Similarity Learning with Spatial Constraints for Person Re-identification

Dapeng Chen, Zejian Yuan, Badong Chen, Nanning Zheng
Xi'an Jiaotong University, China
dapengchenxjtu@foxmail.com, {yuan.ze.jian, chenbd, nnzheng}@mail.xjtu.edu.cn

Abstract

Pose variation remains one of the major factors that adversely affect the accuracy of person re-identification. Such variation is not arbitrary as body parts (e.g. head, torso, legs) have relative stable spatial distribution. Breaking down the variability of global appearance regarding the spatial distribution potentially benefits the person matching. We therefore learn a novel similarity function, which consists of multiple sub-similarity measurements with each taking in charge of a subregion. In particular, we take advantage of the recently proposed polynomial feature map to describe the matching within each subregion, and inject all the feature maps into a unified framework. The framework not only outputs similarity measurements for different regions, but also makes a better consistency among them. Our framework can collaborate local similarities as well as global similarity to exploit their complementary strength. It is flexible to incorporate multiple visual cues to further elevate the performance. In experiments, we analyze the effectiveness of the major components. The results on four datasets show significant and consistent improvements over the state-of-the-art methods.

1. Introduction

Person re-identification refers a task of associating a same person in different camera views. It plays a crucial role for applications such as long-term person tracking [9], multi-person association [28], group behavior analysis [38], etc. Similarity measuring serves as a major step for a person re-identification system. An ideal measurement is a matching rule that yields higher matching score for the image pairs belonging to the same person than the pairs belonging to different persons. The similarity measurement can be pre-defined or be learned. The former type adopts the off-the-shelf distance metric such as Euclidean distance [10], Bhattacharyya distance[6], and covariance distance [22, 1], while the latter type tries to exploit the inherent invariance between image pairs [12, 27, 13, 24]. By making use of the training data, learning-based models generally enjoy better performance than the learning-free methods. However, most similarity learning just focus on a "holistic" measurement, which discards the geometric structure of pedestrians and can not further exploit the discriminative power.

We argue that similarity learning should obey certain spatial constraints, which indicates the matching of certain body part should follow its spatial distribution. For example, the region containing the head of a person should be compared with the region containing the head rather than the region containing the feet. With such constraints, each region has its own similarity measurement that excels at handling the special intra-person variation within it. The combination of multiple measurements is more flexible to exploit the global invariance than a holistic one. Besides, enforcing the matching within the corresponding region can effectively reduce the risk of mismatching and become more robust to partial occlusion.

We combine such constraints with recently proposed polynomial feature maps [4]. As each feature map can describe the matching within each local region, we employ multiple feature maps to represent the different regions simultaneously and inject all of them into a unified learning framework. The framework not only outputs the similarity measurements for each local region, but also makes a better consistency among these measurements. Our framework is able to collaborate the local measurements with global measurement to exploit their complementary power, and it is flexible to incorporate multiple visual cues to further improve the performance.

The main contributions are threefold: (1) We present a novel similarity function to investigate how the spatial constraints can benefit the similarity learning for person re-identification. (2) We propose a convex objective function as well as an efficient optimization algorithm for it. (3) We operate in-depth experiments to analyze various aspects of our approach, and the final results outperform the state-of-the-arts over four benchmarks.

2. Related Work

A comprehensive survey can be found in [11]. Here, we briefly review the most relevant works.

Spatial constraints. As pedestrians have relative stable geometric structure, spatial constraints have been widely adopted for person image representation [29, 10, 36, 17, 20, 30, 26, 34]. Wang et al. [29] utilized shape and appearance context to capture the spatial relations between appearance labels. Farenzena et al. [10] considered the symmetric and asymmetric prior of human body part, and extracted local features from each part. Methods [31, 36] adopting SPM-like [17] representation separated the image into several horizontal stripes, and used unsupervised Bag-of-Words model to represent each stripe region. Many other works extracted dense local features [20, 30, 26, 37], and concatenated those descriptors to implicitly encode the spatial layout of the person. After feature extraction, all these methods usually employ a "holistic" similarity measurement for all the extracted descriptors without further utilizing the spatial relation.

Only a few works imposed the spatial constraints for similarity measuring. Zhao et al. [34, 33] matched each patch in one image with the neighbouring patches in the other image, where the matching rule is pre-defined and computational cost is large. Different from their work, we impose a similarity function with spatial constraints, the similarity function is much more efficient than exhaustive matching and the spatial constraints can better exploit the discriminative ability from data.

Similarity learning. Similarity learning has gradually shown its effectiveness in person re-identification. Most works learn a similarity measurement based upon Mahalanobis distance. Among them, Hirzer et al. [13] relaxed the PSD constraint of the metric and obtained a simplified formulation with reasonable effectiveness. Köstinger et al. [15] proposed an efficient metric computation method motivated by the log likelihood ratio test of two Gaussian distributions. In [18], Li et al. proposed Locally-Adaptive Decision Function, which can be viewed as a joint model of a distance metric and a locally adapted thresholding rule. Zheng et al. [37] made use of the triplet relationship between a positive pair and a negative pair, optimizing the relative distance comparison.

Recently, a similarity measurement built on polynomial feature map has been proposed [4]. The feature map explicitly represents the matching of two images, and its regularized form is connected to Mahalanobis distance and cross-patch similarity. Our work takes advantage of the feature map to represent the matching within each sub-region. The linear form of the feature maps allows us to conveniently exploit the complementary strength of different local regions.

3. Our Approach

In this section, we first revisit the polynomial feature map. Based upon the map, we impose spatial constraints for similarity learning, and formulate the learning problem specifically designed for person re-identification.

3.1. Polynomial Feature Map [4]

To measure the similarity between image descriptors $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^{d\times 1}$, we aim to learn a basic similarity function $f(\mathbf{x}_a, \mathbf{x}_b)$ that can yield high score when $\mathbf{x}_a$ and $\mathbf{x}_b$ are similar. The similarity function is in linear form:

$$f(\mathbf{x}_a, \mathbf{x}_b) = \langle \phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W} \rangle_F. \tag{1}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. To take advantage of both Mahalanobis distance and a bilinear similarity metric, we decompose Eq. 1 to be

$$f(\mathbf{x}_a, \mathbf{x}_b) = \langle \phi_M(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_M \rangle_F + \langle \phi_B(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_B \rangle_F, \tag{2}$$

where $\mathbf{W} = [\mathbf{W}_M, \mathbf{W}_B]$ and $\phi(\mathbf{x}_a, \mathbf{x}_b) = [\phi_M(\mathbf{x}_a, \mathbf{x}_b), \phi_B(\mathbf{x}_a, \mathbf{x}_b)]$. More specifically,

$$\phi_M(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top, \phi_B(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top.$$

The part $\langle \phi_M(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_M \rangle_F = (\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{W}_M (\mathbf{x}_a - \mathbf{x}_b)$, is connected to Mahalanobis distance. As we want to achieve high score when $\mathbf{x}_a$ and $\mathbf{x}_b$ are similar, $\mathbf{W}_M$ should be negative semi-definite. The part $\langle \phi_B(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}_B \rangle_F = \mathbf{x}_a^\top \mathbf{W}_B \mathbf{x}_b + \mathbf{x}_b^\top \mathbf{W}_B \mathbf{x}_a$, corresponds to bilinear similarity [3]. Both parts ensure the effectiveness of $f(\mathbf{x}_a, \mathbf{x}_b)$.

The feature map $\phi(\mathbf{x}_a, \mathbf{x}_b)$ is composed by the elements in $\varphi(\mathbf{x}_a, \mathbf{x}_b) = [\mathbf{x}_a\mathbf{x}_a^\top, \mathbf{x}_b\mathbf{x}_b^\top, \mathbf{x}_a\mathbf{x}_b^\top, \mathbf{x}_b\mathbf{x}_a^\top]$. Particularly, $\varphi(\mathbf{x}_a, \mathbf{x}_b)$ is induced by the second order polynomial kernel $k(\mathbf{z}, \mathbf{z}') = (\mathbf{z}^\top\mathbf{z}')^2 = \langle \varphi(\mathbf{z}), \varphi(\mathbf{z}') \rangle_F$, where $\mathbf{z} = [\mathbf{x}_a^\top, \mathbf{x}_b^\top]$. $\phi(\mathbf{x}_a, \mathbf{x}_b)$ re-organizes the elements in $\varphi(\mathbf{x}_a, \mathbf{x}_b)$, thus it is a regularized form of polynomial feature map.

$\phi(\mathbf{x}_a, \mathbf{x}_a)$ conveys the matching information of $\mathbf{x}_a$ and $\mathbf{x}_b$. In the case that $\mathbf{x}_a$ and $\mathbf{x}_b$ are patch-wise descriptors of an image (each entry or sub-vector corresponds to a block of the image), $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$ focus on measuring the similarity for descriptors at the same position. $\phi_B(\mathbf{x}_a, \mathbf{x}_b)$ matches each patch in one image with all the patches in the other image, and all the cross-patch similarities are attained as $\mathbf{x}_a\mathbf{x}_b^\top$ and $\mathbf{x}_b\mathbf{x}_a^\top$. To reduce the dimensionality of the feature map, method [4] performs PCA for $\mathbf{x}_a$ and $\mathbf{x}_b$ before forming the feature map, which keeps the effectiveness.

3.2. Spatially Constrained Similarity Function

The flowchart of the overall similarity function is illustrated in Fig. 1. We explain it with more details as follows.

Regional feature map. An image is partitioned into R non-overlap horizontal stripe regions. For each region, we divide it into a collection of overlapped patches, and extract color and texture histograms from each patch. The extracted histograms belonging to a same stripe region are concatenated together. After that, we apply PCA to reduce the

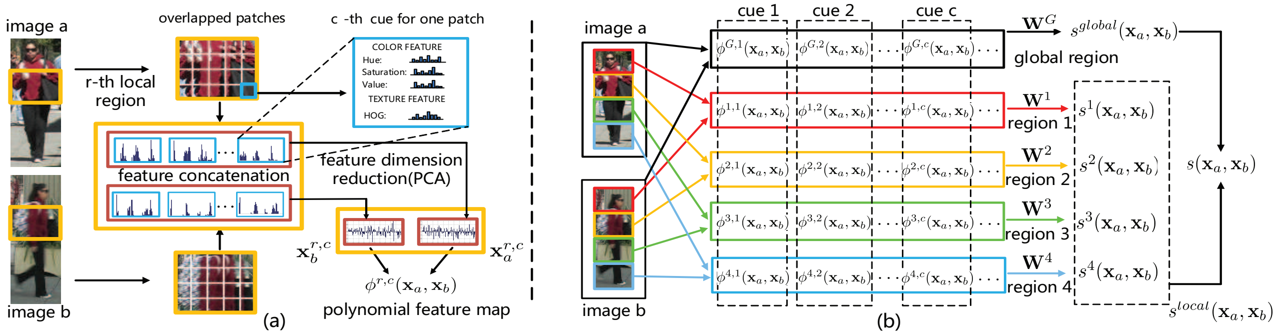
image a
overlapped patches
c -th cue for one patch
r-th local region
COLOR FEATURE
Hue:
Saturation:
Value:
TEXTURE FEATURE
HOG:
feature concatenation
feature dimension reduction(PCA)
$\mathbf{x}_b^{r,c}$
$\mathbf{x}_a^{r,c}$
$\phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$
polynomial feature map
image b
(a)
image a
cue 1
cue 2
cue c
$\phi^{G,1}(\mathbf{x}_a, \mathbf{x}_b)$ $\phi^{G,2}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$ $\phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$
$\phi^{1,1}(\mathbf{x}_a, \mathbf{x}_b)$ $\phi^{1,2}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$ $\phi^{1,c}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$
$\phi^{2,1}(\mathbf{x}_a, \mathbf{x}_b)$ $\phi^{2,2}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$ $\phi^{2,c}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$
$\phi^{3,1}(\mathbf{x}_a, \mathbf{x}_b)$ $\phi^{3,2}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$ $\phi^{3,c}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$
$\phi^{4,1}(\mathbf{x}_a, \mathbf{x}_b)$ $\phi^{4,2}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$ $\phi^{4,c}(\mathbf{x}_a, \mathbf{x}_b)$ $\cdots$
$\mathbf{W}^G$
$s^{global}(\mathbf{x}_a, \mathbf{x}_b)$
global region
$\mathbf{W}^1$
region 1
$s^1(\mathbf{x}_a, \mathbf{x}_b)$
$\mathbf{W}^2$
region 2
$s^2(\mathbf{x}_a, \mathbf{x}_b)$
$\mathbf{W}^3$
region 3
$s^3(\mathbf{x}_a, \mathbf{x}_b)$
$\mathbf{W}^4$
region 4
$s^4(\mathbf{x}_a, \mathbf{x}_b)$
$s(\mathbf{x}_a, \mathbf{x}_b)$
$s^{local}(\mathbf{x}_a, \mathbf{x}_b)$
image b
(b)

Figure 1: System overview. (a) The process of matching representation for r-th local region with c-th visual cue, which includes patch feature extraction, feature concatenation, PCA based dimensionality reduction and feature map generation. (b) The flowchart of our similarity function. Our similarity combines one global similarity for the whole image region and multiple local similarities for associated local regions with multiple visual cues.

dimensionality and obtain the region descriptor $\mathbf{x}^r$ for the r-th stripe, where $r \in \{1, ..., R\}$.

A stripe region r can be described by C visual cues $\{\mathbf{x}^{r,1}, ..., \mathbf{x}^{r,c}, ..., \mathbf{x}^{r,C}\}$, thus $\mathbf{x}_a$ and $\mathbf{x}_b$ accordingly form C polynomial feature maps for the r-th region, i.e., $\{\phi^{r,1}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{r,C}(\mathbf{x}_a, \mathbf{x}_b)\}$, where $\phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b) = \phi(\mathbf{x}_a^{r,c}, \mathbf{x}_b^{r,c})$. As different feature maps can describe the matching in different aspects of view, multiple feature maps can encode more comprehensive information about the matching.

Local similarity integration. In order to exploit the complementary strengths of multiple visual cues within a local region, we employ a linear similarity function to combine them together for the r-th region:

$$s^r(\mathbf{x}_a, \mathbf{x}_b) = \sum_{c=1}^{C} \langle \phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{r,c} \rangle_F, \tag{3}$$

where $\mathbf{W}^{r,c} = [\mathbf{W}_M^{r,c}, \mathbf{W}_B^{r,c}]$, and $\mathbf{W}_M^{r,c}, \mathbf{W}_B^{r,c}$ corresponds to $\phi_M^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B^{r,c}(\mathbf{x}_a, \mathbf{x}_b)$ respectively. For all the R regions, the integrated local similarity score is simply represented as:

$$s^{local}(\mathbf{x}_a, \mathbf{x}_b) = \sum_{r=1}^{R} s^r(\mathbf{x}_a, \mathbf{x}_b). \tag{4}$$

Global-local collaboration. The feature maps of a local region can not describe the matching of large patterns across the stripes. To compensate the insufficiency of local similarity, we also make use of the polynomial feature map for the whole image, yielding global similarity:

$$s^{global}(\mathbf{x}_a, \mathbf{x}_b) = \sum_{c=1}^{C} \langle \phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{G,c} \rangle_F, \tag{5}$$

where $\phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b) = \phi(\mathbf{x}_a^{G,c}, \mathbf{x}_b^{G,c})$ and $\mathbf{x}_a^{G,c}, \mathbf{x}_b^{G,c}$ are the c-th type global visual descriptors for image a and image b. The global similarity and local similarity are linearly combined, and the overall similarity score is given by:

$$s(\mathbf{x}_a, \mathbf{x}_b) = s^{local}(\mathbf{x}_a, \mathbf{x}_b) + \gamma s^{global}(\mathbf{x}_a, \mathbf{x}_b). \tag{6}$$

Here γ is the hyper-parameter that mediates the local similarity and global similarity.

3.3. Learning for Person Re-identification

Regularization. As $\langle \mathbf{W}_M, \phi_M(\mathbf{x}_a, \mathbf{x}_b) \rangle_F$ corresponds to negative Mahalanobis distance (discussed in Sec. 3.1), it is reasonable for $\mathbf{W}_M$ to be negative semi-definite:

$$\mathbf{W}_M^{r,c}, \mathbf{W}_M^{G,c} \in \mathbb{S}_-^d \quad r = 1, ..., R, \quad c = 1, ..., C, \tag{7}$$

where $\mathbb{S}_-^d$ denote the set containing negative semi-definite matrices with the size of $d \times d$.

Considering the construction of $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B(\mathbf{x}_a, \mathbf{x}_b)$, both feature maps are generated by the out product of certain feature vector. If some dimensions of the vector are not effective for discrimination, the elements in corresponding columns or rows of the feature map tend to be less effective. The assumption indicates that the effective elements in polynomial feature map would appear in group, we therefore impose mixed norm for the corresponding coefficient matrices.

$$R(\mathcal{W}) = \sum_{\mathbf{W} \in \mathcal{W}} \|\mathbf{W}\|_{2,1}, \tag{8}$$

where $\|\mathbf{W}\|_{2,1} := \sum_{i=1}^{d} \|\mathbf{W}_{i.}\|_2$[32], and $\mathcal{W}$ is the coefficient set defined by $\mathcal{W} = \{\mathbf{W}_M^{1,c}, \mathbf{W}_B^{1,c}, ..., \mathbf{W}_M^{r,c}, \mathbf{W}_B^{r,c}, ..., \mathbf{W}_M^{R,c}, \mathbf{W}_B^{R,c}, \mathbf{W}_M^{G,c}, \mathbf{W}_B^{G,c}\}_{c=1}^{C}$.

Relaxed loss term. The training data for person re-identification can be organized as follows. Given the descriptors of probe images $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_N\}$, $\mathbf{x}_n$ is associated with two sets of gallery images: a positive set $\mathcal{X}_n^+$ composed of the descriptors about the same person with $\mathbf{x}_n$ and a negative set $\mathcal{X}_n^-$ composed of the descriptors about different persons to $\mathbf{x}_n$. We consider to enforce the relative comparison and propose a relaxed loss term:

$$L(\mathcal{W}) = \frac{1}{N} \sum_{n=1}^{N} [1 - \frac{\sum_{\mathbf{x}_i \in \mathcal{X}_n^+, \mathbf{x}_j \in \mathcal{X}_n^-} s(\mathbf{x}_n, \mathbf{x}_i) - s(\mathbf{x}_n, \mathbf{x}_j)}{|\mathcal{X}_n^+| \cdot |\mathcal{X}_n^-|}]_+, \tag{9}$$

where $[.]_+$ denotes the hinge loss. Given a probe descriptor $\mathbf{x}_n$, instead of forcing every positive pair to achieve a higher score than negative pairs [37, 27], we require the average score of positive pairs should be higher than the average score of the negative pairs at least by a margin 1. The relaxed loss term only consists of N constraints, largely accelerating the training compared with the non-relaxed one.

Objective function. According to Eqs. 7, 8 and 9, the objective function for person re-identification is given by:

$$\begin{aligned}&\min_{\mathcal{W}} L(\mathcal{W}) + \lambda R(\mathcal{W})\\ \text{s.t.}\quad &\mathbf{W}_M^{r,c}, \mathbf{W}_M^{G,c} \in \mathbb{S}_-^d, \quad c = 1, ..., C, r = 1, ..., R.\end{aligned} \tag{10}$$

3.4. Discussion

Our method is related to Spatial Pyramid Matching (SPM) [17]. Both of them employ subregions to encode the spatial layout information for matching. SPM employs the intersection kernel to compute the similarity for each subregion, defines a pyramid match kernel to combine the similarities in different pyramid layers, and takes the direct sum of the similarities of different type features. Instead of designing the three factors, we learn them from the data. The learned similarity measurements can adapt to each regions each visual cue, and can effectively combine them together. From another aspect of view, SPM is a more general measurement that is used for image classification, while our similarity function encodes more details about a specific class thus is suitable for identification or verification.

Our method is compared with previous ensemble approaches [35, 25, 30]. Although employing multiple measurements for re-identification, both the similarity functions and the learning approaches are quite different. Firstly, the multiple measurements in other methods are for the whole person image, while our measurements are with certain spatial attributes. Secondly, approaches[35, 25, 30] build their final similarity function in two-stage. They first learn the similarity measurements of different features independently, then utilize ensemble techniques to combine the independent scores. In contrast, we learn the multiple measurements simultaneously, and can better exploit the consistency between different types of features.

4. Optimization

To clarify the notation, we first concatenate the C feature maps in each sub-region together:

$$\phi^r(\mathbf{x}_a, \mathbf{x}_b) = [\phi^{r,1}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{r,c}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{r,C}(\mathbf{x}_a, \mathbf{x}_b)],$$
$$\phi^G(\mathbf{x}_a, \mathbf{x}_b) = [\phi^{G,1}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{G,c}(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^{G,C}(\mathbf{x}_a, \mathbf{x}_b)].$$

Accordingly, $\mathbf{W}^r = [\mathbf{W}^{r,1}, ..., \mathbf{W}^{r,c}, ..., \mathbf{W}^{r,C}]$ and $\mathbf{W}^G = [\mathbf{W}^{G,1}, ..., \mathbf{W}^{G,c}, ..., \mathbf{W}^{G,C}]$ are the coefficients for $\phi^r(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi^G(\mathbf{x}_a, \mathbf{x}_b)$. Let $\Phi(\mathbf{x}_a, \mathbf{x}_b) = [\phi^1(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^r(\mathbf{x}_a, \mathbf{x}_b), ..., \phi^R(\mathbf{x}_a, \mathbf{x}_b), \gamma\phi^G(\mathbf{x}_a, \mathbf{x}_b)]$ and $\mathbf{U} = [\mathbf{W}^1, ..., \mathbf{W}^r, ..., \mathbf{W}^R, \mathbf{W}^G]$, the similarity function of Eq. 6 can be simply computed as:

$$s(\mathbf{x}_a, \mathbf{x}_b) = \langle \Phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{U} \rangle_F. \tag{11}$$

Algorithm 1 The ADMM optimization.

```
Input: Dateset $\mathcal{D} = \{\mathbf{x}_i, \mathcal{X}_i^+, \mathcal{X}_i^-\}_{i=1}^n$,
Output: Coefficient $\mathbf{U}$
for $l = 0, ..., L-1$ (until convergence) do
    Update $\mathbf{U}_1^{l+1}$ by solving Eq. 16
    Update $\mathbf{U}_2^{l+1}$ by applying prox-operator in Eq. 17
    Update $\mathbf{U}_3^{l+1}$ by projection in Eq. 18
    Update $\mathbf{\Lambda}_1^{l+1}$ and $\mathbf{\Lambda}_2^{l+1}$
end for
    $\mathbf{U} \leftarrow \mathbf{U}_3^L$
```

We note that the content within the hinge loss is linear w.r.t. coefficient $\mathbf{U}$. By defining

$$\Psi(\mathbf{x}_n) = \frac{\sum_{\mathbf{x}_i \in \mathcal{X}_n^+, \mathbf{x}_j \in \mathcal{X}_n^-} \Phi(\mathbf{x}_n, \mathbf{x}_i) - \Phi(\mathbf{x}_n, \mathbf{x}_j)}{|\mathcal{X}_n^+| \cdot |\mathcal{X}_n^-|}, \tag{12}$$

Eq. 9 is rewritten as $L(\mathbf{U}) = \frac{1}{N}\sum_{n=1}^N [1 - \langle \Psi(\mathbf{x}_n), \mathbf{U} \rangle_F]_+$.

4.1. ADMM Optimization

Our objective function forms a convex optimization problem. For the ease of optimization, we transform Eq. 10 to an equivalent problem:

$$\min_{\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3} g_1(\mathbf{U}_1) + g_2(\mathbf{U}_2) + g_3(\mathbf{U}_3), \quad \text{s.t.} \mathbf{U}_1 = \mathbf{U}_2 = \mathbf{U}_3, \tag{13}$$

where $g_1(\mathbf{U}) = L(\mathbf{U})$, $g_2(\mathbf{U}) = \lambda R(\mathbf{U})$, and $g_3(\mathbf{U}) = \infty\delta[\mathbf{U} \notin \mathcal{C}]$. Here, $\mathcal{C}$ is a closed convex set defined from the constraints in Eq. 7, and $\delta[\cdot]$ is an indicator function which takes one if the argument is true and zeros otherwise. By performing ADMM, we have following iterations:

$$\mathbf{U}_1^{l+1} = \arg\min_{\mathbf{U}_1} g_1(\mathbf{U}_1) + \frac{\rho}{2}\|\mathbf{U}_1 - (\mathbf{U}_3^l - \mathbf{\Lambda}_1^l)\|_F^2 \tag{14}$$

$$\mathbf{U}_2^{l+1} = \arg\min_{\mathbf{U}_2} g_2(\mathbf{U}_2) + \frac{\rho}{2}\|\mathbf{U}_2 - (\mathbf{U}_3^l - \mathbf{\Lambda}_2^l)\|_F^2 \tag{15}$$

$$\mathbf{U}_3^{l+1} = \arg\min_{\mathbf{U}_3} g_3(\mathbf{U}_3) + \rho\|\mathbf{U}_3 - \frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \mathbf{\Lambda}_1^l + \mathbf{\Lambda}_2^l)\|_F^2$$

$$\mathbf{\Lambda}_1^{l+1} = \mathbf{\Lambda}_1^l + \mathbf{U}_1^{l+1} - \mathbf{U}_3^{l+1}, \quad \mathbf{\Lambda}_2^{l+1} = \mathbf{\Lambda}_2^l + \mathbf{U}_2^{l+1} - \mathbf{U}_3^{l+1},$$

where ρ is a scalar value called the penalty parameter, and $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are scaled dual variables. The whole update procedures are summarized in Algorithm. 1. The details about the updates of $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$ are presented below:

The update of $\mathbf{U}_1$. Eq. 14 is a convex problem. We consider to optimize $\mathbf{U}_1^{l+1}$ from its dual form.

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2\rho}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} - \mathbf{b}^\top\boldsymbol{\alpha}, \quad \text{s.t.} \quad 0 \leq \alpha_n \leq \frac{1}{N}, \forall n, \tag{16}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{N\times 1}$ are dual variables and α_n is its nth element. The element of $\mathbf{b} \in \mathbb{R}^{N\times 1}$ is defined as: $b_n = \langle \mathbf{U}_3^l - \mathbf{\Lambda}_1^l, \Psi(\mathbf{x}_n)\rangle_F - 1$. $\mathbf{H} \in \mathbb{R}^{N\times N}$ is the kernel matrix with

$\mathbf{H}_{ij} := \langle \Psi(\mathbf{x}_i), \Psi(\mathbf{x}_j) \rangle_F$. Eq. 16 is a standard quadratic programming problem. As $\mathbf{H}$ can be pre-computed, optimizing the dual form is quite efficient. With optimal $\boldsymbol{\alpha}^*$, $\mathbf{U}_1^{l+1}$ is updated by $\frac{1}{\rho}\sum_{n=1}^{N} \alpha_n^* \Psi(\mathbf{x}_n) + \mathbf{U}_3^l - \mathbf{\Lambda}_1^l$.

The update of $\mathbf{U}_2$. As $\mathbf{U}_2 \in \mathbb{R}^{d \times 2dC(R+1)}$, the problem of Eq.15 can be decomposed into $2C(R+1)$ subproblems with each optimizing the coefficients corresponding to a feature map with the size of $d \times d$. Let $\mathbf{U}_2^s \in \mathbb{R}^{d \times d}$ denote the s-th sub-matrix of $\mathbf{U}_2$ and $\mathbf{A}^s \in \mathbb{R}^{d \times d}$ be the corresponding sub-matrix of $\mathbf{U}_3^l - \mathbf{\Lambda}_2^l$, a separate problem of Eq. 15 is: $\min_{\mathbf{U}_2^s} \lambda \|\mathbf{U}_2^s\|_{2,1} + \frac{\rho}{2}\|\mathbf{U}_2^s - \mathbf{A}^s\|_F^2$. The optimal solution is given by a prox-operator[16]:

$$(\mathbf{U}_2^s)_{ij} = \mathbf{A}_{ij}^s \left[1 - \frac{\lambda/\rho}{\|\mathbf{A}_{i.}^s\|_2}\right]_+, \tag{17}$$

where $\mathbf{A}_{i.}^s$ is the i-th row of $\mathbf{A}^s$. We apply the prox-operator for all the sub-matrices, obtaining $\mathbf{U}_2^{l+1}$.

The update of $\mathbf{U}_3$. $\mathbf{U}_3$ is updated through the projection:

$$\mathbf{U}_3^{l+1} = \Pi_{\mathcal{C}}\left[\frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \mathbf{\Lambda}_1^l + \mathbf{\Lambda}_2^l)\right], \tag{18}$$

where $\Pi_{\mathcal{C}}$ is the Euclidean projection onto set $\mathcal{C}$. Note that sub-matrices of $\frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \mathbf{\Lambda}_1^l + \mathbf{\Lambda}_2^l)$ that corresponds to $\{\mathbf{W}_B^{r,c}, \mathbf{W}_B^{G,c}\}_{r=1,c=1}^{R,C}$ may not be symmetric, directly projecting a non-symmetric matrix onto $\mathbb{S}_-^d$ is difficult. We operate a separated ADMM, including two iterative projection steps. One is to project the sub-matrices onto the $\mathbb{S}^d$ by $f(\mathbf{W}) := \frac{1}{2}(\mathbf{W} + \mathbf{W}^\top)$, the other is to project symmetric matrices onto $\mathbb{S}_-^d$ by cropping the positive eigenvalues to be zeros. The details about the updating procedures are relegated to the supplementary files.

5. Experiments

5.1. Experimental Setup

Visual cues. We divide each subregion into a set of local patches as shown in Fig. 1a. From each patch, we extract 6 types of basic feature HSV^1, HSV^2, LAB^1, LAB^2, HOG and SILTP. Among them, HSV^1 and LAB^1 are 8×8×8 joint histograms, and HSV^2 and LAB^2 are 48 bin concatenated histograms with each channel having 16 bins, HOG[7] and SILPT[20] are texture descriptors.

The four visual cues C_1,C_2,C_3,C_4 concatenate both color and texture features, which are organized as HSV^1/HOG, HSV^2/SILPT, LAB^1/SILPT, LAB^2/HOG. We employ PCA to reduce their dimension, and do a whitening process to limit the impact of co-occurrence [14]. The resulting descriptors are normalized to have unit L_2 norms.

Parameter setting. We empirically set the number of local region $R = 4$, the parameter for ADMM learning $\rho = 0.001$. The PCA reduced dimension d depends on the size of training data, we set d to be 120, 60, 60 and 500 for VIPeR [12], GRID [21], 3DPES [2] and Market-1501 [36], respectively. The tradeoff parameter λ in Eq. 10 is selected via cross validation.

Evaluation protocol. Our experiments follow the evaluation protocol in [12]. The dataset is separated into the training set and test set, where images of a same person can only appear in either set. The test set is further divided into probe set and gallery set, and the two sets contains the different images of a same person. We match each probe image with every image in gallery set, and rank the gallery images according to the similarity score. The results are evaluated by CMC curves [12], an estimate of the expectation of finding the correct match in the top n matches.

5.2. Comparison to state-of-the-art Approaches

We term the proposed **S**patially **C**onstrained **S**imilarity function on **P**olynomial feature map as SCSP. Besides, we also report the results of two variants G-All and L-All, where L-All corresponds to the integrated local similarity defined in Eq. 4, and G-All corresponds to the global similarity defined in Eq. 5. All the three methods are equipped with four visual cues.

VIPeR [12]. The VIPeR dataset is a challenging test bed for person re-identification. It contains 632 persons, and each person has 2 images taken from camera A and B with different viewpoints and illumination conditions. We randomly select 316 persons to form the training set, and select the remaining 316 persons to form the test set. The procedure is repeated 10 times to get an average performance.

We present the comparison results in Fig. 2a and Tab. 1. SCSP achieves the new state of art. Its rank-1 matching rate 53.54% outperforms the second best one ME by 7.65%. It also significantly improves Polymap, which is the original method employing polynomial feature map for person re-identification. By comparing L-All and G-All, we find that imposing spatial constraints improves the final performance (51.04% v.s. 48.10%). Such benefit is more significant when only using visual cue C_1 (43.70% v.s. 37.31%) as analyzed in Fig. 4a.

GRID [21]. The GRID dataset consists of 1275 person images. Among them, there are 250 pedestrian image pairs. Images in each pair belongs to a same person but are captured from different camera views. Besides, there are 775 additional person images that do not belong to any of the 250 persons. For the experiment, 10 partitions of the training and test samples have already been provided by the dataset. For each partition, 125 image pairs are used for training, and the remaining 125 image pairs and the 775 irrelevant images are used for testing. They form 125 probe images and 900 gallery images in one test.

Similar to the performance on VIPeR, SCSP significantly outperforms the previous state-of-the-arts, achieving 24.24% rank-1 matching rate. The CMC curves in Fig. 2b

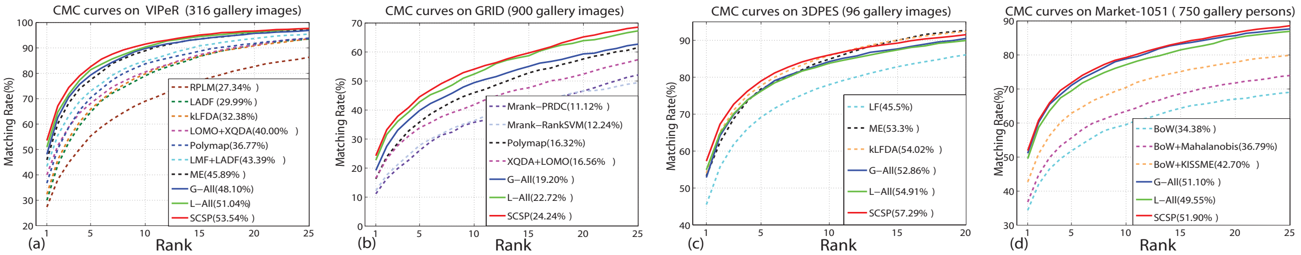

Figure 2: CMC curves for method comparison on (a) the VIPeR dataset with 316 gallery images, (b) the GRID dataset with 900 gallery images, (c) the 3DPES dataset with 96 gallery images, (d) the Market-1501 dataset with 750 gallery persons. The rank-1 matching rates are shown after the method names.

Table 1: Comparison of top-n matching rate(%) on the VIPeR dataset.

Methods	Top n matching rate (%) on VIPeR				
	r = 1	r = 5	r = 10	r = 15	r = 20
RPLM[13]	27.34	55.30	69.02	77.12	82.69
LADF[18]	29.99	64.71	79.00	86.71	91.29
kLFDA[30]	32.38	65.88	79.82	86.79	90.83
LOMO+XQDA[19]	40.00	68.13	80.51	87.37	91.08
Polymap [4]	36.77	70.35	83.70	88.73	91.74
Mirror-KMFA [5]	42.97	75.82	87.28	- -	94.84
LMF+LADF [35]	43.39	73.04	84.87	90.85	93.70
ME [25]	45.89	77.40	88.87	93.52	95.84
G-All	48.10	79.30	89.78	93.48	95.76
L-All	51.04	81.39	90.35	94.49	96.30
SCSP	**53.54**	**82.59**	**91.49**	**95.09**	**96.65**

Table 2: Comparison of top-n matching rate(%) on the GRID dataset.

Methods	Top n matching rate (%) on GRID				
	r = 1	r = 5	r = 10	r = 15	r = 20
Mrank-PRDC[21]	11.12	26.08	35.76	41.76	46.56
Mrank-RankSVM[21]	12.24	27.84	36.32	42.24	46.56
Polymap[4]	16.32	35.84	46.00	52.80	57.60
XQDA+LOMO[19]	16.56	33.84	41.84	47.68	52.40
G-All	19.20	39.84	49.44	55.04	59.36
L-All	22.72	42.80	52.40	58.72	63.92
SCSP	**24.24**	**44.56**	**54.08**	**59.68**	**65.20**

Table 3: Method comparison on the 3DPES and Market-1501 datasets.

Methods	*3DPES*		
	r = 1	r = 5	r = 20
LF[26]	45.50	69.18	86.06
ME[25]	53.30	76.79	**92.78**
kLFDA[30]	54.02	77.74	92.38
G-All	52.86	76.45	90.49
L-All	54.91	76.23	89.93
SCSP	**57.29**	**78.97**	91.51

Methods	*Market-1501*	
	r = 1	mAP
BoW[36]	34.38	14.10
+Mahalanobis	36.79	15.08
+KISSME[15]	42.70	19.55
G-All	51.10	25.47
L-All	49.55	23.83
SCSP	**51.90**	**26.35**

show that L-All performs close to that of SCSP, and is much better than G-All, which indicates spatial constraints make a more important contribution on GRID.

3DPES [2]. The 3DPES dataset includes 1011 images of 192 persons captured from 8 outdoor cameras with significantly different viewpoints. The image number of each person varies from 2 to 26. We utilize the same protocol with [30, 25], where the images of 96 persons are used for training and those of the remaining 96 persons are used for testing. As each person has more than two images, to reduce the computational burden of training, we take the mean descriptor of a person as the probe descriptor $\mathbf{x}_n$ in Eq. 9.

The comparison results are shown in Fig. 2c and Tab. 3. Our method achieves the best on rank-1 and rank-5, but performs worse than ME[25] and kLFDA[30] on rank-20. The reason may be that both methods utilize non-linear kernels, which are effectiveness on this dataset, while our final similarity function (Eq. 6) is linear.

Market-1501 [36]. Market-1501 is a newly proposed large-scale dataset containing the images of 1501 persons. It consists of three parts: the training set containing of 12936 images about 751 persons, the test set containing 19732 images about the remaining 750 persons, and the query set containing 3368 images about the same 750 persons with the test set. In testing, the query set is used as probe set and the test set is used as gallery set. The training process is same with that over 3DPES, but as the gallery set has multiple images of a person, the evaluation process is slightly different. Here, the top-n matching rate indicates the expectation of finding any one of the correct matched images. Besides, mAP [36] is used to evaluate the performance.

Our approach again obtains superior results as shown in Fig. 2d and Tab. 3. However, we find G-All performs better than L-All, indicating imposing spatial constraint may have negative effects on Market-1501. We attribute the less effectiveness of spatial constraints to the severe misalignments of body parts. As shown in Fig. 3, the corresponding local regions may contain totally different parts, which violates our initial assumption about local matching. We introduce spatial constraints to handle local variation, but if the local regions are not even roughly associated, our method will hardly have any effects.

5.3. Empirical Analysis of the Proposed Method

We perform empirical analysis of our approach on the VIPeR dataset with 316 gallery images.

5.3.1 Effect of Major Components

Effect of spatial constraints. We study the effect of spatial constraints by observing how the performance changes with the number of the stripes. In particular, we construct a series of variants by dividing the images into $\{1, 2, 4, 8, 16, 32\}$ horizontal stripes, where the variant with only one stripe corresponds to the global similarity and other variants correspond to local similarities with different extents of spatial constraints. All the variants are trained with cue C_1.

Figure 3: Sample images of VIPeR, GRID, 3DPES and Market-1501. Images in the same column represent the same person.

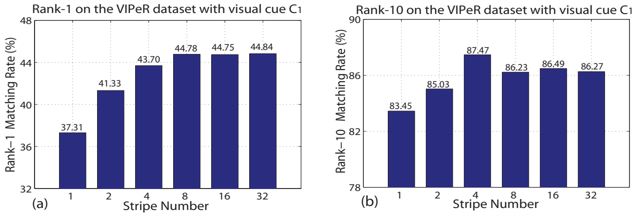

Figure 4: **Local similarity analysis:** We report how the performance changes with the number of stripes on (a) rank-1 matching rate, (b) rank-5 matching rate. All the experiments are trained and tested with cue C_1.

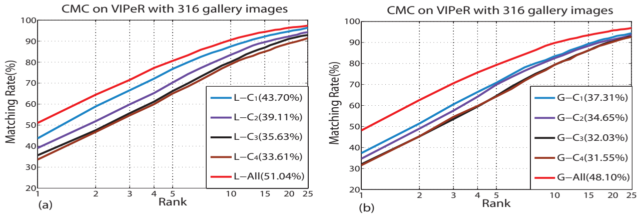

Figure 5: **Multi-cue integration analysis:** We compare the variant using 4 visual cues with the 4 variants using a single visual cue in (a),(b), where variants in (a) measure the local similarities, while the variants in (b) measures the global similarities.

The results in Fig. 4 show that the integrated local similarity is more effective than global similarity even by dividing the images into 2 stripes. Generally, more stripes tend to yield higher rank-1 matching rate. The results increase quickly with the strip number up to 8 and stay stable afterwards. The situation is slightly different for rank-10 matching rate, it shows that employing more than 4 stripes will decrease the performance. One possible reason is that small stripes are less robust when the persons in two images are misaligned along the vertical direction. The computational complexity is linear w.r.t. the number of stripes, we select the stripe number $R = 4$ in this work, which turns out to be a suitable mediation between effectiveness and efficiency.

To better understand the effectiveness of spatial constraints, we compare SCSP with other metric learning methods using a single visual cue C_1, These methods include LADF[18], KISSME[15], MFA-χ^2[30] and XQDA[19]. We also evaluate our global and local similarity using C_1, denoted by G-C_1 and L-C_1. In Tab. 4, G-C_1 performs close to XQDA, L-C_1 evidently improves G-C_1 by considering the spatial constraints. SCSP-C_1 takes advantages of the two, achieving 46.65% rank-1 matching rate.

Table 4: Comparison with other Metric learning method using cue C_1.

	LADF	KISSME	MFA-χ^2	XQDA	G-C_1	L-C_1	SCSP-C_1
r=1	26.96	33.96	35.57	37.09	37.31	43.70	**46.65**
r=10	69.30	78.99	79.81	79.68	83.45	87.47	**88.67**
r=20	81.80	90.32	90.19	90.03	94.40	**96.36**	95.73

Table 5: Comparison between joint learning and sum fusion of multiple similarity measurements.

Methods	r = 1	r = 5	r = 10	r =15	r= 20
sum-SCSP	49.49	79.78	90.09	94.27	96.20
SCSP	53.54	82.59	91.49	95.09	96.65

Effect of multi-cue integration. We investigate the effect of multi-cue integration for both integrated local similarity (R=4) and global similarity. We compare the variants using 4 visual cues (L-All,G-All) with the variants using a single visual cue. In Figs. 5a and 5b, L-C_1, L-C_2, L-C_3, L-C_4 denote the variants using corresponding visual cues independently for the integrated local similarity, and G-C_1,G-C_2, G-C_3, G-C_4 are the variants for the global similarity. The two figures reflect that (1) multiple cue collaboration actually improves the variants using each visual cue individually and (2) the integrated local similarity outperforms global similarity using all kinds of visual cues. In the future, we will incorporate high-level feature descriptors such as CNN[8] and fisher vector [23], which have different properties from low-level features, to further improve the performance.

Effect of global-local collaboration. To verify the benefits of global-local collaboration, we observe the performance by adjusting the hyper-parameter γ in Eq. 6. When $\gamma = 0$, SCSP degenerates to be L-All, the global similarity gradually takes a more important role as γ increases. The trend of rank-1 and rank-10 matching rates with respect to γ are demonstrated Figs. 6a and b. With $\gamma = 1.1$, the collaborative model can achieve 53.54% rank-1 matching rate, which outperforms the rate of local similarity 51.04% and the rate of global similarity 48.10%. The matching rate on rank-10 is not significantly influenced by the collaboration, but keeps consistency with the rank-1 matching rates.

Effect of joint learning. Intuitively, integrating multiple distinguished similarity measurements will generally improve the performance, but how to effectively take their complementary strength still remains an open problem. The proposed joint learning goes beyond the sum fusion. It not only selects effective feature from each feature map but also makes consistencies between different feature maps. To verify this, we decompose SCSP into 20 similarity functions (5 regions, 4 cues), train them independently, and fuse them by sum. The comparison results are shown in Tab. 5, where SCSP consistently outperforms sum-SCSP.

5.3.2 Other Properties

Influence of training parameters. The trade-off parameter λ in Eq. 10 and penalty parameter ρ in Eq. 14 are

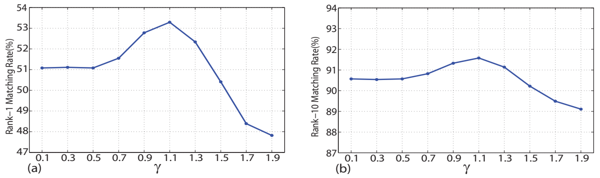

Figure 6: **Global-local collaboration analysis:** (a) The rank-1 matching rate and (b) the rank-10 matching rate of SCSP with respect to γ.

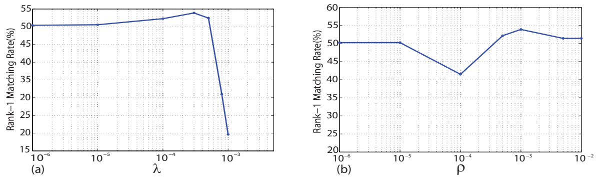

Figure 7: **Parameter analysis:** The rank-1 matching rate with respect to (a) parameter λ when $\rho = 10^{-3}$; (b) parameter ρ when $\lambda = 3 \times 10^{-4}$.

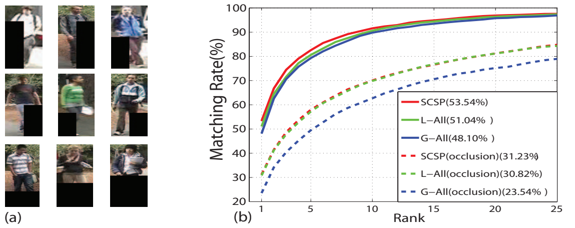

Figure 8: **Sensitivity to Occlusion**. (a) 9 occlusion patterns for probe images when testing. (b) Influences of occlusion for SCSP, L-All, G-All.

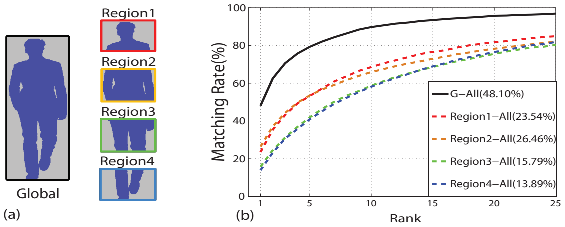

Figure 9: **Effectiveness of each region**. (a) SCSP are decomposed into 5 components associated with 5 regions. (b) Effectiveness of the 5 regions.

analyzed. As λ and ρ are mutually influenced, we show how the performance changes w.r.t. λ in Fig. 7a by fixing $\rho = 10^{-3}$, and show the influence of ρ in Fig. 7b by fixing $\lambda = 3 \times 10^{-4}$. It can be seen that too large or too small λ will lead to inferior results. This is because large λ will impose over-sparse while small λ will cause over-fitting. The influence of ρ is a little complex, but the performance w.r.t. ρ is less sensitive than that w.r.t. λ.

Sensitivity to occlusion. As each similarity measurement in our SCSP is associated with one local region. Once some region is occluded, the similarity measurements for other regions still work. Such mechanism implies that SCSP is potentially robust to occlusion. To verify this point, we design the following experiments to compare the performance of SCSP, L-All and G-All when occlusion happens.

In the experiment, we modify the probe images with various occlusion patterns at the test stage. In particular, 9 occlusion patterns given in Fig. 8a are randomly assigned to each probe image. The CMC curves show that all the three methods decrease heavily due to the occlusion (from solid line to dash line). The rank-1 matching rates of SCSP, L-All and G-All decrease 22.31%, 20.22% and 24.56%, respectively. In particular, L-All, which employs only local similarities, is the least influenced by occlusion.

Effectiveness of different local regions. It is interesting to investigate which region is most effective in SCSP. At the testing stage, we only fire the similarities measurement for a single region and set the similarity scores of other regions to be 0. The CMC curves in Fig. 9b show that the similarity measurement of the whole region evidently outperforms the one of any local region. For local similarity measurements, the ones for upper body are more effective than those for lower body. In particular, the measurement of Region2 including the torso achieves the highest rank-1 matching rate 26.46%, while the measurement of Region1 gradually performs better when the rank increases.

Runtime. Our method was implemented in MATLAB/MEX with a 3.07Ghz, 2 Cores CPU. For 128×48 person images, it takes about 0.02 second(s) per image to extract the raw features. Taken VIPER for example, at training stage, it takes about 300s to learn 20 PCA projection matrices of 632 training images, and further takes about 105s to generate both positive and negative polynomial feature maps for 316 persons. ADMM spends about 6s. At test stage, it requires 0.016s to rank 316 gallery images for a probe image. Note that we don't need to explicitly generate polynomial feature map for testing, because SCSP can be decomposed into basic similarities related to Mahalanobis distance and bilinear similarity. The testing cost is linear w.r.t. (R+1) and C.

6. Conclusions

We have proposed a novel similarity learning approach by imposing spatial constraints. We grounded our similarity function upon the polynomial feature maps, formulated a convex objective function and provided its optimization algorithm. The effectiveness of our method stems from the spatial constraints, which reduces the risk of mismatching, increases robustness to occlusion and is more flexible to handle pose variation. Our method also benefits from the multiple cue integration that is complementary to the spatial constraints. The performance is further improved by local-global similarity collaboration that measures the similarity in different scales. In the future, we will extend our framework by adopting other local region association strategies or by incorporating other types of features, which is expected to achieve better performance.

Acknowledgement. This work was supported by National Basic Research Program of China (No.2015CB351703)National Natural Science Foundation of China (No.61573280, No.61231018), and 111 Project (No.B13043).

References

[1] S. Bak, E. Corvée, F. Brémond, and M. Thonnat. Person re-identification using spatial covariance regions of human body parts. In *International Conference on Advanced Video and Signal Based Surveillance*, 2010. 1

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *International Conference on Image Analysis and Processing*, 2011. 5, 6

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11:1109–1135, 2010. 2

[4] 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 *Conference on Computer Vision and Pattern Recognition*, June 2015. 1, 2, 6

[5] Y. Chen, W. Zheng, and J. Lai. Mirror representation for modeling view-specific transform in person re-identification. In *International Joint Conference on Artificial Intelligence, 2015*, pages 3402–3408, 2015. 6

[6] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, 2011. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition*, 2005. 5

[8] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015. 7

[9] T. D'Orazio and G. Cicirelli. People re-identification and tracking from multiple cameras: A review. In *International Conference on Image Processing*, 2012. 1

[10] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition*, 2010. 1, 2

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors. *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition. Springer, 2014. 1

[12] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *International Workshop on PETS, Rio de Janeiro*, 2007. 1, 5

[13] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012. 1, 2, 6

[14] H. Jégou and O. Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *European Conference on Computer Vision*, 2012. 5

[15] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition*, 2012. 2, 6, 7

[16] M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303C324, 2009. 5

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, 2006. 2, 4

[18] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *Computer Vision and Pattern Recognition*, 2013. 2, 6, 7

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Conference on Computer Vision and Pattern Recognition*, 2015. 6, 7

[20] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Conference on Computer Vision and Pattern Recognition*, 2010. 2, 5

[21] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *International Conference on Image Processing*, 2013. 5, 6

[22] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *British Machine Vision Conference*, 2012. 1

[23] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *ECCV Workshops and Demonstrations*, 2012. 7

[24] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition*, 2012. 1

[25] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Conference on Computer Vision and Pattern Recognition*, 2015. 4, 6

[26] S. Pedagadi, J. Orwell, S. A. Velastin, and B. A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Computer Vision and Pattern Recognition*, 2013. 2, 6

[27] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010. 1, 3

[28] B. Wang, G. Wang, K. Luk Chan, and L. Wang. Tracklet association with online target-specific metric learning. In *Conference on Computer Vision and Pattern Recognition*, 2014. 1

[29] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *International Conference on Computer Vision*, 2007. 2

[30] F. Xiong, M. Gou, O. I. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, 2014. 2, 4, 6, 7

[31] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *European Conference on Computer Vision*, 2014. 2

[32] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. 3

[33] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *International Conference on Computer Vision*, 2013. 2

[34] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Computer Vision and Pattern Recognition*, 2013. 2

[35] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Conference on Computer Vision and Pattern Recognition*, 2014. 4, 6

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, J. Bu, and Q. Tian. Scalable person re-identification: A benchmark. In *International Conference on Computer Vision*, 2015. 2, 5, 6

[37] W. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Trans. Pattern Anal. Mach. Intell*, 35(3):653–668, 2013. 2, 3

[38] W. Zheng, S. Gong, and T. Xiang. Group association: Assisting re-identification by visual context. In *Person Re-Identification*, pages 183–201. 2014. 1