured by eight cameras, where each person has 2 to 26 images, as in Figures 4d and 5a. Note that since there are eight cameras with significantly different views in the dataset, in our experiments, we group cameras with similar views together and form three camera groups. Then, we train a correspondence structure between each pair of camera groups. Finally, three correspondence structures are achieved and utilized to perform Re-ID between different camera groups. For images from the same camera group, we simply utilize adjacency-constrained search [27] to find patch-wise mapping and calculate the image matching score accordingly.

Road. The road dataset is our own constructed dataset which includes 416 image pairs taken by two cameras with camera A monitoring an exit region and camera B monitoring a road region, as in Figures 1 and 5g.[2] Since images in this dataset are taken from a realistic crowd road scene, the interferences from severe occlusion and large pose variation significantly increase the difficulty of this dataset.

For all of the above datasets, we follow previous methods [7, 22, 25] and perform experiments under 50%-training and 50%-testing. All images are scaled to 128×48. The

[2]This dataset will be open to the public soon.

patch size in our approach is 24×18. The stride size between neighboring patches is 6 horizontally and 8 vertically for probe images, and 3 horizontally and 4 vertically for gallery images. Note that we use smaller stride size in gallery images in order to obtain more patches. In this way, we can have more flexibilities during patch-wise matching.

6.1. Results for patch matching

We compare the patch matching results of three methods: (1) The adjacency-constrained search method [27, 26] which finds a best matched patch for each patch in a probe image (probe patch) by searching a fixed neighborhood region around the probe patch's co-located patch in a gallery image (*Adjacency-constrained*). (2) The simple-average method which simply averages the binary mapping structures for different probe images (as in Fig. 4a) to be the correspondence structure and combines it with a global constraint to find the best one-to-one patch matching result (*Simple-average*). (3) Our approach which employs a boosting-based process to learn the correspondence structure and combines it with a global constraint to find the best one-to-one patch matching result.

Fig. 5 shows the patch mapping results of different methods, where solid lines represent matching probabilities in a correspondence structure and red-dashed lines represent patch matching results. Besides, Figures 4e and 4f show one example of the correspondence structure matrix obtained by our approach and the simple-average method, respectively. From Figures 5 and 4e-4f, we can observe:

(1) Since the adjacency-constrained method searches a fixed neighborhood region without considering the correspondence pattern between cameras, it may easily be interfered by wrong patches with similar appearances in the neighborhood (cf. Figures. 5d, 5g). Comparatively, with the indicative matching probability information in the correspondence structure, the interference from mismatched patches can be effectively reduced (cf. Figures. 5f, 5i).

(2) When there are large misalignments between cameras, the adjacency-constrained method may fail to find proper patches as the correct patches may be located outside the neighborhood region, as in Fig. 5a. Comparatively, the large misalignment pattern between cameras can be properly captured by our correspondence structure, resulting in a more accurate patch matching result (cf. Fig. 5c).

(3) Comparing Figures 4e, 4f with the last two columns in Fig. 5, it is obvious that the correspondence structures by our approach is better than the simple average method. Specifically, the correspondence structures by the simple average method include many unsuitable matching probabilities which may easily result in wrong patch matches. In contrast, the correspondence structures by our approach are more coherent with the actual spatial correspondence pattern between cameras. This implies that reliable correspondence structure cannot be easily achieved without suitably integrating the information cues between cameras.

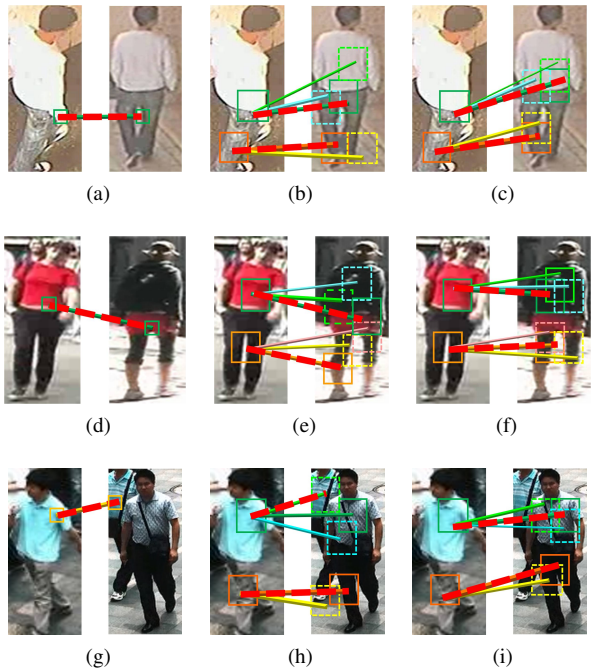


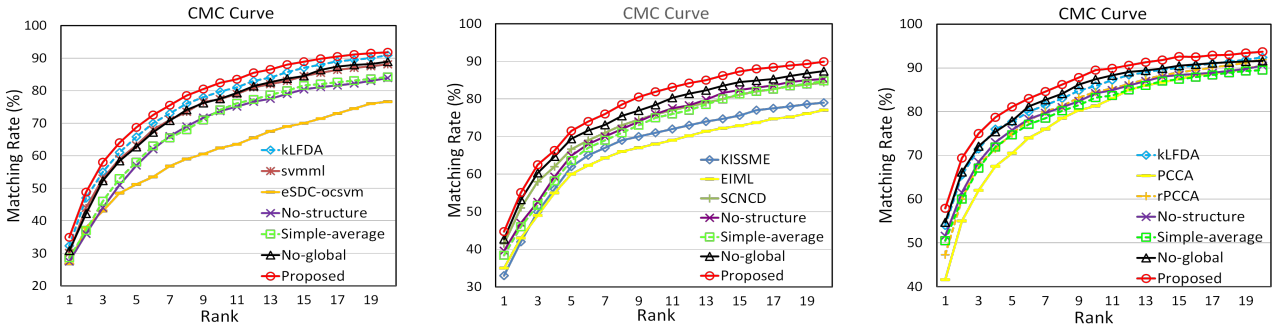
Figure 5. Comparison of different patch mapping methods. Left column: the adjacency-constrained method; Middle column: the simple-average method; Last column: our approach. The solid lines represent matching probabilities in a correspondence structure and the red-dashed lines represent patch matching results. Note that the image pair in (a)-(c) includes the same person (i.e., correct match) while the image pairs in (d)-(i) include different people (i.e., wrong match). (Best viewed in color)

6.2. Results for person re-identification

We evaluate person re-identification results by the standard Cumulated Matching Characteristic (CMC) curve [21] which measures the correct match rates within different Re-ID rank ranges. The evaluation protocols are the same as [7]. That is, for each dataset, we perform 10 randomly-partitioned 50%-training and 50%-testing experiments and average the results.

We compare results of four methods: (1) Not applying correspondence structure and directly using the appearance similarity between co-located patches for person Re-ID (*No-structure*); (2) Simply averaging the binary mapping structures for different probe images as the correspondence structure and utilizing it for Re-ID (*Simple-average*); (3) Using the correspondence structure learned by our approach, but do not include global constraint when performing Re-ID (*No-global*); (4) Our approach (*Proposed*).

We also compare our results with state-of-the-art methods on different datasets: kLFDA [22], eSDC-ocsvm [27], KISSME [19], Salience [26], svmml [13], RankBoost [11]



(a) the VIPeR dataset (b) the PRID 450S dataset (c) the 3DPeS dataset

Figure 6. CMC curves for different methods.

Table 1. CMC results on the VIPeR dataset

Rank	1	5	10	20	30	50
kLFDA[22]	32.3	65.8	79.7	90.9	-	-
KISSME[19]	27	-	70	83	-	95
Salience[26]	30.2	52.3	-	-	-	-
svmml[13]	30.1	63.2	77.4	88.1	-	-
RankBoost[11]	23.9	45.6	56.2	68.7	-	-
eSDC-ocsvm[27]	26.7	50.7	62.4	76.4	-	-
LF[18]	24.2	-	67.1	-	-	94.1
No-structure	27.5	57.0	73.7	83.9	87.7	94.3
Simple-average	28.5	57.9	74.1	84.2	88.3	94.6
No-global	30.8	62.7	77.5	88.9	91.7	95.6
Proposed	**34.8**	**68.7**	**82.3**	**91.8**	**94.9**	**96.2**

Table 2. CMC results on the PRID 450S dataset

Rank	1	5	10	20	30	50
KISSME[19]	33	-	71	79	-	90
EIML[8]	35	-	68	77	-	90
SCNCD[25]	41.5	66.6	75.9	84.4	88.4	92.4
SCNCDFinal[25]	41.6	68.9	79.4	87.8	91.8	95.4
No-structure	39.6	64.9	76.0	85.3	89.3	93.3
Simple-average	38.2	63.6	75.1	84.9	88.9	92.4
No-global	42.7	69.3	78.2	87.4	91.1	95.1
Proposed	**44.4**	**71.6**	**82.2**	**89.8**	**93.3**	**96.0**

Table 3. CMC results on the 3DPeS dataset

Rank	1	5	10	15	20	30
kLFDA[22]	54.0	77.7	85.9	-	92.4	-
rPCCA[22]	47.3	75.0	84.5	-	91.9	-
PCCA[16]	41.6	70.5	81.3	-	90.4	-
No-structure	51.6	75.8	84.2	88.4	90.5	92.6
Simple-average	50.5	74.7	83.2	87.4	89.5	92.6
No-global	54.7	77.9	87.4	90.5	91.6	93.7
Proposed	**57.9**	**81.1**	**89.5**	**92.6**	**93.7**	**94.7**

Table 4. CMC results on the Road dataset

Rank	1	5	10	15	20	30
eSDC-knn[27]	52.4	74.5	83.7	88.0	89.9	91.8
No-structure	50.5	80.3	87.0	91.3	94.2	95.7
Simple-average	49.0	81.7	90.4	92.8	95.7	96.2
No-global	58.2	85.6	94.2	97.1	98.1	98.6
Proposed	**61.5**	**91.8**	**95.2**	**98.1**	**98.6**	**99.0**

and LF [18] on the VIPeR dataset; KISSME [19], EIML [8], SCNCD [25], SCNCDFinal [25] on the PRID 450S dataset; kLFDA [22], rPCCA [22], PCCA [16] on the 3DPeS dataset; and eSDC-knn [27] on the Road dataset.

Tables 1–4 and Fig. 6 show the CMC results of different methods. From the CMC results, we can see that: (1) Our approach has better Re-ID performances than the state-of-the-art methods. This demonstrates the effectiveness of our approach. (2) Our approach has obviously improved results than the no-structure method. This indicates that proper correspondence structures can effectively improve Re-ID performances by reducing patch-wise misalignments. (3) The simple-average method has similar performance to the no-structure method. This implies that unsuitably selected correspondence structures cannot improve Re-ID performance. (4) The no-global method also has good Re-ID performance. This further demonstrates the effectiveness of the correspondence structure learned by our approach. Meanwhile, our approach also has superior performance than the no-global method. This demonstrates the usefulness of introducing global constraint in the patch matching process.

7. Conclusion

In this paper, we propose a novel framework for addressing the problem of cross-view spatial misalignments in person Re-ID. Our framework consists of two key ingredients: 1) introducing the idea of correspondence structure and learning this structure via a novel boosting method to adapt to arbitrary camera configurations; 2) a constrained global matching step to control the patch-wise misalignments between images due to local appearance ambiguity. Extensive experimental results on benchmark show that our approach achieves the state-of-the-art performance.

Under this framework, our future work is devoted to explore new variants of the two components, such as: 1) designing other correspondence structure learning methods that allow for multiple structure candidates to enhance its flexibility; 2) devising and incorporating edge-to-edge similarity metrics for solving the constrained global matching problem as graph matching [4, 24], which has been proven more effective in many computer vision applications.

References

[1] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *ACM workshop Human gesture and behavior understanding*, 2011. 6

[2] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*, 2010. 3

[3] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 2

[4] M. Cho, J. Lee, and K. M. Lee. Reweighted random walks for graph matching. In *ECCV*, 2010. 8

[5] 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

[6] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007. 1, 6

[7] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 1, 2, 6, 7

[8] M. Hirzer, P. M. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *AVSS*, 2012. 1, 2, 8

[9] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 1, 2, 3, 5

[10] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955. 3, 4

[11] C.-H. Kuo, S. Khamis, and V. Shet. Person re-identification using semantic color names and rankboost. In *WACV*, 2013. 1, 2, 7, 8

[12] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Trans. PAMI*, 2013. 1, 2

[13] 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, 8

[14] C. Liu, S. Gong, and C. C. Loy. On-the-fly feature importance mining for person re-identification. *Pattern Recognition*, 2014. 2

[15] L. Ma, X. Yang, Y. Xu, and J. Zhu. A generalized emd with body prior for pedestrian identification. *Journal of Visual Communication and Image Representation*, 2013. 1, 2, 4

[16] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 8

[17] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *CVPR*, 2010. 1, 2

[18] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013. 8

[19] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*. Springer, 2014. 1, 2, 6, 7, 8

[20] H. Wang, S. Gong, and T. Xiang. Unsupervised learning of generative topic saliency for person re-identification. In *BMVC*, 2014. 2

[21] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 5, 7

[22] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 1, 2, 6, 7, 8

[23] Y. Xu, L. Lin, W.-S. Zheng, and X. Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, 2013. 2

[24] J. Yan, C. Zhang, H. Zha, X. Yang, W. Liu, and S. M. Chu. Discrete hyper-graph matching. In *CVPR*, 2015. 8

[25] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *ECCV*, 2014. 1, 2, 6, 8

[26] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013. 2, 3, 7, 8

[27] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 1, 2, 3, 4, 5, 6, 7, 8