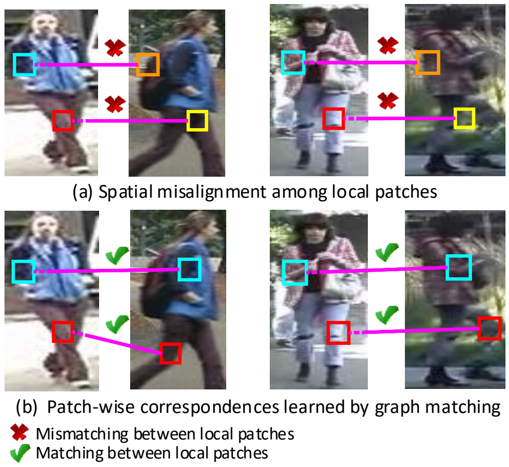
Graph Correspondence Transfer for Person Re-identification

Qin Zhou[1,3] **Heng Fan**[3] **Shibao Zheng**[1] **Hang Su**[4] **Xinzhe Li**[1] **Shuang Wu**[5] **Haibin Ling**[2,3,*]
[1]Institute of Image Processing and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[3]Department of Computer & Information Sciences, Temple University, Philadelphia 19122, USA
[4]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[5]YouTu Lad, Tencent, Shanghai 200233, China

Abstract

In this paper, we propose a graph correspondence transfer (GCT) approach for person re-identification. Unlike existing methods, the GCT model formulates person re-identification as an *off-line* graph matching and *on-line* correspondence transferring problem. In specific, during training, the GCT model aims to learn *off-line* a set of correspondence templates from positive training pairs with various pose-pair configurations via patch-wise graph matching. During testing, for each pair of test samples, we select a few training pairs with the most similar pose-pair configurations as references, and transfer the correspondences of these references to test pair for feature distance calculation. The matching score is derived by aggregating distances from different references. For each probe image, the gallery image with the highest matching score is the re-identifying result. Compared to existing algorithms, our GCT can handle spatial misalignment caused by large variations in view angles and human poses owing to the benefits of patch-wise graph matching. Extensive experiments on five benchmarks including VIPeR, Road, PRID450S, 3DPES and CUHK01 evidence the superior performance of GCT model over other state-of-the-art methods.


(a) Spatial misalignment among local patches

(b) Patch-wise correspondences learned by graph matching

✗ Mismatching between local patches
✓ Matching between local patches

Figure 1: Illustration of misalignment problem in Re-ID. Image (a) shows misalignment among local patches caused by viewpoint changes. The proposed GCT model can capture the correct semantic matching among patches using patch-wise graph matching, as shown in image (b).

1 Introduction

Person re-identification (Re-ID), which aims to associate a probe image to each individual in a gallery set (usually across different non-overlapping camera views), plays a crucial role in various applications including video surveillance, human retrieval, etc. Despite great successes in recent years, accurate Re-ID remains challenging due to many factors such as large appearance changes in different camera views and heavy body occlusions. To deal with these issues, numerous Re-ID approaches are proposed (Farenzena et al. 2010; Karanam, Li, and Radke 2015; Zhao, Ouyang, and Wang 2013; Chen et al. 2016; Köstinger et al. 2012; Li et al. 2013; Liao et al. 2015; Mignon and Jurie 2012; Paisitkriangkrai, Shen, and van den Hengel 2015; Pedagadi et al. 2013; Martinel et al. 2016).

For Re-ID task, one major challenge is to deal with the inevitable spatial misalignments between image pairs caused by large variations in camera views and human poses, as shown in Fig. 1. Most existing methods (Yang et al. 2014; Xiong et al. 2014; Chen et al. 2015), nevertheless, focus on addressing the problem of Re-ID by comparing the holistic differences between images, which ignores spatial misalignment problem. To alleviate this issue, there are some attempts to apply part-based approaches to handle misalignment (Shen et al. 2015; Oreifej, Mehran, and Shah 2010; Zhao, Ouyang, and Wang 2013; Yang et al. 2017). These methods divide objects into local patches and perform an *online* patch-level matching for Re-ID. Though these approaches can handle spatial misalignment to some extent, being in lack of spatial and visual context information among local patches, they still fail in presence of visually similar body appearances or occlusions.

In this paper, we propose to learn patch-level matching templates for positive training pairs via graph matching, and transfer the learned patch-level correspondences to test pairs with similar pose-pair configurations. In the formulation of graph matching, both spatial and visual context information are utilized to solve misalignment problem. The cor-

[*] Corresponding author.

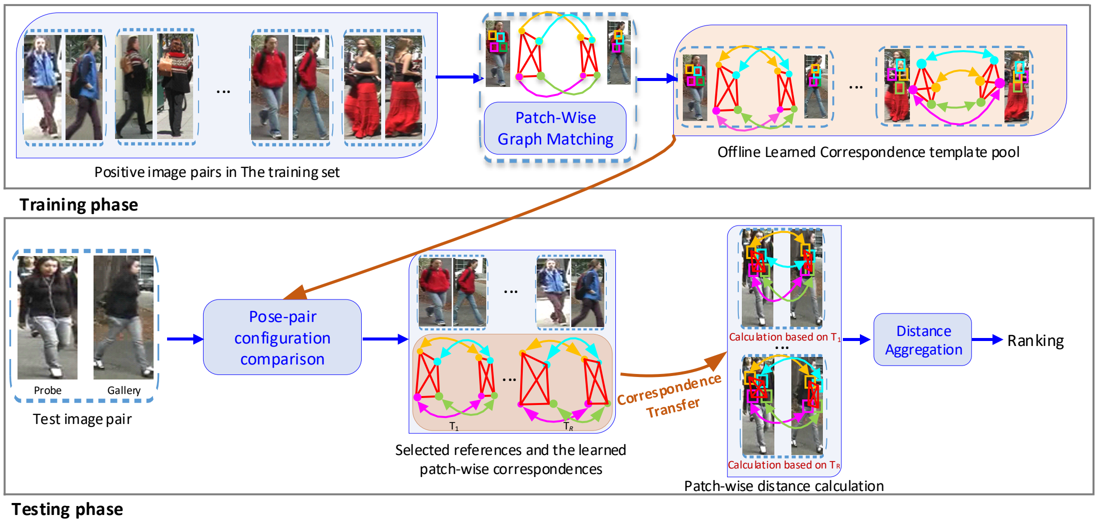

Positive image pairs in The training set
Patch-Wise Graph Matching
Offline Learned Correspondence template pool
Training phase
Probe
Gallery
Test image pair
Pose-pair configuration comparison
T_1
T_R
Selected references and the learned patch-wise correspondences
Correspondence Transfer
Calculation based on T_1
Calculation based on T_R
Patch-wise distance calculation
Distance Aggregation
Ranking
Testing phase

Figure 2: Illustration of GCT model. During training, spatial and visual context information are embedded into patch-wise graph matching to establish patch-level correspondences for different positive training pairs with various pose-pair configurations. During testing, for a pair of test samples, we use a simple pose-pair configuration comparison method to 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 to compute local feature distances, which are then aggregated to calculate the overall feature distance.

respondence transfer in testing pahse is based on the observation that two image pairs with similar pose-pair configurations tend to share similar patch-level correspondences. Benefiting from part-based strategy and implicitly modeling body context information into graph matching procedure, our GCT algorithm is able to well deal with spatial misalignment. Besides, the regularization term imposed on the learned correspondences in patch-wise graph matching helps to improve the robustness of GCT model when occlusions occur. During testing, for each pair of samples, we present a simple but effective pose-pair configuration comparison method to select the training pairs with the most similar pose-pair configurations as references, and the learned patch-wise correspondences of these references are utilized to compute the overall feature distance between this test pair. Fig 2 illustrates the details of GCT model. Experiments on five benchmarks show the effectiveness of GCT method.

In summary, we make the following contributions:

- Spatial and visual context information are utilized to exploit the patch-wise semantic correspondences between positive image pairs to handle spatial misalignment. And for the first time, a novel graph correspondence transfer (GCT) model is presented for person re-identification.
- Based on the observation that image pairs with similar pose-pair configurations tend to share similar patch-level correspondences, we introduce a simple but effective pose-pair configuration comparison method to find the best references for each test image pair.
- Extensive experiments on five challenging benchmarks demonstrate that the proposed GCT model performs favorably against state-of-the-art approaches, and in fact even better than many deep learning based solutions.

2 Related Work

Being extensively studied, numerous approaches have been proposed for Re-ID in recent years (Zheng et al. 2015). Here we briefly review some related works of this paper.

To handle the problem of spatial misalignment, there are some attempts to apply part-based strategies for Re-ID. This kind of methods divide images into several parts and perform on-line patch-level matching to discard misalignments. In (Liao et al. 2015), Liao *et al.* propose to incorporate body prior into Re-ID by decomposing human body into a few fixed stripes, however it still fails in the scenarios of large view changes. Cheng *et al.* (Cheng et al. 2016) take the advantages of body part detection, and propose a part-based appearance model to alleviate the influence of misalignment in Re-ID. However, this approach heavily relies on body part detection performance, leading to degradation in presence of occlusion. In (Zhao, Ouyang, and Wang 2013), Zhao *et al.* propose to mine the salient scores of different local patches for Re-ID by building dense correspondences between image pairs. The aforementioned algorithms, however, neglect body context information during feature designing or metric learning, resulting in performance deteriorating.

The most relevant work to ours is (Shen et al. 2015) (CSL), which aims to mine the patch-wise matching structure for each pair of cameras. Nevertheless, our GCT differs from CSL in two aspects: (i) Instead of learning a holistic optimal correspondence structure for each camera pair in CSL, we utilize graph matching to establish accurate local feature correspondences for each positive image pair, and then transfer the learned correspondence templates for Re-ID. (ii) We implicitly model the body context information, which is neglected in CSL, in the affinity matrix for graph matching to improve the Re-ID performance in our work.

3 The Proposed Approach

In this section, we describe the detailed GCT model, which consists of correspondence learning with patch-wise graph matching in training phase, reference selection via pose-pair configuration comparison and patch-wise feature distance calculation and aggregation based on correspondence transfer.

3.1 Patch-wise correspondence learning with graph matching

To deal with the misalignment problem, a part-based strategy is adopted to represent human appearance. In specific, we decompose the images into many overlapping patches, and represent each image with an undirected attribute graph $G = (V, E, A^V)$, where each vertex v_i in the vertex set $V = \{v_i\}_{i=1}^{n}$ denotes an image patch, and each edge encodes the context information of the connected vertex pair. $A^V = \{A^{V_P}, A^{V_F}\}$ are vertex attributes representing spatial and visual features of the local patches.

During training, given a positive image pair I_1 and I_2 with identify labels l_1 and l_2, where $l_1 = l_2$ (i.e., I_1 and I_2 belong to same person), they can be represented with attribute graphs $G_1 = (V_1, E_1, A_1^V)$ and $G_2 = (V_2, E_2, A_2^V)$, 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 vertexes and V_2 with n_2 vertexes, such that the intra-person similarity (i.e., $l_1 = l_2$) is maximized on the training set.

In Re-ID, $X_{ia} = 1$ means the i^{th} patch in I_1 semantically corresponds to the a^{th} patch in I_2. Mathematically, the patch-wise correspondence learning is formulated as an Integer Quadratic Programming (IQP) problem as follows:

$$\arg\max_{\mathbf{x}} \ \mathbf{x}^{\mathrm{T}} K \mathbf{x},$$
$$s.t. \begin{cases} X_{ia} \in \{0,1\}, \forall i \in \{1, \cdots, n_1\}, \forall a \in \{1, \cdots, n_2\} \\ \sum_i X_{ia} \leq 1, \forall a \in \{1, \cdots, n_2\}, \\ \sum_a X_{ia} \leq 1, \forall i \in \{1, \cdots, n_1\}, \end{cases} \tag{1}$$

where $\mathbf{x} = \mathrm{vec}(X) \in \{0,1\}^{n_1 \times n_2}$ denotes the vectorized version of matrix X and $K \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$ represents the corresponding affinity matrix between G_1 and G_2, which encodes the relational similarities between edges and vertices.

Affinity matrix design Due to large variations in human body configuration caused by serious 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 $K^{ia,ia}$ of the affinity matrix K (which capture the node compatibility between vertex $v_i \in V_1$ and vertex $v_a \in V_2$) are calculated as follows:

$$K^{ia,ia} = S_{ia}^{P} \cdot S_{ia}^{F}, \tag{2}$$

where S_{ia}^{P} and S_{ia}^{F} refer to the *spatial proximity* and *visual similarity* between v_i and v_a respectively. The S_{ia}^{P} and S_{ia}^{F} can be mathematically computed with

$$\begin{aligned} S_{ia}^{P} &= \exp(-\|A_i^{V_P} - A_a^{V_P}\|_2), \\ S_{ia}^{F} &= \exp(-\|A_i^{V_F} - A_a^{V_F}\|_2), \end{aligned} \tag{3}$$

where $A_i^{V_P}$ and $A_a^{V_P}$ denote spatial positions of v_i and v_a, and $A_i^{V_F}$ and $A_a^{V_F}$ represent their visual features.

Likewise, for non-diagonal element $K^{ia,jb}$ in K, which encodes the compatibility between two edges e_{ij} connecting $\{v_i \in V_1, v_j \in V_1\}$ and e_{ab} connecting $\{v_a \in V_2, v_b \in V_2\}$, it can be obtained as the following

$$K^{ia,jb} = S_{ij,ab}^{P} \cdot S_{ij,ab}^{F}, \tag{4}$$

where $S_{ij,ab}^{P}$ and $S_{ij,ab}^{F}$ represent spatial and visual compatibilities between edges $e_{ij} \in E_1$ and $e_{ab} \in E_2$, and they are calculated by

$$\begin{aligned} S_{ij,ab}^{P} &= \exp(-\|(A_i^{V_P} - A_j^{V_P}) - (A_a^{V_P} - A_b^{V_P})\|_2), \\ S_{ij,ab}^{F} &= \exp(-\|(A_i^{V_F} - A_j^{V_F}) - (A_a^{V_F} - A_b^{V_F})\|_2). \end{aligned} \tag{5}$$

In this way, the calculated affinity matrix K implicitly embeds the spatial and visual context information into the graph matching procedure, such that the matched vertices and edges have larger similarities and are more compatible with each other both spatially and visually. Therefore, we can obtain a satisfying patch-wise matching result for Re-ID.

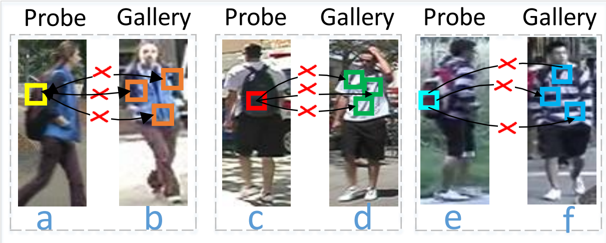


Figure 3: Sample images to demonstrate the fact that local patches visible in one view may not appear in the other.

Outlier removal Due to spatial misalignment, some patches visible in one view may not appear in the other (see Figure 3). Therefore, imposing a global one-to-one match between patches could bring in noise and deteriorate the performance. In this case, only establishing correspondences between commonly visible parts of positive image pairs is more reasonable. To this end, we adopt the same strategy as in (Suh, Adamczewski, and Lee 2015) by incorporating a regularization term on the number of activated vertices. In this way, the probe patches that match with high spatial and visual similarities are activated, while the outliers that do not co-exist in the two images are excluded. Thus, the objective function (1) can be rewritten as follows:

$$\arg\max_{\mathbf{x}} \ \mathbf{x}^{\mathrm{T}} K \mathbf{x} - \lambda ||\mathbf{x}||_2^2,$$
$$s.t. \begin{cases} X_{ia} \in \{0,1\}, \forall i \in \{1, \cdots, n_1\}, \forall a \in \{1, \cdots, n_2\} \\ \sum_i X_{ia} \leq 1, \forall a \in \{1, \cdots, n_2\}, \\ \sum_a X_{ia} \leq 1, \forall i \in \{1, \cdots, n_1\}, \end{cases} \tag{6}$$

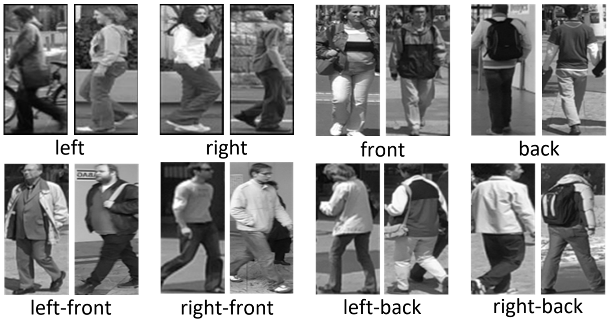


Figure 4: Sample images in eight classes in the TUD dataset (Andriluka, Roth, and Schiele 2010). Note that the TUD dataset is only used for training the reference selection model, and is different from the benchmarks in experiments.

where λ is a trade-off parameter to control the difficulty of a new probe vertex being activated. Larger λ means more extra similarity should be brought to activate a new vertex. We adopt the method in (Suh, Adamczewski, and Lee 2015) to solve Eq.(6).

In existing part-based Re-ID methods (Shen et al. 2015; Zhao, Ouyang, and Wang 2013), an image is typically decomposed into hundreds of patches to capture detailed local visual information, leading to intractability in solving Eq. (6). To reduce the search space and inhibit potential matching ambiguity, we adopt the commonly utilized spatial constraints (Liao et al. 2015; Chen et al. 2016) to lower the computational load, as well as to improve the patch-wise matching results. More specifically, a probe image is divided into a few horizontal stripes, and for each stripe in the probe image, patch-wise correspondences are established between the corresponding gallery stripe within the search range in the gallery image by optimizing Eq.(6).

By optimizing Eq.(6), we can obtain a set of graph correspondence templates for the positive image pairs in the training set.

3.2 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 are directly utilized to compute the distance between probe and gallery images. To this end, we need to find out the best references for each pair of test images from the training set. Since pose configurations are closely related to body orientations, we calculate the similarities between different pose pairs by comparing their related body orientations.

We propose to utilize a simple but effective random forest method (Liaw and Wiener 2002) 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 Figure 4. 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 (Liaw and Wiener 2002). Once the random forest $M = \{tree_1, tree_2, \cdots, tree_T\}$ is built, where T denotes the number of trees in M, the body orientation proximity O between two images I_i and I_j can be calculated as:

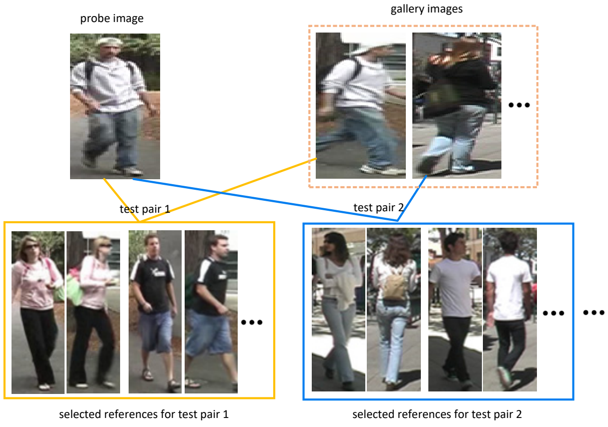

probe image
gallery images
...
test pair 1
test pair 2
...
... ...
selected references for test pair 1
selected references for test pair 2

Figure 5: Demonstration of reference selection results.

$$\mathrm{O}(I_i, I_j) = \frac{1}{T}\sum_{t=1}^{T} y_{ij}^t, \tag{7}$$

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

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 follows

$$S(P, P') = \mathrm{O}(I_p, I_p^{'}) \cdot \mathrm{O}(I_g, I_g^{'}) \tag{8}$$

With Eq.(8), we can calculate the body configuration similarities between an test image pair and all the positive training image pairs, and select R training image pairs with highest similarities as the best references for the test pair. Figure 5 shows some selected references of the sample test pairs.

3.3 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 3.2) 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.

After obtaining the best references for each test image pair, we transfer their learned correspondences to compute the distance between the test images. 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, 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 template set of these R references, where each template $T_i = \{c_{ij}\}_{j=1}^{Q_i}$ contains Q_i patch-wise correspondences, and each correspondence $c_{ij} = (w_{ij}^{p}, w_{ij}^{g})$ denotes the positions of matched local patches in the probe and gallery image.

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

$$D(\bar{I}_p, \bar{I}_g) = \sum_{i=1}^{R} \sum_{j=1}^{Q_i} \delta(f_p^{w_{ij}^{p}}, f_g^{w_{ij}^{g}}) \tag{9}$$

where $\delta(\cdot,\cdot)$ denotes the KISSME metric (Köstinger et al. 2012), and $f_p^{w_{ij}^{p}}$ and $f_g^{w_{ij}^{g}}$ represent features of local patches located at w_{ij}^{p} and w_{ij}^{g} in probe image $\bar{I}_p$ and gallery image $\bar{I}_g$. In this paper, we use Local Maximal Occurrence features (Liao et al. 2015) to represent each image patch.

With Eq.(9), we can calculate the patch-wise feature distances between each correspondence (a semantically matched patch pair between the probe and gallery images) of the selected references, these local feature distances are then equally aggregated to obtain the overall distance between the test image pairs. The gallery image with the smallest distance is determined to be the re-identifying result.

4 Experimental Results

4.1 Experimental setup

Datasets We conduct extensive 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 (Gray, Brennan, and Tao 2007) dataset consists 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 (Shen et al. 2015), 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 (Roth et al. 2014) dataset contains 450 pairs of images from two camera views. The very similar appearances in images make it very challenging for person re-identification.

3DPES dataset: The 3DPES dataset (Baltieri, Vezzani, and Cucchiara 2011) contains 1011 images of 192 persons captured from 8 different cameras. 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 (Li, Zhao, and Wang 2012) 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.

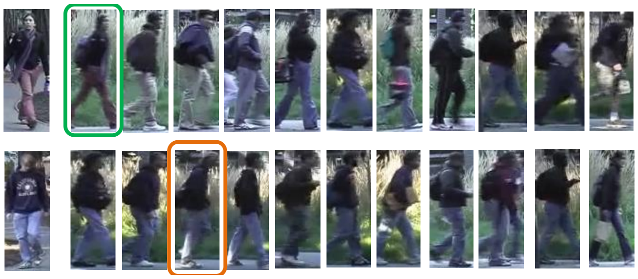


Figure 6: Top ranked images selected by our GCT algorithm. The first column are two probe images, and the following are the top ranked gallery images of each probe obtained by GCT. Images marked by green/orange bounding boxes in each row are the ground-truth matches of each probe.

Parameter setup The proposed algorithm is implemented in Matlab on an Intel(R) Core(TM) i7-5820K CPU of 3.30GHz. The λ in Eq.(6) is set to 2. The number of trees in the random forest model is 500. The best R for VIPeR, Road, PRID450S, 3DPES and CUHK01 datasets are 20, 5, 10, 20 and 20 respectively. The size of local patch is 32×24. All the parameters will be available in the source code to be released for accessible reproducible research.

Evaluation We adopt the common half-training and half-testing setting (Köstinger et al. 2012), 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 Figure 6. As shown in Figure 6, the proposed GCT 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 second row of Figure 6, the top ranked gallery images by GCT are all with dark coats and blue jeans, and to some extent, the rank1 image is visually more similar than the correct match marked by orange bounding box w.r.t. the probe image. Therefore, the proposed GCT algorithm is able to handle the spatial misalignment problem and generate satisfying ranking results for practical applications.

4.2 Overall performance

The GCT algorithm is implemented in Matlab on a PC with i7-5820K. The average training and testing time for an image pair are 0.096s and 0.0024s. To validate our approach, we evaluate the GCT model on five challenging datasets and compare it with several state-of-the-art approaches. The detailed comparison results are presented as follows.

On the VIPeR dataset, we compare the GCT with other fifteen algorithms, including SalMatch (Zhao, Ouyang, and

Table 1: Comparisons of top r matching rate using CMC (%) on VIPeR dataset. The best and second best results are marked in red and blue, respectively.

Methods	r=1	r=5	r=10	r=20
SalMatch	30.2	52.3	65.5	79.2
Semantic	41.6	71.9	86.2	**95.1**
LSSCDL	42.7	–	84.3	91.2
KISSME	27.3	55.3	69.0	82.7
SVMML	30.0	64.7	79.0	91.3
kLFDA	32.4	65.9	79.8	90.8
Polymap	36.8	70.4	83.7	91.7
LMF+LADF	43.4	73.0	84.9	93.7
LOMO+XQDA	40.0	68.1	80.5	91.1
DCSL	44.6	73.4	82.6	–
TMA	48.2	–	**87.7**	**95.5**
TCP	47.8	**74.7**	84.8	91.1
DGD	35.4	62.3	69.3	–
Spindle-Net	**53.8**	74.1	83.2	92.1
CSL	34.8	68.7	82.3	91.8
Our GCT	**49.4**	**77.6**	**87.2**	94.0

Table 2: Comparison of top r matching rate using CMC (%) on Road dataset. The best and second best results are marked in red and blue, respectively.

Methods	r=1	r=5	r=10	r=20
eSDC-knn	52.4	74.5	83.7	89.9
CSL	**61.5**	**91.8**	**95.2**	**98.6**
Our GCT	**88.8**	**96.7**	**98.4**	**99.6**

Wang 2013), Semantic (Shi, Hospedales, and Xiang 2015), LSSCDL (Zhang et al. 2016b), KISSME (Köstinger et al. 2012), SVMML (Li et al. 2013), kLFDA (Xiong et al. 2014), Polymap (Chen et al. 2015), LMF+LADF (Zhao, Ouyang, and Wang 2014), LOMO+XQDA (Liao et al. 2015), DCSL (Zhang et al. 2016a), TMA (Martinel et al. 2016), TCP (Cheng et al. 2016), DGD (Xiao et al. 2016), Spindle Net (Zhao et al. 2017) and CSL (Shen et al. 2015). The comparison results are presented in Table 1. As illustrated in Table 1, the proposed GCT algorithm achieves the best recognition rate at rank 5, and competitive performances at rank 1, 10 and 20.

The Road dataset is proposed in CSL (Shen et al. 2015). For comprehensive comparison, we also report the result on this dataset, and compare it with eSDC-knn (Zhao, Ouyang, and Wang 2013) and CSL (Shen et al. 2015). As shown in Table 2, compared to CSL (Shen et al. 2015), our algorithm obtains significant improvements of 27.3 percents on rank 1 recognition rate, 4.9 percents on rank 5, 3.2 percents on rank 10 and 1 percent on rank 20 recognition rate, respectively.

On PRID450S, we compare with six state-of-the-art algorithms, including KISSME (Köstinger et al. 2012), SCNCDFinal (Yang et al. 2014), Semantic (Shi, Hospedales, and Xiang 2015), TMA (Martinel et al. 2016), NSFT (Zhang, Xiang, and Gong 2016) and CSL (Shen et al. 2015). As shown in Table 3, our algorithm achieves the best results on recognition rates of small ranks, which is critical in practice.

Table 3: Comparison of top r matching rate using CMC (%) on PRID450S dataset. The best and second best results are marked in red and blue, respectively.

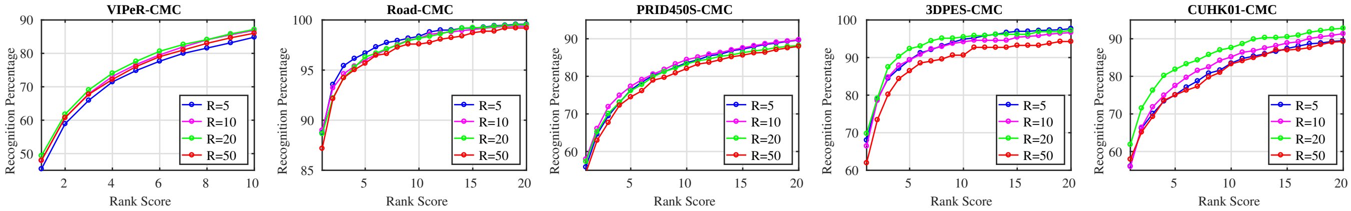
Methods	r=1	r=5	r=10	r=20
KISSME	33	–	71	79
SCNCDFinal	41.6	68.9	79.4	87.8
Semantic	44.9	71.7	77.5	86.7
TMA	**54.2**	**73.8**	**83.1**	**90.2**
NSFT	40.9	64.7	73.2	81.0
CSL	44.4	71.6	82.2	**89.8**
Our GCT	**58.4**	**77.6**	**84.3**	**89.8**

Table 4: Comparison of top r matching rate using CMC (%) on 3DPES dataset. The best and second best results are marked in red and blue, respectively.

Methods	r=1	r=5	r=10	r=20
LFDA	45.5	69.2	–	86.1
ME	53.3	76.8	–	92.8
kLFDA	54.0	77.7	85.9	92.4
PCCA	41.6	70.5	81.3	90.4
rPCCA	47.3	75.0	84.5	91.9
SCSP	57.3	79.0	–	91.5
WARCA	51.9	75.6	–	–
DGD	56.0	–	–	–
Spindle-Net	**62.1**	**83.4**	**90.5**	**95.7**
CSL	57.9	81.1	89.5	93.7
Our GCT	**69.8**	**92.4**	**95.5**	**97.2**

On 3DPES, we compare the GCT method with ten algorithms, including LFDA (Pedagadi et al. 2013), ME (Paisitkriangkrai, Shen, and van den Hengel 2015), kLFDA (Xiong et al. 2014), PCCA (Mignon and Jurie 2012), rPCCA (Xiong et al. 2014), SCSP (Chen et al. 2016), WARCA (Jose and Fleuret 2016), DGD (Xiao et al. 2016), Spindle Net (Zhao et al. 2017) and CSL (Shen et al. 2015). As shown in Table 4, we can see that the proposed algorithm outperforms existing state-of-the-art algorithms, and even deep learning based algorithm (Zhao et al. 2017). 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 patch-wise graph matching, our GCT model is robust to deal with these issues.

On CUHK01, we compare with nine state-of-the-art algorithms, including Semantic (Shi, Hospedales, and Xiang 2015), kLFDA (Xiong et al. 2014), IDLA(Ahmed, Jones, and Marks 2015), DeepRanking (Chen, Guo, and Lai 2016), ME (Paisitkriangkrai, Shen, and van den Hengel 2015), GOG (Matsukawa et al. 2016), SalMatch (Zhao, Ouyang, and Wang 2013), CSBT (Chen et al. 2017) and TCP (Cheng et al. 2016). The detailed comparison results are summarized in Table 5. As shown in Table 5, the proposed GCT method can achieve competitive results on this dataset. More specifically, we obtain the best rank 1 recognition rate (has a 8.2



(a) VIPeR (b) Road (c) PRID450S (d) 3DPES (e) CUHK01

Figure 7: Evaluation of different numbers of selected references R on the five datasets.

Table 5: Comparison of top r matching rate using CMC (%) on CUHK01. The best and second best results are marked in red and blue, respectively.

Methods	r=1	r=5	r=10	r=20
Semantic	32.7	51.2	–	76.3
kLFDA	32.8	59.0	69.6	–
IDLA	47.5	71.5	80.0	–
DeepRanking	50.4	75.9	84.1	–
ME	53.4	76.3	84.4	–
GOG	**57.8**	79.1	86.2	–
SalMatch	28.5	46.0	–	67.3
CSBT	51.2	76.3	–	91.8
TCP	53.7	**84.3**	**91.0**	**96.3**
Our GCT	**61.9**	**81.9**	**87.6**	**92.8**

percent gain over TCP (Cheng et al. 2016), a part based deep learning algorithm). And the recognition rates of other ranks are also competitive.

4.3 Analysis on the number of selected references

Different number of selected references (R) for calculating the distances between test pairs have different influences on the re-identification performance. With a small R, bad references may have a large impact on the patch-wise distance calculation, resulting in 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. We record the best R on the VIPeR, Road, PRID450S, 3DPES and CUHK01 are 20, 5, 10, 20 and 20, respectively. As shown in Figure 7, $R = 50$ performs the worst on all the other four datasets (except for VIPeR, on which $R = 50$ performs slightly better than $R = 5$, but is still inferior than $R = 10$ and $R = 20$). A relatively small optimal R on Road dataset can be attributed to the similar view of images in this dataset (small R can include enough meaningful patch-wise correspondences for each test pair). While a bigger optimal R on VIPeR, 3DPES and CUHK01 datasets indicates that the pose variations on these datasets are more serious, therefore more references should be incorporated together to include enough correct correspondences.

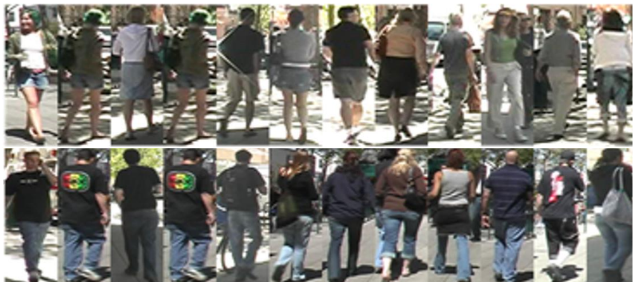


Figure 8: Typical failure cases. The first image in each row is the probe image, the second one is the correctly matched gallery image, followed by the ranking list by GCT.

4.4 Analysis on Typical Failure Cases

We record the typical failure cases to explore the limitations of the proposed GCT algorithm. As shown in Figure 8, when severe self-occlusion occurs, the appearances of the same person may be dramatically different across different camera views, rendering it difficult for GCT to establish enough local correspondences between matched image pairs. Even though, the proposed GCT algorithm can rank visually similar images with the probe ahead of others, which is valuable for further manual verification.

5 Conclusion

This paper proposes a novel GCT model for Re-ID task. The GCT model aims to learn a set of patch-wise correspondence templates from positive image pairs in the training set, and then transfer these correspondences to test image pairs with similar pose-pair configurations for distance computation. Owing to the part-based strategy as well as the incorporation of the body context information, the GCT model is capable of dealing with the problem of spatial misalignment caused by large variations in viewpoints and human poses. Extensive experiments on five challenging datasets demonstrate the effectiveness of the GCT model.

Acknowledgments This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 61671289, 61221001, 61528204, 61771303 and 61571261), and by US National Science Foundation (NSF) (Grant No. 1618398 and 1350521). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Ahmed, E.; Jones, M. J.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*.

Andriluka, M.; Roth, S.; and Schiele, B. 2010. Monocular 3d pose estimation and tracking by detection. In *CVPR*.

Baltieri, D.; Vezzani, R.; and Cucchiara, R. 2011. 3dpes: 3d people dataset for surveillance and forensics. In *Joint ACM Workshop on Human Gesture and Behavior Understanding*.

Chen, D.; Yuan, Z.; Hua, G.; Zheng, N.; and Wang, J. 2015. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*.

Chen, D.; Yuan, Z.; Chen, B.; and Zheng, N. 2016. Similarity learning with spatial constraints for person re-identification. In *CVPR*.

Chen, J.; Wang, Y.; Qin, J.; Liu, L.; and Shao, L. 2017. Fast person re-identification via cross-camera semantic binary transformation. In *CVPR*.

Chen, S.; Guo, C.; and Lai, J. 2016. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions. Image Processing* 25(5):2353–2367.

Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; and Zheng, N. 2016. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *CVPR*.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*.

Gray, D.; Brennan, S.; and Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshop*.

Jose, C., and Fleuret, F. 2016. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*.

Karanam, S.; Li, Y.; and Radke, R. J. 2015. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*.

Köstinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*.

Li, Z.; Chang, S.; Liang, F.; Huang, T. S.; Cao, L.; and Smith, J. R. 2013. Learning locally-adaptive decision functions for person verification. In *CVPR*.

Li, W.; Zhao, R.; and Wang, X. 2012. Human reidentification with transferred metric learning. In *ACCV*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Liaw, A., and Wiener, M. 2002. Classification and regression by random forest. *R News*.

Martinel, N.; Das, A.; Micheloni, C.; and Roy-Chowdhury, A. K. 2016. Temporal model adaptation for person re-identification. In *ECCV*.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*.

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

Oreifej, O.; Mehran, R.; and Shah, M. 2010. Human identity recognition in aerial images. In *CVPR*.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2015. Learning to rank in person re-identification with metric ensembles. In *CVPR*.

Pedagadi, S.; Orwell, J.; Velastin, S. A.; and Boghossian, B. A. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*.

Roth, P. M.; Hirzer, M.; Köstinger, M.; Beleznai, C.; and Bischof, H. 2014. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*.

Shen, Y.; Lin, W.; Yan, J.; Xu, M.; Wu, J.; and Wang, J. 2015. Person re-identification with correspondence structure learning. In *ICCV*.

Shi, Z.; Hospedales, T. M.; and Xiang, T. 2015. Transferring a semantic representation for person re-identification and search. In *CVPR*.

Suh, Y.; Adamczewski, K.; and Lee, K. M. 2015. Subgraph matching using compactness prior for robust feature correspondence. In *CVPR*.

Xiao, T.; Li, H.; Ouyang, W.; and Wang, X. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*.

Xiong, F.; Gou, M.; Camps, O. I.; and Sznaier, M. 2014. Person re-identification using kernel-based metric learning methods. In *ECCV*.

Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; and Li, S. Z. 2014. Salient color names for person re-identification. In *ECCV*.

Yang, Y.; Wen, L.; Lyu, S.; and Li, S. Z. 2017. Unsupervised learning of multi-level descriptors for person re-identification. In *AAAI*.

Zhang, Y.; Li, X.; Zhao, L.; and Zhang, Z. 2016a. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*.

Zhang, Y.; Li, B.; Lu, H.; Irie, A.; and Ruan, X. 2016b. Sample-specific SVM learning for person re-identification. In *CVPR*.

Zhang, L.; Xiang, T.; and Gong, S. 2016. Learning a discriminative null space for person re-identification. In *CVPR*.

Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wng, X.; and Tang, X. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *CVPR*.

Zhao, R.; Ouyang, W.; and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.