Metadata of the article that will be visualized in OnlineFirst

Please note: Images will appear in color online but will be printed in black and white.

ArticleTitle	Exemplar-Guided Similarity Learning on Polynomial Kernel Feature Map for Person Re-identification
Article Sub-Title	
Article CopyRight	Springer Science+Business Media New York (This will be the copyright line in the final PDF)
Journal Name	International Journal of Computer Vision

Corresponding Author	Family Name	**Yuan**
	Particle	
	Given Name	**Zejian**
	Suffix	
	Division	
	Organization	Xi’an Jiaotong University
	Address	Xian Shi, China
	Phone	
	Fax	
	Email	yuan.ze.jian@mail.xjtu.edu.cn
	URL	
	ORCID	http://orcid.org/0000-0001-5548-3634
Corresponding Author	Family Name	**Wang**
	Particle	
	Given Name	**Jingdong**
	Suffix	
	Division	
	Organization	Microsoft Research Asia
	Address	Beijing, China
	Phone	
	Fax	
	Email	jingdw@microsoft.com
	URL	
	ORCID	
Author	Family Name	**Chen**
	Particle	
	Given Name	**Dapeng**
	Suffix	
	Division	
	Organization	Xi’an Jiaotong University
	Address	Xian Shi, China
	Phone	
	Fax	
	Email	dapengchen@xjtu.edu.cn

	URL	
	ORCID	
Author	Family Name	**Chen**
	Particle	
	Given Name	**Badong**
	Suffix	
	Division	
	Organization	Xi’an Jiaotong University
	Address	Xian Shi, China
	Phone	
	Fax	
	Email	chenbd@mail.xjtu.edu.cn
	URL	
	ORCID	
Author	Family Name	**Hua**
	Particle	
	Given Name	**Gang**
	Suffix	
	Division	
	Organization	Microsoft Research Asia
	Address	Beijing, China
	Phone	
	Fax	
	Email	ghua@stevens.edu
	URL	
	ORCID	
Author	Family Name	**Zheng**
	Particle	
	Given Name	**Nanning**
	Suffix	
	Division	
	Organization	Xi’an Jiaotong University
	Address	Xian Shi, China
	Phone	
	Fax	
	Email	nnzheng@mail.xjtu.edu.cn
	URL	
	ORCID	

Schedule	Received	10 September 2015
	Revised	
	Accepted	20 January 2017

Abstract

Person re-identification is a crucial problem for video surveillance, aiming to discover the correct matches for a probe person image from a set of gallery person images. To directly describe the image pair, we present a novel organization of polynomial kernel feature map in a high dimensional feature space to break down the variability of positive person pairs. An exemplar-guided similarity function is built on the map, which consists of multiple sub-functions. Each sub-function is associated with an “exemplar” image being responsible for a particular type of image pair, thus excels at separating the persons with similar appearance. We formulate a unified learning problem including a relaxed loss term as well as two kinds of regularization strategies particularly designed for the feature map. The corresponding optimization algorithm jointly optimizes the coefficients of all the sub-functions and selects the proper exemplars for a better discrimination. The proposed method is extensively evaluated on six public datasets, where we thoroughly analyze the contribution of each component and verify the generalizability of our approach by cross-dataset experiments. Results show that the new method can achieve consistent improvements over state-of-the-art methods.

Keywords (separated by '-')	Explicit polynomial kernel feature map - Exemplar-guided similarity function - Multiple visual cues - Similarity learning - Person re-identification
Footnote Information	Communicated by Zhouchen Lin.

Int J Comput Vis
DOI 10.1007/s11263-017-0991-0

Exemplar-Guided Similarity Learning on Polynomial Kernel Feature Map for Person Re-identification

Dapeng Chen[1] · Zejian Yuan[1] · Jingdong Wang[2] · Badong Chen[1] · Gang Hua[2] · Nanning Zheng[1]

Received: 10 September 2015 / Accepted: 20 January 2017

Abstract Person re-identification is a crucial problem for video surveillance, aiming to discover the correct matches for a probe person image from a set of gallery person images. To directly describe the image pair, we present a novel organization of polynomial kernel feature map in a high dimensional feature space to break down the variability of positive person pairs. An exemplar-guided similarity function is built on the map, which consists of multiple sub-functions. Each sub-function is associated with an "exemplar" image being responsible for a particular type of image pair, thus excels at separating the persons with similar appearance. We formulate a unified learning problem including a relaxed loss term as well as two kinds of regularization strategies particularly designed for the feature map. The corresponding optimization algorithm jointly optimizes the coefficients of all the sub-functions and selects the proper exemplars for a better discrimination. The proposed method is extensively evaluated on six public datasets, where we thoroughly analyze the contribution of each component and verify the generalizability of our approach by cross-dataset experiments. Results show that the new method can achieve consistent improvements over state-of-the-art methods.

Keywords Explicit polynomial kernel feature map · Exemplar-guided similarity function · Multiple visual cues · Similarity learning · Person re-identification

1 Introduction

Person re-identification refers to a task of associating the persons through different camera views located at different physical sites. It serves as a crucial step for video surveillance and has a wide range of applications for public security including cross-camera people tracking or behavior analysis. In real scenarios, the camera views are often significantly disjoint and the transition time between two cameras may vary greatly, making the temporal information insufficient to approach the problem. So far a lot of efforts have been devoted to investigating the solution through human appearance modeling (e.g. Li et al. 2013; Mignon and Jurie 2012; Köstinger et al. 2012; Liao et al. 2015; Wu et al. 2015).

Current solutions to re-identification generally utilize the person images cropped or detected from recorded videos. Nevertheless, finding the images of the target person from a large gallery set remains very challenging. The visual features, mainly dependent on clothes, can be very similar for different persons, leading to inherent ambiguity of the re-identification problem. Meanwhile, the features of a same person can appear quite different, often because of the changes of poses, view angles, illumination conditions or surrounding environments. Existing work tackle this problem from two paths: one relies on distinctive and robust visual

Communicated by Zhouchen Lin.

✉ Zejian Yuan
yuan.ze.jian@mail.xjtu.edu.cn

✉ Jingdong Wang
jingdw@microsoft.com

Dapeng Chen
dapengchen@xjtu.edu.cn

Badong Chen
chenbd@mail.xjtu.edu.cn

Gang Hua
ghua@stevens.edu

Nanning Zheng
nnzheng@mail.xjtu.edu.cn

[1] Xi'an Jiaotong University, Xian Shi, China

[2] Microsoft Research Asia, Beijing, China

descriptors that is invariant to inter-camera variations, the other learns a similarity measurement to separate the matched pairs from unmatched ones. Despite great progress, how to describe the matching between two images and how to distinguish different persons with similar appearance still remain open problems.

Given two image descriptors $\mathbf{x}_a$ and $\mathbf{x}_b$, we perform similarity learning over the a high dimensional feature space, which is induced by the feature map $\phi(\mathbf{x}_a, \mathbf{x}_b)$, a regularized form of the second-order explicit polynomial kernel feature map of the concatenated descriptor $[\mathbf{x}_a^\top, \mathbf{x}_b^\top]^\top$. $\phi(\mathbf{x}_a, \mathbf{x}_b)$ describes the matching between $\mathbf{x}_a$ and $\mathbf{x}_b$, and is composed of two parts related to Mahalanobis distance and bilinear similarity, respectively. The basic similarity function $f(\mathbf{x}_a, \mathbf{x}_b)$ is linear over $\phi(\mathbf{x}_a, \mathbf{x}_b)$. Taking advantages of the polynomial kernel feature map, $f(\mathbf{x}_a, \mathbf{x}_b)$ is robust to pose variation and illumination change, meanwhile discriminative enough to separate one person from another. The similarity function can be extended by incorporating multiple feature maps, each describes the matching of the image pair using a different visual cue. Learning these feature maps together exploits effective features from different information sources simultaneously, reducing the risk of mismatching.

However, a single linear similarity function is difficult to account for all the persons, especially when the target needs to be distinguished from the persons with similar appearance. We therefore propose an exemplar-guided similarity function $s(\mathbf{x}_a, \mathbf{x}_b)$ by extending $f(\mathbf{x}_a, \mathbf{x}_b)$ with a mixture of linear sub-functions. Each sub-function is associated with an exemplar selected from the training set, and is weighted by the affinity between the probe image and the exemplar image. In this way, persons with similar appearance tend to rely on a same sub-function, which is trained to specialize in distinguishing the matched images pair with that kind of appearance.

For the similarity learning, we follow the learning-to-rank methodology with the goal of ranking the image pair about the same person before the image pair about different persons. To accelerate the training process, the loss term averages the constraints with respect to a same probe image, which is relaxed from the one that imposes all the triplet constraints. Two specialized regularization strategies are imposed regarding the structure of the feature map. Firstly, parts of the coefficients are required to be negative semi-definite matrices due to the connection between the explicit polynomial kernel feature map and the Mahalanobis distance. Secondly, the coefficients of each sub-matrix are imposed by group sparse for more effective feature selection. We optimize the coefficients via an iterative algorithm consisting of two complementary steps: one is to select exemplars for better discrimination, and the other is to learn the coefficients of all the linear functions. The two steps alteratively decrease the value of objective function, and can finally achieve the convergence.

We validate our method on six publicly available datasets with different configurations in different scenarios. By constructing a number of variant methods, we perform in-depth experiments to investigate the nature of the performance and to analyze the impact of the choice of parameters. Our method achieves high accuracy within each dataset, and shows good generalization ability in the experiments where the training data and test data are from different datasets.

In summary, the main contributions are:

1. A novel organization of the explicit polynomial kernel feature map is proposed to describe the image pair in a high dimensional feature space. The organization takes the strength of both Mahalanobis distance and bilinear similarity, and is convenient to exploit the effectiveness of multiple visual cues.
2. An exemplar-guided similarity function is built upon the feature map. It employs the exemplars to break down the variability of the positive feature maps, and softly combines the similarity in each decomposed subspace weighted by the affinity between probe image and the exemplar images.
3. We present a unified learning framework. The objective function consists of a relaxed loss term that ensures optimization efficiency and two specialized regularization terms particularly designed for the feature map. The optimization algorithm decreases the objective by alternatively solving the coefficients and selecting the exemplars, guaranteed to achieve convergence.

Parts of this work have been appeared in Chen et al. (2015). Compared with the previous work, this paper has made significant changes in similarity function as well as learning strategy. Both similarity functions compromise a mixture of linear sub-functions, the previous one is with a latent formulation to select the most discriminative sub-function, while the current one weights each sub-function regarding an exemplar image. For the learning strategy, the current relaxed loss term largely reduces the number of loss constraints of the previous one, accelerating the learning procedure. Group sparse rather than the L_1 sparse is imposed for the coefficients of each sub-function, in order to better consider the internal structure of the polynomial kernel feature map. We additionally study the effectiveness of employing multiple feature maps. A substantial number of explanations, analyses and experiments are included in this paper, in order to investigate broader aspects of our method.

2 Related Work

The problem of re-identifying persons across multiple non-overlap cameras has received increasing attentions. To address the problem, existing work mainly focus on feature extraction and similarity measuring. We refer the reader to the work of Gong et al. (2014) for a comprehensive survey, and briefly summarize the most related work in this section.

2.1 Feature Descriptors

A lot of efforts have been devoted to seeking very stable and discriminative feature descriptors.

To obtain the robust features, Farenzena et al. (2010) considered the symmetric and asymmetric prior of human body to extract the local features from different regions. Xu et al. (2013) and Cheng et al. (2011) employed pre-learned part models organized by pictorial structure to localize the body part more accurately. These methods build spatial correspondence before matching and alleviate the influence of spatial misalignment. Ma et al. (2012a) presented the person image via covariance descriptors, making the representation be less affected by illumination changes and background variations. To obtain the distinctive features, Bak et al. (2010) extracted the covariance descriptor (Tuzel et al. 2006) over the interested region; Zhao et al. (2013) exploited unsupervised salience feature to distinguish the correct matched person from others; Zheng et al. (2009) extended the traditional person re-identification to group re-identification, and they have shown that using group contextual descriptors can help to improve the performance of re-identification .

Meanwhile, efficient region descriptors are usually adopted by methods that focus on similarity learning. These descriptors are built in a more straightforward way: most of them extract color and texture histograms from pre-defined image regions in a “block” shape or “strip” shape (e.g. Xiong et al. 2014; Mignon and Jurie 2012; Köstinger et al. 2012; Liao et al. 2015; Zheng et al. 2013; Pedagadi et al. 2013; Chen et al. 2016); several researches further process the extracted descriptors, including methods that apply max pooling (Liao et al. 2015), learn mid-level filters (Zhao et al. 2014) and encode the descriptors to form high level representation (Ma et al. 2012b; Li et al. 2013).

The proposed explicit polynomial kernel feature map is based upon the efficient region descriptors. Unlike other descriptors that characterize a single image, the feature map describes the relationship within an image pair in a high dimensional space. It takes the linear forms of Mahalanobis distance and a bilinear similarity, flexible to incorporate other forms of polynomial kernel feature map. From the feature map, we can exploit the intra-person invariance directly, facilitating the employment of the mixture model and the multiple visual cues.

2.2 Similarity Measurement

Methods stressing on feature design often adopt off-the-shelf distance metrics, such as Euclidean distance (e.g. Farenzena et al. 2010), Bhattacharyya distance (e.g. Cheng et al. 2011), and covariance distance (e.g. Bak et al. 2010; Ma et al. 2012a). Meanwhile, many studies aim to learn more effective similarity measurements specialized for person re-identification.

Similarity measurement can be learned by selecting the features in the original feature space. Gray et al. (2007) utilized Adaboost to select a subset of effective features. Schwartz and Davis (2009) assigned weights to features with partial least squares. Prosser et al. (2010) stressed the importance of ranking and described the triplet relations between a positive pair and a negative pair. They didn’t stick to separating all the correctly matched pairs from all the incorrectly matched pairs, but were interested in the rank of these scores that reflects how likely they match to a given probe image. However, it is computationally expensive to optimize all the triplet constraints. We provide a relaxed loss term that makes the constraint number be only related to the number of person identities, largely accelerating the training process.

Metric learning has been widely developed for person re-identification (e.g. Li et al. 2013; Mignon and Jurie 2012; Köstinger et al. 2012; Liao et al. 2015; Li and Wang 2013; Li et al. 2015). The formulations of most metrics originate from the Mahalanobis distance, trying to obtained a transformed feature descriptor that is robust to intra-person variation. Among them, Hirzer et al. (2012) relaxed the positive semi-definite(PSD) constraint required in Mahalanobis metric learning, and obtained a simplified formulation with reasonable effectiveness. Köstinger et al. (2012) proposed an efficient metric computation method motivated by the log-likelihood ratio test of two Gaussian distributions. Li et al. (2013) proposed Locally-Adaptive Decision Function (LADF) to jointly model a distance metric and a locally adaptive threshold rule, extending the form of metric learning. Our method is comparable to LADF: both of them represent linear learning over explicit polynomial kernel feature map, but the organization of the feature map and the learning strategy are quite different. Our feature map directly encodes the relationship between two images, yielding better performance.

Employing a mixture of similarity functions began to show its effectiveness. For example, Li and Wang (2013) learned a mixture of experts, where samples were softly distributed into different experts (similarity functions) via a gating function. Ma et al. (2014) divided the data according to the camera label, and utilized multi-task learning to obtain the distance metric for each camera pair. The previous version of our method (Chen et al. 2015) combined multiple similarity functions in a latent fashion, which discovers different matching patterns to increase the model's discriminability. In this paper, the mixture model has been formulated in a different manner. Each sub-function is weighted by the affinity between the probe image and a selected exemplar, thus different persons with similar appearance tend to rely on a same similarity function, which excels at handling that kind of appearance.

Making use of multiple visual features can further improve the performance. For example, Farenzena et al. (2010) combined the matching score of three distinguished feature descriptors by pre-defined weights. Zhao et al. (2014) directly fused the results of Middle Level Filter with the one of LADF (Li et al. 2013). Paisitkriangkrai et al. (2015) attempted to find the optimal linear weights of multiple distance functions in structure learning framework. All of them firstly learned the similarity metrics independently, and then combined their results together. In contrast to the two-stage approaches, our method takes advantages of the explicit form of the polynomial kernel feature map, which is flexible to be extended with multiple visual cues by encoding them into multiple feature maps. We cast the feature maps into a unified similarity learning framework, achieving a better consistency among different visual cues.

3 Linear Similarity Function on Explicit Polynomial Kernel Feature Map

Given the descriptors of two images $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^d$, we aim to measure the similarity between $\mathbf{x}_a$ and $\mathbf{x}_b$ to determine whether they belong to a same person. Let $\mathbf{z}^\top = [\mathbf{x}_a^\top, \mathbf{x}_b^\top]^\top$ be the vector representing an image pair, a desirable similarity function should be able to separate positive pairs (two images belong to a same person) from negative pairs (two images belong to different persons) despite the influence from pose variation and illumination change. However, vector $\mathbf{z}$ itself is not sufficiently separable, we therefore project $\mathbf{z}$ into a high dimensional feature space. In particular, the projection is via a feature map $\varphi(\mathbf{z})$, which is induced by a second-order polynomial kernel $k(\mathbf{z}, \mathbf{z}') = (\mathbf{z}^\top \mathbf{z}')^2$ and has an explicit form, i.e.,

$$\begin{aligned}\varphi(\mathbf{z}) &= \varphi(\mathbf{x}_a, \mathbf{x}_b)\\ &= [\mathrm{vec}(\mathbf{x}_a\mathbf{x}_a^\top)^\top, \mathrm{vec}(\mathbf{x}_b\mathbf{x}_b^\top)^\top, \mathrm{vec}(\mathbf{x}_a\mathbf{x}_b^\top)^\top, \mathrm{vec}(\mathbf{x}_b\mathbf{x}_a^\top)^\top]^\top.\end{aligned}$$

In $\varphi(\mathbf{x}_a, \mathbf{x}_b)$, the terms $\mathbf{x}_a\mathbf{x}_a^\top$, $\mathbf{x}_b\mathbf{x}_b^\top$, $\mathbf{x}_a\mathbf{x}_b^\top$ and $\mathbf{x}_b\mathbf{x}_a^\top$ are outer products of image descriptors, encoding the relationship within and between the two images.

However, directly learning linear classifiers over the feature map would easily lead to overfitting. To achieve more effective similarity learning, we need to perform reorganization and regularization to $\varphi(\mathbf{x}_a, \mathbf{x}_b)$. Before that, it is necessary to revisit LADF (Li et al. 2013), another similarity function also built upon second-order polynomial kernel feature map.

3.1 A Retrospect of LADF

LADF extends the Mahalanobis distance with a local decision rule, aiming to increase the adaptivity to the local structure of data. The formulated similarity function is linear over the feature map $\hat{\phi}(\mathbf{x}_a, \mathbf{x}_b)$:

$$\hat{f}(\mathbf{x}_a, \mathbf{x}_b) = \hat{\phi}(\mathbf{x}_a, \mathbf{x}_b) \cdot \zeta, \tag{1}$$

where

$$\hat{\phi}(\mathbf{x}_a, \mathbf{x}_b) = \begin{bmatrix} \frac{1}{2}\mathrm{vec}(\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top) \\ \frac{1}{2}\mathrm{vec}(\mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top) \\ \mathbf{x}_a + \mathbf{x}_b \end{bmatrix}, \zeta = \begin{bmatrix} \mathrm{vec}(\hat{\mathbf{A}}) \\ \mathrm{vec}(\hat{\mathbf{B}}) \\ \hat{\mathbf{c}} \end{bmatrix}.$$

$\hat{\phi}(\mathbf{x}_a, \mathbf{x}_b)$ is a re-organized polynomial kernel feature map considering both first-order and second-order terms. Due to the re-organization, $\mathbf{x}_a\mathbf{x}_a^\top$ and $\mathbf{x}_b\mathbf{x}_b^\top$ share the coefficients $\hat{\mathbf{A}}$, while $\mathbf{x}_a\mathbf{x}_b^\top$ and $\mathbf{x}_b\mathbf{x}_a^\top$ share the coefficient $\hat{\mathbf{B}}$. Such property ensures the similarity function is symmetric, i.e., $\hat{f}(\mathbf{x}_a, \mathbf{x}_b) = \hat{f}(\mathbf{x}_b, \mathbf{x}_a)$. $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are regularized to be negative semi-definite (NSD) and positive semi-definite (PSD) respectively, i.e., $\hat{\mathbf{A}} \in \mathbb{S}_-^d$ and $\hat{\mathbf{B}} \in \mathbb{S}_+^d$, which fits the semi-definite requirement for precision matrix of Mahalanobis distance and is found to be useful in practice.

Although LADF is quite effective, the advantages of explicit polynomial kernel feature maps have not been fully exploited. By transforming $\hat{f}(\mathbf{x}_a, \mathbf{x}_b)$ into its equivalent form:

$$\frac{1}{2}\left(\mathbf{x}_a^\top\hat{\mathbf{A}}\mathbf{x}_a + \mathbf{x}_b^\top\hat{\mathbf{A}}\mathbf{x}_b + \mathbf{x}_a^\top\hat{\mathbf{B}}\mathbf{x}_b + \mathbf{x}_b^\top\hat{\mathbf{B}}\mathbf{x}_a\right) + \hat{\mathbf{c}}^\top(\mathbf{x}_a + \mathbf{x}_b), \tag{2}$$

we find the ranking of similarity score with respect to a probe image $\mathbf{x}_a$ is determined by $\mathbf{x}_a^\top\hat{\mathbf{B}}\mathbf{x}_b + \mathbf{x}_b^\top\hat{\mathbf{B}}\mathbf{x}_a + \mathbf{x}_b^\top\hat{\mathbf{A}}\mathbf{x}_b + \hat{\mathbf{c}}^\top\mathbf{x}_b$. Among them, the score of $\mathbf{x}_b^\top\hat{\mathbf{A}}\mathbf{x}_b + \hat{\mathbf{c}}^\top\mathbf{x}_b$ only depends on the appearance of $\mathbf{x}_b$, thus it hardly tells whether $\mathbf{x}_a$ and $\mathbf{x}_b$ belong to a same identity or notk. Thus the remaining part $\mathbf{x}_a^\top\hat{\mathbf{B}}\mathbf{x}_b + \mathbf{x}_b^\top\hat{\mathbf{B}}\mathbf{x}_a$, related to the bilinear similarity, contributes most to the effectiveness of LADF.

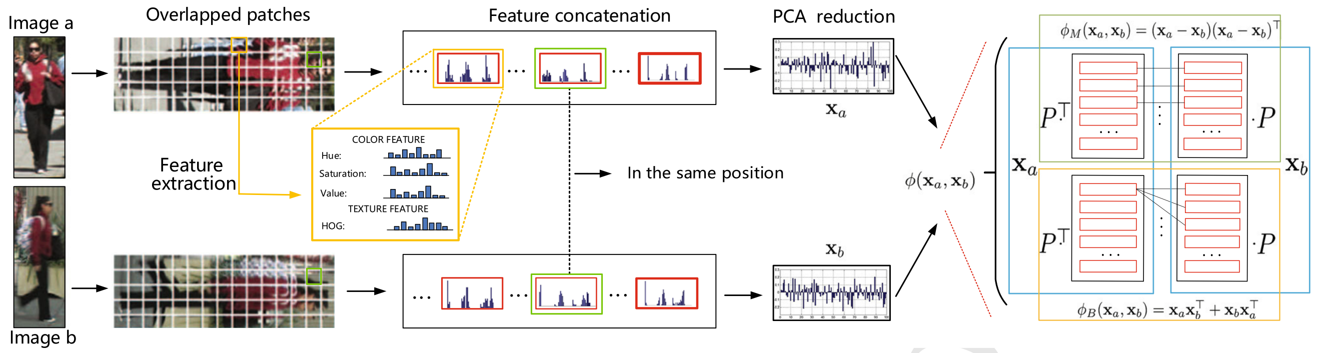
Image a
Overlapped patches
Feature concatenation
PCA reduction
Feature extraction
COLOR FEATURE
Hue:
Saturation:
Value:
TEXTURE FEATURE
HOG:
In the same position
$\mathbf{x}_a$
$\mathbf{x}_b$
$\phi(\mathbf{x}_a, \mathbf{x}_b)$
$\phi_M(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top$
$P^\top$
$\cdot P$
$\phi_B(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top$
Image b

Fig. 1 Feature map generation. We extract color and texture histogram from overlapped patches of a person image, and concatenate these descriptors together following by PCA. For the feature map, one part is related to Mahalanobis distance capturing the correlation of feature difference; the other part corresponds to a bilinear similarity that builds the connection between each patch in one image and all the patches in the other image

3.2 New Organization of the Feature Map

To better exploit the polynomial kernel feature map, we propose a new organization, where each separated part of the feature map directly encodes the relationship between the two images rather than describe an single image. In this way, the learned similarity is more determined by the appearance of both images, reducing the bias caused by a single image. In particular, our polynomial kernel feature map stems from two kinds of inter-image relationships.

We firstly consider Mahalanobis distance, a widely used, quite effective distance metric in the form of:

$$d_M(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{M}(\mathbf{x}_a - \mathbf{x}_b), \tag{3}$$

where $\mathbf{M} \in \mathbb{R}^{d\times d}$ is the precision matrix constrained to be positive semi-definite. According to the cyclic property of the trace, i.e., $(\mathbf{x}_a-\mathbf{x}_b)^\top \mathbf{M}(\mathbf{x}_a-\mathbf{x}_b) = \mathrm{tr}\big(\mathbf{M}(\mathbf{x}_a-\mathbf{x}_b)(\mathbf{x}_a-\mathbf{x}_b)^\top\big)$, $d_M(\mathbf{x}_a, \mathbf{x}_b)$ can be represented via Frobenius inner product: $\langle \mathbf{M}, (\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top\rangle_F$, which is a linear function over a feature map:

$$\phi_M(\mathbf{x}_a, \mathbf{x}_b) = (\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top. \tag{4}$$

As $(\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top = \mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top - \mathbf{x}_a\mathbf{x}_b^\top - \mathbf{x}_b\mathbf{x}_a^\top$, $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$ is re-organized by the elements in the second-order polynomial kernel feature map. In a similar manner, the bilinear similarity,

$$s_B(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{x}_a^\top \mathbf{B}\mathbf{x}_b + \mathbf{x}_b^\top \mathbf{B}\mathbf{x}_a, \tag{5}$$

is equivalent to $\mathbf{x}_a^\top \mathbf{B}\mathbf{x}_b + \mathbf{x}_b^\top \mathbf{B}\mathbf{x}_a = \langle \mathbf{B}, \mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top\rangle_F$, where $\mathbf{B}$ is the coefficients need to be learned. $s_B(\mathbf{x}_a, \mathbf{x}_b)$ is a linear function over the feature map:

$$\phi_B(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top, \tag{6}$$

which is another organization of $\varphi(\mathbf{x}_a, \mathbf{x}_b)$.

We concatenate $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B(\mathbf{x}_a, \mathbf{x}_b)$ together, yielding:

$$\phi(\mathbf{x}_a, \mathbf{x}_b) = [\phi_M(\mathbf{x}_a, \mathbf{x}_b), \phi_B(\mathbf{x}_a, \mathbf{x}_b)], \tag{7}$$

where $\phi(\mathbf{x}_a, \mathbf{x}_b) \in \mathbb{R}^{d\times 2d}$. In general, $\phi(\mathbf{x}_a, \mathbf{x}_b)$ is not restricted to the listed two kinds of feature maps, and could be expanded with broader forms (Fig. 1).

3.3 Basic Similarity Function Over $\boldsymbol{\phi}(\mathbf{x}_a, \mathbf{x}_b)$

The basic similarity function $f(\mathbf{x}_a, \mathbf{x}_b)$ is linear over $\phi(\mathbf{x}_a, \mathbf{x}_b)$, i.e.,

$$f(\mathbf{x}_a, \mathbf{x}_b; \mathbf{W}) = \langle \phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}\rangle_F, \tag{8}$$

where $\langle \cdot, \cdot\rangle_F$ is the Frobenius inner product and $\mathbf{W} \in \mathbb{R}^{d\times 2d}$ is the coefficient matrix. $f(\mathbf{x}_a, \mathbf{x}_b)$ consistently takes advantage of both Mahalanobis distance and bilinear similarity. With respect to Eq. (7), we decompose $\mathbf{W}$ as $[\mathbf{W}_M, \mathbf{W}_B]$, where $\mathbf{W}_M$ and $\mathbf{W}_B$ are the coefficients for $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B(\mathbf{x}_a, \mathbf{x}_b)$, respectively. $f(\mathbf{x}_a, \mathbf{x}_b)$ is equivalent to:

$$(\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{W}_M(\mathbf{x}_a - \mathbf{x}_b) + \mathbf{x}_a^\top \mathbf{W}_B\mathbf{x}_b + \mathbf{x}_b^\top \mathbf{W}_B\mathbf{x}_a. \tag{9}$$

We impose two kinds of regularization over $\mathbf{W}_M$ and $\mathbf{W}_B$. One originates from the fact that the precision matrix of Mahalanobis distance is usually PSD. As smaller Mahalanobis distance indicates more similarity, the term $(\mathbf{x}_a - \mathbf{x}_b)^\top \mathbf{W}_M(\mathbf{x}_a - \mathbf{x}_b)$ corresponds to $-d_M(\mathbf{x}_a, \mathbf{x}_b)$, thus $\mathbf{W}_M$ should be NSD, i.e.,

$$\mathbf{W}_M \in \mathbb{S}^d_-. \tag{10}$$

The other regards the structure of $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_B(\mathbf{x}_a, \mathbf{x}_b)$. Specifically, both feature maps are generated by the outer product of two vectors (e.g., descriptor difference, two sample descriptors). If some entries of the vectors are not encoding the information beneficial to matching, the elements of the feature map in corresponding columns and rows tend to be less effective. The assumption indicates that the effective elements in polynomial kernel feature map would appear in group, thus we regularize $\mathbf{W}_M$ and $\mathbf{W}_B$ by the mixed-norm:

$$\|\mathbf{W}\|_{2,1} := \sum_{i=1}^{d} \|\mathbf{W}_{i\cdot}\|_2, \tag{11}$$

where $\mathbf{W}_{i\cdot}$ is a d-dimensional vector taken from the i-th row of matrix $\mathbf{W}$.

3.4 Incorporating with Multiple Visual Cues

As an image can be described by multiple visual cues $\{\mathbf{x}_n^1, \ldots, \mathbf{x}_n^c, \ldots, \mathbf{x}_n^C\}$, the matching of two images can be described in different aspects of view by forming multiple feature maps $\{\phi(\mathbf{x}_a^1, \mathbf{x}_b^1), \ldots, \phi(\mathbf{x}_a^c, \mathbf{x}_b^c), \ldots, \phi(\mathbf{x}_a^C, \mathbf{x}_b^C)\}$. To reduce the risk of mismatching, the similarity learning can be performed over all the feature maps simultaneously, to exert their complementary strengths. Let $\mathbf{W}^c$ be the coefficient of $\phi(\mathbf{x}_a^c, \mathbf{x}_b^c)$, the similarity function with multiple visual cues is:

$$f'\left(\mathbf{x}_a, \mathbf{x}_b; \{\mathbf{W}^c\}_{c=1}^C\right) = \sum_{c=1}^{C} \langle \phi^c(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^c \rangle_F, \tag{12}$$

where $\phi^c(\mathbf{x}_a, \mathbf{x}_b) = \phi(\mathbf{x}_a^c, \mathbf{x}_b^c)$.

3.5 Visual Cues

The visual cues are obtained by applying the PCA over the spatially arranged patch descriptors.

We divide the image into a collection of overlapped patches, and extract different color and texture histograms for different visual cues in each patch. The image descriptor of one visual cue is obtained by concatenating the corresponding histograms of all the patches, forming a d'-dimensional (around $40K$) descriptor. However, it is computationally infeasible to directly use such descriptors to generate the feature map, as the number of necessary parameter is at the order of $O(d'^2)$. Because of this, PCA is employed for dimensionality reduction, which tries to find m principle components to reflect the intrinsic variability of the data. However, when the feature dimension d' is much higher than the training sample number l, there may exist concerns about the feasibility of PCA on both scalability and effectiveness.

3.5.1 Feasibility

For scalability, the typical algorithm for finding the eigenvectors of a $d' \times d'$ matrix has a computational cost that scales like $O(d'^3)$, while PCA in dual form makes a transformation and solves the eigenvector problem in a lower dimensional space with computational cost $O(l^3)$.

We also defend the rationality of training PCA over only hundreds of person images. Whether l is sufficient for PCA does not depend on d', but depends on the number of intrinsic variabilities m (Kjems et al. 2000). In our case, m is much less than d'. This is because: (1) great redundancy exists in the spatially arranged patch descriptors, e.g., the person clothes are usually with coherent color or textures, thus the descriptors of one patch may strongly correlate with the ones of other patches. (2) The variation among the training images is restricted, e.g., a lot of dimensions in the original descriptors characterize the shape of the person, but the shape only has several prototypes corresponding to different poses. Besides, despite m is unknown, the optimal reduced dimensionality obtained by cross-validation is much less than the training sample number l in our experiment, which indicates the great possibility that m is less than l.

3.5.2 Limitations

Nevertheless, PCA is not a panacea for dimensionality reduction. It is hard to determine which principle component encodes the interested structures for a specific task and which one just reflects the noise signals. In general, d eigenvectors with largest eigenvalues are preserved and are supposed to be more significant to characterize the samples than those with smaller eigenvalues. Besides, it is unclear how many intrinsic variabilities need to be estimated for a specific task, the reduced dimensionality d is usually determined empirically in practice. Overall, PCA is a linear dimensionality reduction method that is agnostic to the source of the data, and it will be less effective when the data follow non-linear distribution.

4 Exemplar-Guided Similarity Learning

4.1 Exemplar-Guided Similarity Function on Explicit Polynomial Kernel Feature Map

To further increase the discriminative ability, we extend $f(\mathbf{x}_a, \mathbf{x}_b)$ with a mixture of linear sub-functions, where each sub-function is expected to handle a particular appearance type. The main idea is that, instead of seeking a single sophisticate function which can separate the entire variability space of the positive pairs, we break down the variability of positive pair into manageable pieces, where each piece can be better separated from negative pairs by a sub-function.

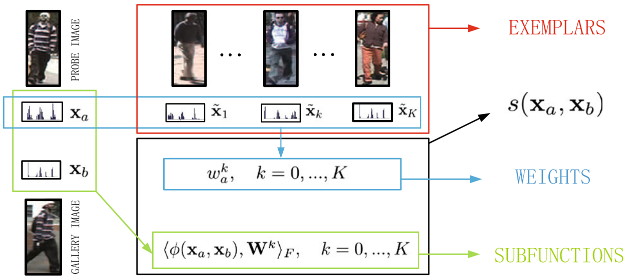
PROBE IMAGE
EXEMPLARS
$\mathbf{x}_a$
$\tilde{\mathbf{x}}_1$
$\tilde{\mathbf{x}}_k$
$\tilde{\mathbf{x}}_K$
$s(\mathbf{x}_a, \mathbf{x}_b)$
$\mathbf{x}_b$
$w_a^k, \quad k = 0, ..., K$
WEIGHTS
GALLERY IMAGE
$\langle \phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^k \rangle_F, \quad k = 0, ..., K$
SUBFUNCTIONS

Fig. 2 The flowchart of our similarity function. The final function is the sum of weighted sub-functions, where the weights are determined by the probe image and exemplar images

Let $\mathbf{x}_a$ and $\mathbf{x}_b$ be the probe image and gallery image descriptors, the exemplar-guided similarity function is:

$$s(\mathbf{x}_a, \mathbf{x}_b) = f(\mathbf{x}_a, \mathbf{x}_b) + \sum_{k=1}^{K} w_a^k f^k(\mathbf{x}_a, \mathbf{x}_b), \qquad (13)$$

$$= \sum_{k=0}^{K} w_a^k \langle \phi(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^k \rangle_F,$$

where $w_a^0 = 1$ is the weight associated with the shared function $f(\mathbf{x}_a, \mathbf{x}_b)$, and $\{w_a^k\}_{k=1}^K$ are the weights for sub-functions $\{f^k(\mathbf{x}_a, \mathbf{x}_b)\}_{k=1}^K$. The weight w_a^k reflects the affinity between the probe descriptor $\mathbf{x}_a$ and the k-th exemplar descriptor $\tilde{\mathbf{x}}_k$, which is computed as:

$$w_a^k = \exp^{\beta \langle \mathbf{x}_a, \tilde{\mathbf{x}}_k \rangle}. \qquad (14)$$

During training, all the exemplars are selected from the probe set, and each exemplar represents a particular appearance type of the probe images. In this way, given probe images with similar appearance, the overall similarities tend to depend heavier on a same sub-function, which excels at separating the positive pairs with that kind of appearance (Fig. 2).

The exemplar-guided similarity function can also be expanded with multiple visual cues. In particular, each visual cue has K exemplars, and the similarity function is given by:

$$s'(\mathbf{x}_a, \mathbf{x}_b; \mathcal{W}, \mathcal{E}) = \sum_{k=0}^{K} \sum_{c=1}^{C} w_a^{c,k} \langle \phi^c(\mathbf{x}_a, \mathbf{x}_b), \mathbf{W}^{c,k} \rangle_F. \qquad (15)$$

In Eq. (15), $\{w_a^{c,k}\}_{k=0}^K$ are the weights of linear functions, they are computed from $\mathbf{x}_a^c$ and $\tilde{\mathbf{x}}_k^c$ according to Eq. (14), where $\mathbf{x}_a^c$ and $\tilde{\mathbf{x}}_k^c$ are the probe image and exemplar image of the c-th visual cue. We collect all the coefficients and all the exemplars in set $\mathcal{W}$ and set $\mathcal{E}$, i.e,

$$\mathcal{W} = \{\mathbf{W}^{c,k} | c = 1, \ldots, C, k = 0, \ldots, K\}$$
$$\mathcal{E} = \{\tilde{\mathbf{x}}_k^c | c = 1, \ldots, C, k = 0, \ldots, K\}, \tilde{\mathbf{x}}_k^c \in \mathcal{X}_e^c. \qquad (16)$$

In particular, the exemplar $\tilde{\mathbf{x}}_k^c$ is selected from $\mathcal{X}_e^c$, which contains the exemplar candidates with the c-th visual cue. Without loss of generality, we simply set it as the probe images in the training set.

4.2 Learning for Person Re-identification

We aim to learn $\mathcal{W}, \mathcal{E}$ for $s'(\mathbf{x}_a, \mathbf{x}_b)$. The proposed learning strategy is also applicable to $s(\mathbf{x}_a, \mathbf{x}_b)$, $f'(\mathbf{x}_a, \mathbf{x}_b)$ and $f(\mathbf{x}_a, \mathbf{x}_b)$ as they are reduced versions of $s'(\mathbf{x}_a, \mathbf{x}_b)$.

4.2.1 Regularization Term

Set $\mathcal{W}$ contains $(K + 1) \times C$ different coefficient matrices. For each of them, i.e., $\mathbf{W}^{c,k}$, is decomposed as $[\mathbf{W}_M^{c,k}, \mathbf{W}_B^{c,k}]$. We impose two kinds of regularization over the coefficients according to Eqs. (10) and (11). The overall group sparse regularization term is:

$$R_1(\mathcal{W}) = \sum_{c=1}^{C} \sum_{k=0}^{K} \left(\|\mathbf{W}_M^{c,k}\|_{2,1} + \|\mathbf{W}_B^{c,k}\|_{2,1} \right). \qquad (17)$$

Let $\mathcal{C}$ be the set of the coefficients $\mathcal{W}$ that satisfy the NSD constraints:

$$\mathcal{C} = \{\mathcal{W} | \mathbf{W}_M^{c,k} \in \mathbb{S}_-^d, c = 1, \ldots, C, k = 0, \ldots, K\}. \qquad (18)$$

The overall semi-definite regularization term is:

$$R_1(\mathcal{W}) = \infty \delta[\mathcal{W} \notin \mathcal{C}], \qquad (19)$$

where $\delta[\cdot]$ is an indicator function which takes one if the argument is true and zero otherwise.

4.2.2 Relaxed Loss Term

The training data for person re-identification can be organized as follows. Given probe images $\{\mathbf{x}_n\}_{n=1}^N$, each image $\mathbf{x}_n$ is associated with two sets of gallery images: a positive set $\mathcal{X}_n^+$ composed of image descriptors about the same person with $\mathbf{x}_n$, and a negative set $\mathcal{X}_n^-$ composed of image descriptors about different persons.

As person re-identification is usually formulated as a ranking problem, it is natural to impose a triplet loss $\ell_{\text{triplet}}(\mathbf{x}_n, \mathbf{x}_i, \mathbf{x}_j) = [1 - (s'(\mathbf{x}_n, \mathbf{x}_i) - s'(\mathbf{x}_n, \mathbf{x}_j))]_+$, where $\mathbf{x}_i \in \mathcal{X}_n^+$ and $\mathbf{x}_j \in \mathcal{X}_n^-$. The triplet loss forces every positive pair to be scored higher than negative pairs at least by 1. However, as the number of triplets may be extremely large, it is costly to optimize over all the triplet constraints. We make a relaxation by comparing the average score of positive pairs and the average score of the negative pairs considering the same probe image, i.e.,

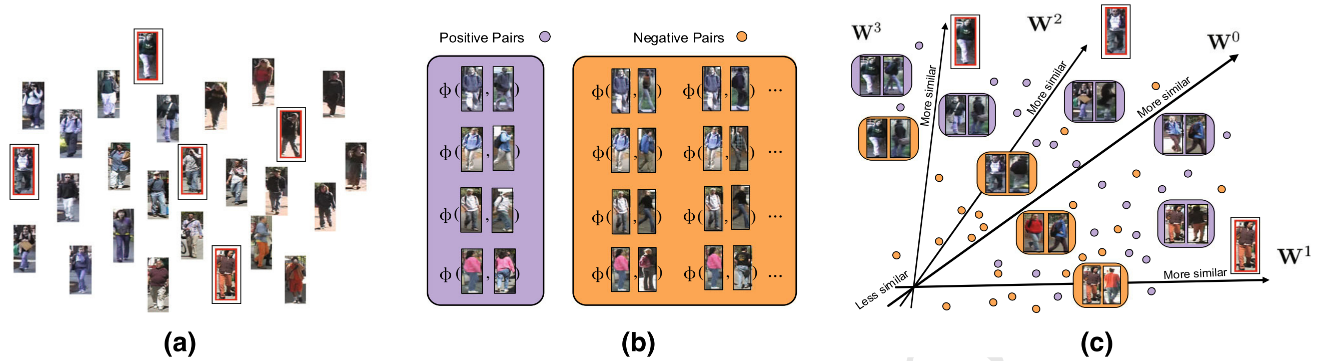

Fig. 3 The interpretation of our similarity learning. The learning includes two complementary tasks: one is to select the representative exemplars that facilitates the discrimination, the other is to learn the coefficients for a mixture of linear functions to distinguish the positive image pairs from the negative image pairs. Especially, each sub-function excels at discriminating different persons with similar appearance

$$\ell(\mathbf{x}_n) = \left[1 - \frac{\sum_{\mathbf{x}_i\in\mathcal{X}_n^+,\mathbf{x}_j\in\mathcal{X}_n^-}\left(s'(\mathbf{x}_n,\mathbf{x}_i) - s'(\mathbf{x}_n,\mathbf{x}_j)\right)}{|\mathcal{X}_n^+|\cdot|\mathcal{X}_n^-|}\right]_+ .$$

By defining

$$\psi^c(\mathbf{x}_n) = \frac{\sum_{\mathbf{x}_i\in\mathcal{X}_n^+}\sum_{\mathbf{x}_j\in\mathcal{X}_n^-}\left(\phi^c(\mathbf{x}_n,\mathbf{x}_i) - \phi^c(\mathbf{x}_n,\mathbf{x}_j)\right)}{|\mathcal{X}_n^+|\cdot|\mathcal{X}_n^-|},$$

$\ell(\mathbf{x}_n)$ is formulated as:

$$\ell(\mathbf{x}_n,\mathcal{W},\mathcal{E}) = \left[1 - \sum_{c=1}^{C}\sum_{k=0}^{K} w_n^{c,k}\left\langle \psi^c(\mathbf{x}_n), \mathbf{W}^{c,k}\right\rangle_F\right]_+ . \tag{20}$$

The whole loss term accumulates $\ell(\mathbf{x}_n;\mathcal{W},\mathcal{E})$ over all the probe images:

$$L(\mathcal{W},\mathcal{E}) = \frac{1}{N}\sum_{n=1}^{N}\ell(\mathbf{x}_n,\mathcal{W},\mathcal{E}). \tag{21}$$

4.2.3 Objective Function

With Eqs. (17) and (21), we train $\mathcal{W}$ and $\mathcal{E}$ by minimizing the objective function:

$$Q(\mathcal{W},\mathcal{E}) = L(\mathcal{W},\mathcal{E}) + \lambda R_1(\mathcal{W}) + R_2(\mathcal{W}). \tag{22}$$

Different coefficient matrices in $\mathcal{W}$ are regularized separately in $R_1(\mathcal{W})$ and $R_2(\mathcal{W})$, but they together contribute to the global similarity score in the loss term $L(\mathcal{W},\mathcal{E})$. In this way, different coefficient matrices are mutually influenced (Fig. 3) .

4.3 Discussion

We compare the exemplar-guided similarity learning with the latent similarity learning in our conference paper (Chen et al. 2015). The previous similarity function is given by:

$$\tilde{s}(\mathbf{x}_a,\mathbf{x}_b) = f(\mathbf{x}_a,\mathbf{x}_b) + \max_{h=1,\ldots,H} f^h(\mathbf{x}_a,\mathbf{x}_b). \tag{23}$$

Both $s(\mathbf{x}_a,\mathbf{x}_b)$ and $\tilde{s}(\mathbf{x}_a,\mathbf{x}_b)$ enhance $f(\mathbf{x}_a,\mathbf{x}_b)$ with a mixture of sub-functions, which facilitate the separation of the positive and negative image pairs by breaking down the variability of positive pairs. Nevertheless, the ways they separate the variability are different.

- $\tilde{s}(\mathbf{x}_a,\mathbf{x}_b)$ assigns the image pair to the highest scored linear sub-function, using the most discriminative similarity measurement to enhance the separability of the positive pairs. Note that the finally selected sub-function is determined by both probe image and gallery image.
- $s(\mathbf{x}_a,\mathbf{x}_b)$ softly combines the scores of different sub-functions. The sub-functions are weighted by the affinity between the probe image and the exemplar images, in order to increase the discriminability for similar images. In particular, the weight of each sub-function is only determined by the probe image.

Compared with $\tilde{s}(\mathbf{x}_a,\mathbf{x}_b)$, $s(\mathbf{x}_a,\mathbf{x}_b)$ gets rid of the dependence on the gallery image, which accelerates the testing process. Once the probe image is given, $s(\mathbf{x}_a,\mathbf{x}_b)$ is written as:

$$s(\mathbf{x}_a,\mathbf{x}_b) = \left\langle \phi(\mathbf{x}_a,\mathbf{x}_b), \sum_{k=0}^{K} w_a^k\mathbf{W}^k\right\rangle_F . \tag{24}$$

We first compute $\sum_{k=0}^{K} w_a^k\mathbf{W}^k$, then apply it to measure the similarity between the probe image and all the gallery images.

Table 1 The rank-1 and rank-5 matching rates (%) for $\tilde{s}(\mathbf{x}_a, \mathbf{x}_b)$ and $s(\mathbf{x}_a, \mathbf{x}_b)$ on VIPER, GIRD and PRID2011

Datasets	VIPER		GRID		PRID2011	
Methods	r=1	r=5	r=1	r=5	r=1	r=5
$\tilde{s}(\mathbf{x}_a, \mathbf{x}_b)$	39.75	70.44	16.40	34.48	14.10	30.60
$s(\mathbf{x}_a, \mathbf{x}_b)$	41.01	70.85	16.96	36.24	14.40	32.90

Meanwhile, $\tilde{s}(\mathbf{x}_a, \mathbf{x}_b)$ has to compute the scores of all the sub-functions, which is K times slower than $s(\mathbf{x}_a, \mathbf{x}_b)$.

The training of the similarity function is also simplified. For example, we don't need to infer the latent variable for every image pair during the training. Moreover, we can merge the triplet losses with a same probe image, forming the relaxed loss term. The relaxed loss term reduces the number of hinge loss from $\sum_{n=1}^{N} |\mathcal{X}_n^+| \cdot |\mathcal{X}_n^-|$ to N, making the training feasible especially for large training sets.

It is hard to tell whether the relaxed loss term is beneficial to the prediction accuracy. On one hand, such relaxation will not be better than the original triplet loss in terms of training accuracy. On the other hand, the robustness can increase as it potentially reduces the overfitting in the previous training procedure. According to our implementation, the exemplar-guided similarity learning slightly improves the latent similarity learning as shown in Table 1.

5 Optimization

The optimization of Eq. (22) includes two tasks. One is to select exemplars for better discrimination, the other is to learn the coefficients of all the linear functions. Both tasks are mutually beneficial, thus a joint optimization is desirable.

Considering the objective function, we have the following observations: $R_1(\mathcal{W})$ is defined by a closed convex set $\mathcal{C}$; $R_2(\mathcal{W})$ is convex with respect to $\mathcal{W}$; but the loss term $L(\mathcal{W}, \mathcal{E})$ is not convex due to the coupling of $\mathcal{W}$ and $\mathcal{E}$. Once the exemplar set is fixed, denoted by $\mathcal{E}^t$, $L(\mathcal{W}; \mathcal{E}^t)$ becomes convex w.r.t. $\mathcal{W}$, yielding an auxiliary objective function:

$$Q(\mathcal{W}; \mathcal{E}^t) = L(\mathcal{W}; \mathcal{E}^t) + \lambda R_1(\mathcal{W}) + R_2(\mathcal{W}). \tag{25}$$

The minimum of $Q(\mathcal{W}, \mathcal{E}^t)$ is an upper bound for the global minimum of $Q(\mathcal{W}; \mathcal{E})$. This justifies optimizing the objective function in Eq. (22) by minimizing Eq. (25), which is a convex problem and is convenient for optimization. In practice, a two-step iterative algorithm is applied to alteratively optimizing $\mathcal{W}^t$ from the convex subproblem by fixing $\mathcal{E}^t$ and estimating $\mathcal{E}^{t+1}$ by fixing $\mathcal{W}^t$. The whole algorithm is summarized in Algorithm 1.

Algorithm 1 The main algorithm.

```
Input: Dateset $\mathcal{D} = \{\mathbf{x}_n, \mathcal{X}_n^+, \mathcal{X}_n^-\}_{n=1}^N$,
Randomly initialize $\mathcal{W}^0$
Initialize $\mathcal{E}$ via affinity propagation.
for $t = 0, ..., T-1$ (until convergence) do
  Initialize $\mathbf{U}_3^0$ via $\mathcal{W}^t$
  for $l = 0, ..., L-1$ (until convergence) do
    Update $\mathbf{U}_1^{l+1}$ by Eq. (36)
    Update $\mathbf{U}_2^{l+1}$ by applying prox-operator in Eq. (38)
    Update $\mathbf{U}_3^{l+1}$ by projection in Eq. (39)
    Update $\boldsymbol{\Lambda}_1^{l+1}$,
    Update $\boldsymbol{\Lambda}_2^{l+1}$,
  end for
  $\mathcal{W}^{t+1} \leftarrow \mathbf{U}_3^L$
  Estimate $\mathcal{E}^{t+1}$ according to Eq. (27)
end for
$\mathcal{W} \leftarrow \mathcal{W}^T, \mathcal{E} \leftarrow \mathcal{E}^T$
Output: $\mathcal{W}, \mathcal{E}$
```

5.1 Estimating the Exemplars Set $\mathcal{E}^{t+1}$ by Fixing $\mathcal{W}^t$

Given the coefficient set $\mathcal{W}^t$ at the t-th iteration, we aim to estimate the exemplar set $\mathcal{E}^{t+1}$ at the $(t+1)$-th iteration. The estimation of $\mathcal{E}^{t+1}$ is based upon:

$$\mathcal{E}^{t+1} = \arg\min_{\mathcal{E}} Q(\mathcal{E}; \mathcal{W}^t) = \arg\min_{\mathcal{E}} L(\mathcal{E}; \mathcal{W}^t). \tag{26}$$

However, to achieve the best $\mathcal{E}^{t+1}$, we need to test over all the exemplar combinations, whose number is extremely large. Fortunately, selecting the exemplar for one sub-function will not influence the weights of other sub-functions, allowing us to estimate each exemplar sequentially.

To obtain $\mathcal{E}^{t+1}$, we initialize the exemplars by $\mathcal{E}^t$, and then replace each exemplar sequentially via multiple iterations. During this process, the exemplars for c-th visual cue at the l-th iteration are denoted by $\{\tilde{\mathbf{x}}_1^{c,l}, \ldots, \tilde{\mathbf{x}}_k^{c,l}, \ldots, \tilde{\mathbf{x}}_K^{c,l}\}$, and the corresponding exemplars in $(l+1)$-th iteration are estimated as:

$$\begin{aligned}
\tilde{\mathbf{x}}_1^{c,l+1} &= \underset{\tilde{\mathbf{x}}_1^c}{\text{argmin}}\, L(\tilde{\mathbf{x}}_1^c; \mathcal{W}^t, \ldots, \{\tilde{\mathbf{x}}_2^{c,l}, \tilde{\mathbf{x}}_3^{c,l}, \ldots, \tilde{\mathbf{x}}_K^{c,l}\}, \ldots) \\
&\quad s.t. \quad \tilde{\mathbf{x}}_1^c \in \mathcal{X}_e^c / \{\tilde{\mathbf{x}}_2^{c,l}, \tilde{\mathbf{x}}_3^{c,l}, \ldots, \tilde{\mathbf{x}}_K^{c,l}\} \\
\tilde{\mathbf{x}}_2^{c,l+1} &= \underset{\tilde{\mathbf{x}}_2^c}{\text{argmin}}\, L(\tilde{\mathbf{x}}_2^c; \mathcal{W}^t, \ldots, \{\tilde{\mathbf{x}}_1^{c,l+1}, \tilde{\mathbf{x}}_3^{c,l}, \ldots, \tilde{\mathbf{x}}_K^{c,l}\}, \ldots) \\
&\quad s.t. \quad \tilde{\mathbf{x}}_2^c \in \mathcal{X}_e^c / \{(\tilde{\mathbf{x}}_1^c)^{l+1}, (\tilde{\mathbf{x}}_3^c)^l, \ldots, (\tilde{\mathbf{x}}_K^c)^l\} \\
&\quad \vdots \\
(\tilde{\mathbf{x}}_K^c)^{l+1} &= \underset{\tilde{\mathbf{x}}_K^c}{\text{argmin}}\, L(\tilde{\mathbf{x}}_K^c; \mathcal{W}^t, \ldots, \{(\tilde{\mathbf{x}}_1^c)^{l+1}, (\tilde{\mathbf{x}}_2^c)^{l+1}, \ldots, \\
&\quad (\tilde{\mathbf{x}}_{K-1}^c)^{l+1}\}, \ldots) \\
&\quad s.t. \quad \tilde{\mathbf{x}}_K^c \in \mathcal{X}_e^c / \{(\tilde{\mathbf{x}}_1^c)^{l+1}, (\tilde{\mathbf{x}}_2^c)^{l+1}, \ldots, (\tilde{\mathbf{x}}_{K-1}^c)^{l+1}\}.
\end{aligned} \tag{27}$$

In this way, the exemplars are sequentially updated for each visual cue. For each update, the value of the objective function either decreases or stays the same. We stop the updates after 5 iterations. The obtained $\mathcal{E}^{t+1}$ is not the optimum for $\min_{\mathcal{E}} Q(\mathcal{E}; \mathcal{W}^t)$, but it guarantees $Q(\mathcal{W}^t, \mathcal{E}^t) \geq Q(\mathcal{W}^t, \mathcal{E}^{t+1})$.

$\mathcal{E}^0$ needs to be initialized at the very beginning. We adopt affinity propagation (Frey and Dueck 2007) to select K exemplars for every visual cue from the exemplar candidates. For the affinity propagation, the pairwise affinity between two image descriptors $\mathbf{x}_a^c$ and $\mathbf{x}_{a'}^c$ is computed by $\exp^{\beta\langle \mathbf{x}_a^c, \mathbf{x}_{a'}^c \rangle}$.

5.2 Optimizing Coefficients Set $\mathcal{W}^{t+1}$ by Fixing $\mathcal{E}^{t+1}$

Given $\mathcal{E}^{t+1}$, the estimation of $\mathcal{W}^{t+1}$ is a convex problem:

$$\mathcal{W}^{t+1} = \arg\min_{\mathcal{W}} L(\mathcal{W}; \mathcal{E}^{t+1}) + \lambda R_1(\mathcal{W}) + R_2(\mathcal{W}). \tag{28}$$

To clarify the notations, we define $\mathbf{U}$ as a large matrix concatenating all the matrices in $\mathcal{W}$, and define $\Omega(\mathbf{x}_n)$ as the corresponding feature matrix related to the probe image $\mathbf{x}_n$. In $\Omega(\mathbf{x}_n)$, the feature map corresponds to $\mathbf{W}^{c,k}$ equals $w_n^{c,k}\psi^c(\mathbf{x}_n)$. Using these notations, we rewrite $L(\mathcal{W}; \mathcal{E}^{t+1})$ as:

$$L(\mathbf{U}) = \frac{1}{N}\sum_{n=1}^{N}[1 - \langle \Omega(\mathbf{x}_n, \mathbf{U})\rangle_F]_+. \tag{29}$$

Similarly, $R_1(\mathcal{W})$ and $R_2(\mathcal{W})$ can be rewritten as $R_1(\mathbf{U})$ and $R_2(\mathbf{U})$. By defining $g_1(\mathbf{U}) = L(\mathbf{U})$, $g_2(\mathbf{U}) = \lambda R_1(\mathbf{U})$ and $g_3(\mathbf{U}) = R_2(\mathbf{U})$, Eq. (28) becomes $\min_{\mathbf{U}} g_1(\mathbf{U}) + g_2(\mathbf{U}) + g_3(\mathbf{U})$, also equivalent to:

$$\begin{aligned} &\min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3} \; g_1(\mathbf{U}_1) + g_2(\mathbf{U}_2) + g_3(\mathbf{U}_3), \\ &\text{s.t.} \qquad \mathbf{U}_1 = \mathbf{U}_2 = \mathbf{U}_3. \end{aligned} \tag{30}$$

By introducing Lagrange multipliers $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$, we have the augmented Lagrangian:

$$\begin{aligned} La(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2) =& g_1(\mathbf{U}_1) + g_2(\mathbf{U}_2) + g_3(\mathbf{U}_3) \\ &+ \rho\langle \boldsymbol{\Lambda}_1, \mathbf{U}_1 - \mathbf{U}_3\rangle_F + \frac{\rho}{2}\|\mathbf{U}_1 - \mathbf{U}_3\|_F^2 \\ &+ \rho\langle \boldsymbol{\Lambda}_2, \mathbf{U}_2 - \mathbf{U}_3\rangle_F + \frac{\rho}{2}\|\mathbf{U}_2 - \mathbf{U}_3\|_F^2, \end{aligned}$$

where $\rho > 0$ is a scaling parameter. The ADMM in the scaled form consists of the following iterations:

$$\mathbf{U}_1^{l+1} = \arg\min_{\mathbf{U}_1} g_1(\mathbf{U}_1) + \frac{\rho}{2}\|\mathbf{U}_1 - (\mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l)\|_F^2 \tag{31}$$

$$\mathbf{U}_2^{l+1} = \arg\min_{\mathbf{U}_2} g_2(\mathbf{U}_2) + \frac{\rho}{2}\|\mathbf{U}_2 - (\mathbf{U}_3^l - \boldsymbol{\Lambda}_2^l)\|_F^2 \tag{32}$$

$$\begin{aligned} \mathbf{U}_3^{l+1} =& \arg\min_{\mathbf{U}_3} g_3(\mathbf{U}_3) \\ &+ \rho\|\mathbf{U}_3 - \frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \boldsymbol{\Lambda}_1^l + \boldsymbol{\Lambda}_2^l)\|_F^2 \end{aligned} \tag{33}$$

$$\begin{aligned} \boldsymbol{\Lambda}_1^{l+1} &= \boldsymbol{\Lambda}_1^l + \mathbf{U}_1^{l+1} - \mathbf{U}_3^{l+1}, \\ \boldsymbol{\Lambda}_2^{l+1} &= \boldsymbol{\Lambda}_2^l + \mathbf{U}_2^{l+1} - \mathbf{U}_3^{l+1}. \end{aligned} \tag{34}$$

More details about the derivations can be found in the work of Boyd et al. (2011). Here, we will introduce the update procedures of $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$.

Update of $\mathbf{U}_1$. Equation (31) is a convex problem. As $\mathbf{U}_1$ may contain over hundreds of thousands of coefficients, optimization from its primal form is infeasible. Therefore we consider to optimize $\mathbf{U}_1^{l+1}$ from its dual form:

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2\rho}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} - \mathbf{b}^\top\boldsymbol{\alpha}, \quad \text{s.t.} \quad 0 \leq \alpha_n \leq \frac{1}{N}, \quad \forall n, \tag{35}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{N\times 1}$ is the dual variable vector and α_n is its n-th element. The element of $\mathbf{b} \in \mathbb{R}^{N\times 1}$ is $b_n = \langle \mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l, \Omega(\mathbf{x}_n)\rangle_F - 1$. $\mathbf{H} \in \mathbb{R}^{N\times N}$ is a Gram matrix, whose entries are given by $H_{ij} := \langle \Omega(\mathbf{x}_i), \Omega(\mathbf{x}_j)\rangle_F$. The deviation from Eqs. (31) to (35) are given in the Appendix 8.

Equation (35) is a standard quadratic programming problem. With the optimal solution $\boldsymbol{\alpha}^*$, we have:

$$\mathbf{U}_1^{l+1} = \frac{1}{\rho}\sum_{n=1}^{N}\alpha_n^*\Omega(\mathbf{x}_n) + \mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l. \tag{36}$$

Update of $\mathbf{U}_2$. As $g_2(\mathbf{U}) = \lambda\sum_{c=1}^{C}\sum_{k=0}^{K}(\|\mathbf{W}_M^{c,k}\|_{2,1} + \|\mathbf{W}_B^{c,k}\|_{2,1})$, the problem of Eq. (32) is decomposed into $2C(K+1)$ subproblems, and each subproblem optimizes a sub-matrix belonging to $\mathbb{R}^{d\times d}$. Let $\mathbf{U}_2^s \in \mathbb{R}^{d\times d}$ be 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 $\mathbf{U}_3^l - \boldsymbol{\Lambda}_2^l$, the subproblem of optimizing $\mathbf{U}_2^s$ is:

$$\min_{\mathbf{U}_2^s} \lambda\|\mathbf{U}_2^s\|_{2,1} + \frac{\rho}{2}\|\mathbf{U}_2^s - \mathbf{A}^s\|_F^2, \tag{37}$$

whose solution is obtained by a prox-operator:

$$(\mathbf{U}_2^s)_{ij} = \mathbf{A}_{ij}^s\left[1 - \frac{\lambda/\rho}{\|\mathbf{A}_{i.}^s\|_2}\right]_+, \tag{38}$$

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

Update of $\mathbf{U}_3$. According to the definition of $g_3(\mathbf{U})$ and Eq. (33), the update of $\mathbf{U}_3$ is to project $\frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \boldsymbol{\Lambda}_1^l + \boldsymbol{\Lambda}_2^l)$ onto set $\mathcal{C}$:

$$\mathbf{U}_3^{l+1} = \Pi_{\mathcal{C}}\left[\frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \boldsymbol{\Lambda}_1^l + \boldsymbol{\Lambda}_2^l)\right]. \tag{39}$$

For $\frac{1}{2}(\mathbf{U}_1^{l+1}+\mathbf{U}_2^{l+1}+\boldsymbol{\Lambda}_1^l+\boldsymbol{\Lambda}_2^l)$, its sub-matrix corresponding to $\mathbf{W}_M^{c,k}$ is usually not symmetric. As directly projecting a non-symmetric matrix onto $\mathbb{S}_-^d$ is difficult, we operate a separated ADMM algorithm consisting of two projection steps. One is to project the sub-matrices onto $\mathbb{S}^d$ by the projection rule $f(\mathbf{W}) := \frac{1}{2}(\mathbf{W}+\mathbf{W}^\top)$, the other is to project symmetric matrices onto $\mathbb{S}_-^d$ by cropping the positive eigenvalues to be 0. More details about the updating procedures are relegated to the Appendix 9.

We update $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$ iteratively until convergence, which finally outputs $\mathcal{W}^{t+1}$. As Eq. (28) is a convex problem, the updates can guarantee $Q(\mathcal{W}^t,\mathcal{E}^{t+1}) \geq Q(\mathcal{W}^{t+1},\mathcal{E}^{t+1})$.

5.3 Discussion

The Convergence The estimation of $\mathcal{E}^{t+1}$ in Sect. 5.1 and the estimation of $\mathcal{W}^{t+1}$ in Sect. 5.2 decrease the value of Q, we have $Q(\mathcal{W}^t,\mathcal{E}^t) \geq Q(\mathcal{W}^t,\mathcal{E}^{t+1}) \geq Q(\mathcal{W}^{t+1},\mathcal{E}^{t+1})$. Let $Q^t = Q(\mathcal{W}^t,\mathcal{E}^t)$, we obtain a decreasing sequences $\{Q^t\}$, and the sequence is bounded below by 0 because the objective function is nonnegative. According to the monotone convergence theorem, it will converge to some value Q^*. As Eq. (22) is a non-convex problem, our method Q^* is largely dependent on the initial selection of the exemplars. Similar to many other EM-like methods, our algorithm can achieve moderate accuracy but cannot guarantee the global optimum.

The efficiency Compared with the optimization of our conference version, the current one is much more efficient. The efficiency mainly stems from the update of $\mathbf{U}_1$: Instead of updating $\mathbf{U}_1$ by subgradient descent method, we solve a convex problem from its dual form, transforming the coefficient number from $O(d^2)$ to $O(N)$. The practicability of the dual solution benefits from two factors: (1) Unlike $\tilde{s}(\mathbf{x}_a,\mathbf{x}_b)$ of Eq. (23), the exemplar-guided function doesn't need to estimate the latent variables for every image pair during the update, thus the Gram matrix $\mathbf{H}$ can be pre-computed. (2) The employment of the relaxed loss term largely reduces the number of hinge loss constraints, making $\mathbf{H}$ be a small matrix with the size of $N \times N$.

6 Experiments

In this section, we conduct an extensive set of experiments on six public available datasets. We present more implementation details in Sect. 6.1, then delve into the effectiveness of each component by constructing a series of variant methods in Sects. 6.2 and 6.3. In Sect. 6.4, we compare with other state-of-the-art methods over various datasets. Some other properties are also studied in Sect. 6.5, including the cross-dataset experiments showing the generalization ability of our method.

6.1 Implementation Details

6.1.1 The Generation of Visual Cues

The visual cues are obtained by both feature extraction and dimensionality reduction. Firstly, image descriptors are extracted from local patches, then PCA is applied to the concatenated descriptors for dimensionality reduction.

Feature Extraction: Images are divided into 15 × 5 overlapped patches. From each patch, we extract 6 types of basic histogram features, i.e., HSV^1, HSV^2, LAB^1, LAB^2, HOG and SILPT. Among them, HSV^1 and LAB^1 are joint $8 \times 8 \times 8$ histograms for HSV channels and LAB channels, respectively. HSV^2 and LAB^2 are concatenated histograms with each channel having 16 bins. HOG (Dalal and Triggs 2005) and SILTP (Liao et al. 2010) are texture features. HOG feature represents the occurrences of gradient orientation in each region and is invariant to illumination changes. SILTP is an improved descriptor over the LBP (Ojala et al. 2002), achieving invariance to intensity scale changes and robustness to image noises.

Dimensionality Reduction: To effectively represent the person image, each visual cue concatenates both color and texture features. We made four visual cues Cue_1, Cue_2, Cue_3 and Cue_4, composed by HSV^1/SILTP, HSV^2/HOG, LAB^1/SILTP and LAB^2/HOG, respectively. PCA is applied over the concatenated descriptors for dimensionality reduction. To limit the impact of co-occurrence (Jégou and Chum 2012), we do a whitening process after PCA, which divides each dimension by the inverse of the square root of the corresponding eigenvalue. The resulting feature vectors are normalized, making the L_2 norm of each visual cue be 1.

6.1.2 Parameter Setting

We have $C = 4$ visual cues, and each visual cue has $K = 5$ exemplars. In optimization, we empirically set the maximal iteration number $T = 5$, and the scaling parameter $\rho = 0.1$. The PCA reduced dimension d and the sparsity controlling parameter λ are determined by cross-validation. We separate the training data into 10 folders, where 9 folders are used for training while 1 folder is used for validation. Once d and λ are determined, we use all the training data for learning, and repeat the training and testing for multiple partitions. The determined parameters for six datasets are in Table 2.

Table 2 Parameters for different datasets

	VIPER	GRID	PRID2011	3DPeS	CUHK03	Market-1501
d	100	70	60	70	550	500
λ	0.005	0.003	0.0003	0.005	0.02	0.001

Table 3 Empirical analysis with Cue_1 on VIPER. We report the top-n matching rate (%) for different configurations about feature maps, sparsity regularization, semi-definite regularization, loss function and dimensionality reduce methods

Methods	Feature	K	Sparsity	Semi-definite	Dim-reduce	r = 1	r = 5	r = 10	r = 20	r = 25	r = 50
F1	$\phi_1(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_M^k \preceq 0$	PCA	34.56	63.26	75.70	86.23	89.15	95.92
F2	$\phi_2(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	*None*	PCA	28.96	59.97	74.78	87.44	90.41	96.17
F3	$\phi_3(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_{3,1}^k \preceq 0, \mathbf{W}_{3,2}^k \succeq 0,$	PCA	30.38	60.13	72.53	85.03	87.72	94.08
F4	$\phi_4(\mathbf{x}_a, \mathbf{x}_b)$	5	$L2$ norm	*None*	PCA	10.44	25.47	34.08	45.82	49.81	66.77
L1R	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	$L1$ norm	$\mathbf{W}_M^k \preceq 0$	PCA	36.65	67.94	80.89	90.28	92.56	97.31
L2R	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	$L2$ norm	$\mathbf{W}_M^k \preceq 0$	PCA	34.94	66.36	79.02	89.78	91.93	96.65
Semi-0	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	None	PCA	23.10	55.85	70.73	83.89	86.55	94.02
Semi-$\mathrm{M}_N\mathrm{B}_P$	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_M^k \preceq 0, \mathbf{W}_B^k \succeq 0$	PCA	35.85	65.60	77.53	87.69	90.35	96.52
Semi-$\mathrm{M}_N\mathrm{B}_N$	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_M^k \preceq 0, \mathbf{W}_B^k \preceq 0$	PCA	35.09	64.40	76.87	86.93	89.97	95.98
Single	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	0	Group sparse	$\mathbf{W}_M^k \preceq 0$	PCA	**36.77**	*68.58*	*81.04*	*90.98*	*93.32*	*97.82*
LDA-D	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_M^k \preceq 0$	LDA	35.19	63.45	75.09	85.54	88.70	94.87
CCA-D	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_M^k \preceq 0$	CCA	18.86	47.66	63.51	78.73	83.16	92.22
proposed	$\phi(\mathbf{x}_a, \mathbf{x}_b)$	5	Group sparse	$\mathbf{W}_M^k \preceq 0$	PCA	**41.01**	**70.85**	**83.64**	**92.37**	**94.05**	**98.01**

The best two results are denoted in bold and italics

However, it is desirable to determine d and λ directly by a practical parameter selection guidance rather than cross-validation. Currently, such guidance is difficult because the hyper-parameters are sensitive to the specific dataset. Nevertheless, automatically determining the hyper-parameter is an interesting and promising problem that needs to be explored in the future.

6.1.3 Evaluation Scheme

For evaluation, the persons in each dataset are separated into the training set and the test set so that the images of a same person can only be used for training or used for testing. The single-shot evaluation scheme (Li et al. 2013; Pedagadi et al. 2013; Xiong et al. 2014) is adopted for VIPER, GRID, PRID2011, 3DPeS and CUHK03. In the scheme, the testing set is further separated into two parts: the probe set and the gallery set. Each person has one image in the gallery set, and has one image (VIPER, GRID, PRID2011) or multiple images (3DPeS, CUHK03) in the gallery set. The results are displayed by cumulative matching characteristic (CMC) curve (Gray et al. 2007), which is an estimate of the expectation of finding the correct match in the top n matches. The evaluation scheme for Market-1501 is slightly different as described in the work of Zheng et al. (2015), where the provided gallery set contains multiple images for one person.

6.2 Empirical Analysis with a Single Visual Cue

We investigate how various factors affect the proposed similarity learning when using a single visual cue. For this purpose, we construct multiple variants with different settings, and all these variants are with 100-dimensional PCA reduced descriptor of Cue_1.

The experiments in this section are evaluated on VIPER (Gray et al. 2007), a challenging person re-identification dataset that has been widely used for benchmark evaluation. It contains 632 pairs of person images, which are scaled to be 128 × 48 pixels. Images are with large variations in viewpoint, illumination and background content as shown in Fig. 8. The evaluated methods are trained and tested with non-overlapped 316 image pairs, and their results are averaged over 10 trails of different training and test set partitions.

6.2.1 The Effect of the Feature Map

To study the effectiveness of the proposed feature map $\phi(\mathbf{x}_a, \mathbf{x}_b)$, we construct 4 variants F1, F2, F3 and F4 by replacing $\phi(\mathbf{x}_a, \mathbf{x}_b)$ with $\phi_1(\mathbf{x}_a, \mathbf{x}_b)$, $\phi_2(\mathbf{x}_a, \mathbf{x}_b)$, $\phi_3(\mathbf{x}_a, \mathbf{x}_b)$ and $\phi_4(\mathbf{x}_a, \mathbf{x}_b)$, where

$$\begin{aligned}
\phi_1(\mathbf{x}_a, \mathbf{x}_b) &= (\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top, \\
\phi_2(\mathbf{x}_a, \mathbf{x}_b) &= \mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top, \\
\phi_3(\mathbf{x}_a, \mathbf{x}_b) &= [\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top, \mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top]. \\
\phi_4(\mathbf{x}_a, \mathbf{x}_b) &= [|(\mathbf{x}_a)_1 - (\mathbf{x}_b)_1|, \ldots, |(\mathbf{x}_a)_d - (\mathbf{x}_b)_d|]^\top.
\end{aligned} \qquad 835$$

Among them, $\phi_1(\mathbf{x}_a, \mathbf{x}_b)$ corresponds to $\phi_M(\mathbf{x}_a, \mathbf{x}_b)$, and $\phi_2(\mathbf{x}_a, \mathbf{x}_b)$ corresponds to $\phi_B(\mathbf{x}_a, \mathbf{x}_b)$. $\phi_3(\mathbf{x}_a, \mathbf{x}_b)$ is another organization of explicit polynomial kernel feature map. $\phi_4(\mathbf{x}_a, \mathbf{x}_b)$ is a feature vector that concatenates the absolute values of element-wise differences between $\mathbf{x}_a$ and $\mathbf{x}_b$. The corresponding regularization strategies for the feature maps are shown in Table 3.

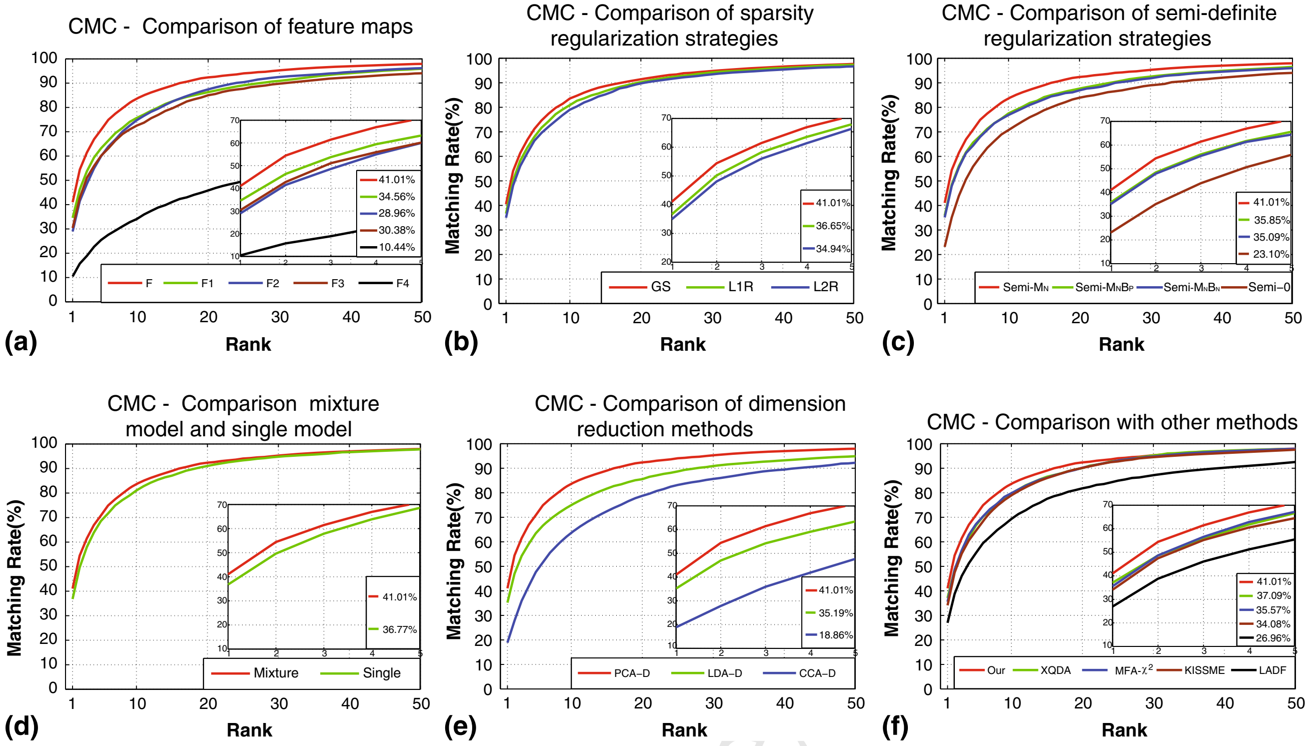

Fig. 4 Empirical analysis with visual cue Cue_1 on VIPER. We display the average CMC curves for different configurations about (**a**) feature map, **b** sparsity regularization strategy, **c** semi-definite regularization strategy, **d** mixture model, **e** dimensionality reduction method. Besides, we also compare with other similarity learning methods using the same visual cue

Denoting the method employing $\phi(\mathbf{x}_a, \mathbf{x}_b)$ by F, we evaluate it along with F1, F2, F3 and F4. As shown in Fig. 4a, we find F1 performs better than F2 at the first several ranks of the CMC curve, but when the rank increases, the accuracy of F2 can surpass F1. This phenomenon verifies the complementary properties of the Mahalanobis distance and bilinear similarity. F4 performs much worse than other methods, which in turn proves the second-order polynomial kernel features can encode more abundant information than the first order ones.

$\phi_3(\mathbf{x}_a, \mathbf{x}_b)$ is exactly the second-order part of the polynomial kernel feature map employed by LADF. The coefficients of the k-th sub-function $\mathbf{W}^k_{3,1}$ and $\mathbf{W}^k_{3,2}$, corresponding to $\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top$ and $\mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top$, are imposed to be NSD and PSD. To investigate why our feature map is superior than that of LADF, we conducted a series of experiments as shown in Table 4.

Table 4 Analysis of $\phi_3(\mathbf{x}_a, \mathbf{x}_b)$. We perform different regularization strategies for $\phi_3(\mathbf{x}_a, \mathbf{x}_b)$, and compare their performance, where the coefficients of the k-th sub-function corresponding to $\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top$ and $\mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top$ are denoted by $\mathbf{W}^k_{3,1}$ and $\mathbf{W}^k_{3,2}$, respectively

Configuration		Results			
		r = 1	r = 5	r = 10	r = 20
$\mathbf{W}^k_{3,1} \preceq 0$	$\mathbf{W}^k_{3,2} = 0$	0.57	1.99	3.83	7.22
None	$\mathbf{W}^k_{3,2} = 0$	0.51	2.31	4.94	9.49
$\mathbf{W}^k_{3,1} \succeq 0$	$\mathbf{W}^k_{3,2} = 0$	0.38	1.71	3.35	6.74
$\mathbf{W}^k_{3,1} \succeq 0$	$\mathbf{W}^k_{3,2} \succeq 0$	30.35	59.91	72.50	84.75
None	$\mathbf{W}^k_{3,2} \succeq 0$	30.63	60.16	72.91	84.94
$\mathbf{W}^k_{3,1} = 0$	$\mathbf{W}^k_{3,2} \succeq 0$	30.41	60.22	72.50	84.97
$\mathbf{W}^k_{3,1} \preceq 0$	$\mathbf{W}^k_{3,2} \succeq 0$	30.38	60.13	72.53	85.03
$\mathbf{W}^k_{3,1} \preceq 0$	None	27.41	57.97	72.22	84.62
$\mathbf{W}^k_{3,1} \preceq 0$	$\mathbf{W}^k_{3,2} \preceq 0$	2.53	11.14	20.32	33.92

From the results, we conclude that the effectiveness of $\phi_3(\mathbf{x}_a, \mathbf{x}_b)$ is only related to $\mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top$. The similarity function can hardly work when we set the corresponding coefficients to be zero. It also shows that imposing PSD for $\mathbf{x}_a\mathbf{x}_b^\top + \mathbf{x}_b\mathbf{x}_a^\top$ is a little better than imposing no constraints, much better than imposing NSD. In contrast, regularization has little influence on $\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top$: various regularization are imposed over the coefficients including setting the coefficient to be zero, but they have little influence over the performance.

In our opinion, the feature map for linear similarity function should directly encode the connections between the descriptors of two images. In $\phi_3(\mathbf{x}_a, \mathbf{x}_b)$, $\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top$ just

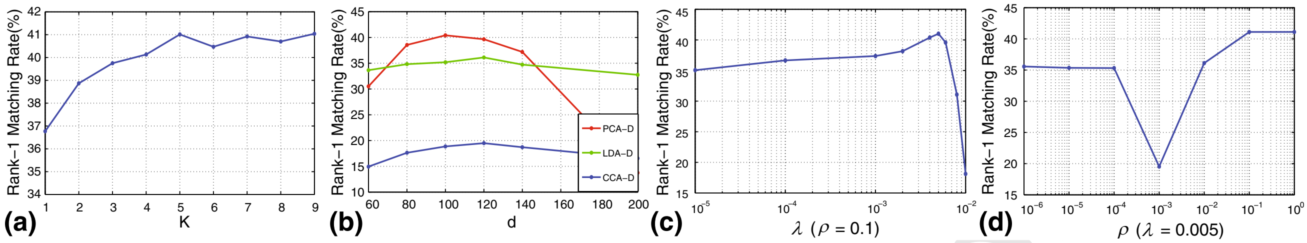

Fig. 5 The rank-1 matching rate (%) on the VIPER dataset with respect to (**a**) the exemplar number K, **b** the reduced dimensionality d, **c** the parameter λ when ρ is fixed to be 0.1, **d** the parameter ρ when λ is fixed to be 0.005

relies on individual image descriptor rather than their relationship. Its similarity score $\langle \mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top, \mathbf{W}_{3,1}^k \rangle_F$, which equals $\mathbf{x}_a^\top \mathbf{W}_{3,1}^k \mathbf{x}_a + \mathbf{x}_b^\top \mathbf{W}_{3,1}^k \mathbf{x}_b$, only depends on the individual appearance of the two images, thus cannot measure the similarity between $\mathbf{x}_a$ and $\mathbf{x}_b$. Instead of $\mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top$, we employ $(\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top$ to directly encode the inter-image information, showing much better accuracy.

The organization of the feature map also imposes strong prior over the coefficients, thus is critical to the similarity learning. For example, as $(\mathbf{x}_a - \mathbf{x}_b)(\mathbf{x}_a - \mathbf{x}_b)^\top = \mathbf{x}_a\mathbf{x}_a^\top + \mathbf{x}_b\mathbf{x}_b^\top - \mathbf{x}_a\mathbf{x}_b^\top - \mathbf{x}_b\mathbf{x}_a^\top$, the learned coefficients are shared by $\mathbf{x}_a\mathbf{x}_a^\top$, $\mathbf{x}_b\mathbf{x}_b^\top$, $\mathbf{x}_a\mathbf{x}_b^\top$, $\mathbf{x}_b\mathbf{x}_a^\top$, where $\mathbf{x}_a\mathbf{x}_a^\top$ and $\mathbf{x}_b\mathbf{x}_b^\top$ are with the positive signs while $\mathbf{x}_a\mathbf{x}_b^\top$ and $\mathbf{x}_b\mathbf{x}_a^\top$ are with the negative signs. It can be seen that the organization largely restricts the formation of coefficients.

6.2.2 The Effect of the Regularization.

Two kinds of regularization are imposed upon the feature map. The coefficients in $\mathcal{W}$ should be sparse while the sub-matrices corresponding to the Mahalanobis distance are imposed to be NSD.

Sparsity plays an important role in the feature selection. We compare group sparse(GS) with two variants L1R and L2R, whose coefficients are regularized by L_1 and L_2 norm. In greater detail, GS penalizes $\|\mathbf{W}\|_{2,1} := \sum_{i=1}^{d} \|\mathbf{W}_{i.}\|_2$ for each sub-matrix $\mathbf{W}$ as shown in Eq. (17), while L1R and L2R penalize the sub-matrix using $\|\mathbf{W}\|_1 := \sum_{i,j} |\mathbf{W}_{ij}|$ and $\|\mathbf{W}\|_2 := \sqrt{\sum_{i,j} |\mathbf{W}_{ij}|^2}$. The results in Fig. 4b show that L1R improves L2R from 34.94 to 36.65% on the rank-1 matching rate, and GS further improves L1R from 36.65 to 41.01%.

Proper semi-definite regularization can further improve the effectiveness. The current similarity function only requires $\{\mathbf{W}_M^k\}_{k=0}^K$ to be NSD, thus is denoted by Semi-M$_N$. In addition, we construct Semi-0, Semi-M$_N$B$_P$ and Semi-M$_N$B$_N$, where Semi-0 imposes no semi-definite regularization, while Semi-M$_N$B$_P$ and Semi-M$_N$B$_N$ not only enforce $\{\mathbf{W}_M^k\}_{k=0}^K$ to be NSD but also let $\{\mathbf{W}_B^k\}_{k=0}^K$ be PSD and NSD, respectively. Figure 4c shows that imposing NSD for $\{\mathbf{W}_M^k\}_{k=0}^K$ boosts the accuracy significantly, while the situation is different for $\{\mathbf{W}_B^k\}_{k=0}^K$: imposing either NSD or PSD over $\{\mathbf{W}_B^k\}_{k=0}^K$ will decrease the rank-1 matching rate about $5 \sim 6\%$.

6.2.3 The Effect of the Exemplar-Guided Mixture Model

We introduce an exemplar-guided mixture model consisting of K sub-functions. As the mixture is trained with weighted feature map $\Omega(\mathbf{x}_n)$, it will increase the discriminative ability for the persons with similar appearance. To evaluate the effectiveness, we compare the exemplar-guided similarity function $s(\mathbf{x}_a, \mathbf{x}_b)$ with $f(\mathbf{x}_a, \mathbf{x}_b)$, the variant using a single linear function. Results in Fig. 4d show that employing the mixture model can actually help to increase the matching rate, where the rank-1 matching rate has been improved from 36.77 to 41.01%.

To evaluate how much the exemplar-guided mixture model can benefit our method, we demonstrate the rank-1 matching rates with respect to the number of sub-functions K in Fig. 5a. It can be seen that a mixture model with multiple sub-functions performs better than the model using a single linear similarity function. The performance increases quickly with K up to 5 and stays stable afterwards.

6.2.4 The Effect of Dimensionality Reduction Methods

Dimensionality reduction is an important procedure that enables our similarity function to be practical. To study the effect of different dimension reduction methods, we construct three variants PCA-D, LDA-D and CCA-D, which apply PCA, LDA (Fisher 1936) and CCA (Hotelling 1936) to reduce the concatenated image descriptors of each visual cue. Among them, PCA is an unsupervised method, while CCA and LDA are supervised methods. Figure 4e compare their performance in terms of CMC curve with the reduced dimensionality $d = 100$, and Fig. 5b displays how their rank-1 matching rates change with d. Although PCA does not need any label for training, it can still outperforms LDA and CCA when $d = 100$. One reason is that the discriminative learning of LDA and CCA may lose important information that could have been exploited by our approach.

Table 5 Comparison with different methods using Cue_1

Methods	r = 1	r = 5	r = 10	r = 20	r = 50
LADF	26.96	55.41	69.30	81.80	92.53
KISSME	34.08	64.46	78.99	90.13	97.56
MFA-χ^2	35.57	67.15	79.81	90.19	97.91
XQDA	37.09	66.58	79.68	90.03	97.88
Our	**41.01**	**70.85**	**83.64**	**92.37**	**98.01**

6.2.5 Comparisons with Other Methods Using Cue_1

In order to get a better understanding of the proposed similarity learning, we compare the exemplar-guided similarity function with the existing state-of-the-art methods using the same visual cue Cue_1. The compared methods include XQDA(Liao et al. 2015), MFA-χ^2(Xiong et al. 2014), KISSME (Köstinger et al. 2012) and LADF (Li et al. 2013). In the experiment, we adopt the implementation provided by the authors, replace their features with Cue_1 and evaluate them on VIPER. The results are reported in Fig. 4f and Table 5 . It can be seen that our method consistently outperforms the others over all ranks.

6.2.6 The Influence of Parameters

The sparsity controlling parameter λ in Eq. (22) and the penalty parameter ρ in ADMM algorithm are analyzed. As ρ and λ interact with each other, we show how the performance changes with respect to λ by fixing $\rho = 0.1$, and show the influence of ρ by fixing $\lambda = 0.005$. All the experiments are with $d = 100$, and the results are demonstrated in Fig. 5c, d. It can be seen that both too large and too small λ will lead to inferior results. This is because small λ imposes little sparsity while large λ makes the coefficients too much sparse. In contrast, the performance w.r.t. ρ is less sensitive than that w.r.t. λ.

6.3 Empirical Analysis on Multi-cue Collaboration

6.3.1 The Results with Multiple Visual Cues

We analyze how the collaboration of multiple visual cues can elevate our approach. To show the effectiveness of the collaborative model, we compare it with the variants Cue1, Cue2, Cue3 and Cue4, which employ only one the visual cue Cue_1, Cue_2, Cue_3 and Cue_4, respectively. To study the necessity of each visual cue, we further construct no-Cue1, no-Cue2, no-Cue3 and no-Cue4 that exclude the corresponding visual cue from the collaborative model. All these models are trained with 316 persons and 100 persons, respectively. The corresponding CMC curves are demonstrated in Fig. 6, and the related top-n matching rates are reported in Table 6.

Figure 6a, c reveal the advantages of multi-cue collaboration. The collaborative model outperforms the variants using a single visual cue by a significant large margin, and it is quite robust as the performance is not strongly dependent on a certain visual cue. Multi-cue collaboration consistently improves the performance despite the changes of training data size. The results indicate that employing more visual cues is helpful, and it is expected to utilize more distinguished visual cues to further improve the accuracy (Table 7).

6.3.2 The Utilization of Multiple Visual Cues

How to make use of multiple visual cues is another important problem that needs to be investigated. We compare two solutions. One is to learn a single similarity function, and its features are concatenated by the different components of the visual cues (1-sim-of-4-cues). The other is to firstly learn 4 independent similarity functions and then fuse their scores together (sum-of-4-sims). The results in Fig. 7a show that both solutions can benefit from the strength of multiple visual cues, while sum-of-4-sims is more effective than 1-sim-of-4-cues (46.49 vs. 42.88%).

However, the conclusion depends on the specific metric learning methods. For examples, XQDA (Liao et al. 2015) prefers 1-sim-of-4-cues, while KISSME (Köstinger et al. 2012) prefers sum-of-4-cues. Both KISSE and our method employ the PCA reduced visual descriptors, and they can benefit more from the sum of different similarity scores (Table 8).

6.3.3 The Evaluation of Joint Learning

To further take advantage of multiple visual cues, the proposed joint learning goes beyond the sum fusion. It not only selects effective feature from each feature map, but also makes consistency among different feature maps. As shown in Fig. 7, the joint learning consistently outperforms the sum fusion. For example, the rank-1 matching rate is improved from 46.49 to 50.28%.

6.4 Comparison with the State-of-the-Art Approaches

We term our **m**ulti-cue **e**xample-guided **s**imilarity function on **p**olynomial kernel feature map as MESP. To evaluate the effectiveness of the exemplar-guided mixture model as well as the collaboration of multiple visual cues on various datasets, we construct two variants ESP and MSP. ESP corresponds to similarity function $s(\mathbf{x}_a, \mathbf{x}_b)$ in Eq. (13), which exploits the effectiveness of multiple exemplars but uses only one visual cue Cue_1, while MSP corresponds to $f'(\mathbf{x}_a, \mathbf{x}_b)$ in Eq. (12) that employs all the four visual cues but learns a single linear function.

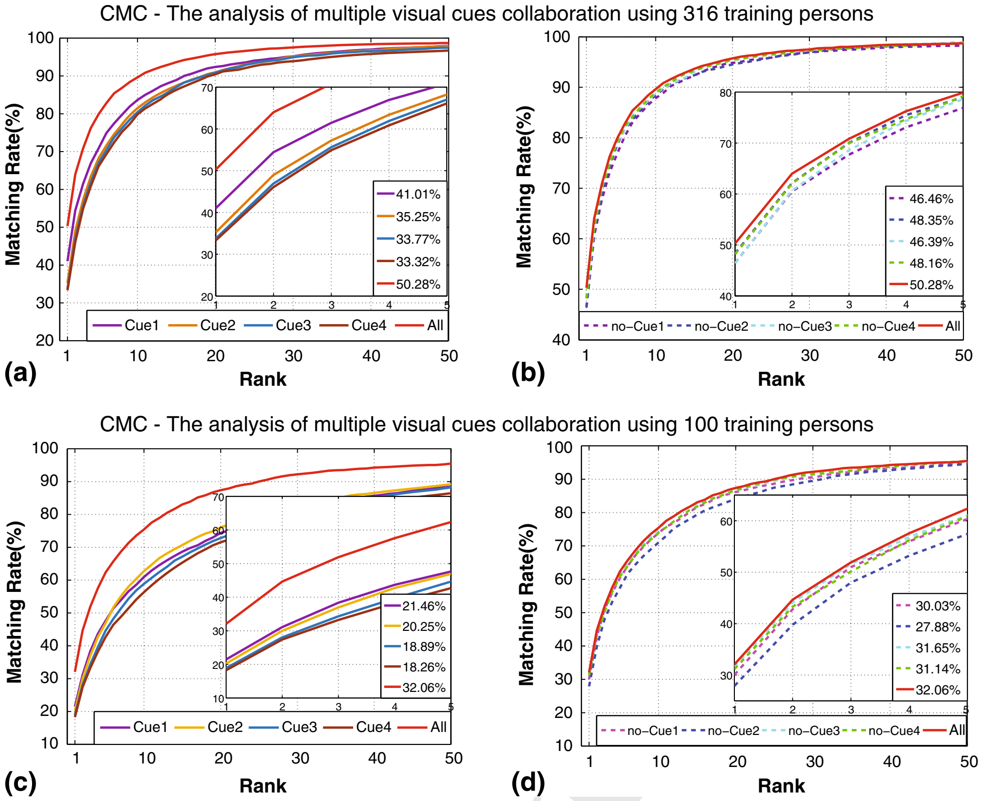

Fig. 6 Empirical analysis of multi-cue collaboration. We display CMC curves with different visual cue configurations, where the models in (**a**, **b**) are trained by 316 identities, and the models in (**c**, **d**) are trained by 100 identities. **a**, **c** compare the model using 4 visual cues with the models using 1 visual cue, while (**b**, **d**) compare the model using 4 visual cues with the models using 3 visual cues

Table 6 Empirical analysis of multi-cue collaboration

Method	Cue configuration	The results of models trained with 316 individuals						The results of models trained with 100 individuals					
		r = 1	r = 5	r = 10	r = 20	r = 25	r = 50	r = 1	r = 5	r = 10	r = 20	r = 25	r = 50
Cue1	Cue_1	41.01	70.85	83.64	92.37	94.05	98.01	21.46	47.63	60.95	74.40	78.96	88.70
Cue2	Cue_2	35.25	68.23	81.39	90.95	93.42	97.91	20.25	46.90	62.72	76.01	79.65	89.43
Cue3	Cue_3	33.77	67.03	80.41	90.85	93.35	97.56	18.89	44.62	58.86	72.82	76.80	88.54
Cue4	Cue_4	33.32	66.11	79.84	90.41	92.63	96.77	18.26	42.69	56.23	71.52	75.89	86.65
no-Cue1	Cue_2, Cue_3, Cue_4	46.46	76.87	87.85	94.84	95.85	98.35	30.03	60.35	73.83	86.14	88.64	95.06
no-Cue2	Cue_1, Cue_3, Cue_4	*48.35*	78.99	88.86	95.44	*96.71*	**98.96**	27.88	57.44	71.01	84.08	87.47	94.65
no-Cue3	Cue_1, Cue_2, Cue_4	46.39	78.61	88.42	94.53	95.89	*98.77*	31.65	61.04	75.13	86.55	89.49	95.54
no-Cue4	Cue_1, Cue_2, Cue_3	48.16	*79.18*	*88.89*	*95.47*	96.46	98.73	*31.14*	*60.85*	*73.96*	**86.71**	*89.91*	*95.54*
All	$Cue_1, Cue_2, Cue_3, Cue_4$	**50.28**	**79.87**	**89.62**	**95.73**	**96.80**	*98.77*	**32.06**	**62.34**	**75.41**	**87.37**	**90.03**	**95.57**

We report the top-n matching rate (%) for different visual cue configurations. The best two results are denoted in bold and italics

6.4.1 Results on VIPER

We compare MESP with the state-of-the-art approaches on the VIPER dataset. All these methods take 316 persons for training and 316 persons for testing. The average results over 10 trails are summarized in Table 9, and the corresponding CMC curves are displayed in Fig. 9a.

MESP outperforms all the other methods, achieving 50.28% on the rank-1 matching rate. It is even better than the combinational approach LMF+LADF (Zhao et al. 2014) and ensemble method ME (Paisitkriangkrai et al. 2015). We emphasize that MESP is quite different from LMF+LADF and ME. Instead of combining the results of different kinds of similarity functions, MESP forms a unified learning problem

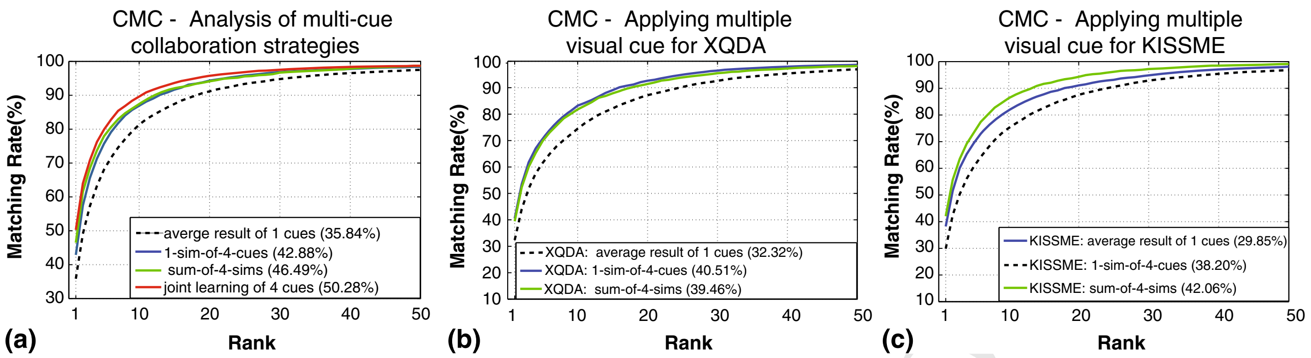

Fig. 7 The analyses of multi-cue collaboration for (**a**) Our method, (**b**) XQDA, (**c**) KISSME

Table 7 The experimental protocol of GRID and PRID2011

Dataset	Training set	Test set Probe set	Gallery set
GRID	125 image pairs	125 images	900 images
PRID2011	100 image pairs	100 images	649 images

to collaborate multiple visual cues in the feature level rather than in the similarity score level. Besides, MESP only utilizes the low level descriptors directly extracted from the image patch, much efficient than clustering-based mid-level features (Zhang et al. 2014; Ma et al. 2012b) and CNN-learned high level features (Ding et al. 2015; Li et al. 2014).

To analyze the effectiveness of our similarity function, we compare MESP, MSP and ESP. Figure 9a shows that the collaboration of multiple visual cues, which improves the orange curve of ESP to the red curve of MESP, contributes more to the performance gain. Meanwhile, employing exemplar-guided mixture model is indispensable, which improves the rank-1 matching rate from 48.10 to 50.28%.

6.4.2 Results on GRID and PRID2011

GRID (Loy et al. 2013) and PRID2011 (Hirzer et al. 2011) are two difficult datasets containing the images of pedestrians in the underground station and street, respectively. As each person in these datasets has only one or two images, the two datasets adopt a very challenging experimental protocol. They provide a small number of image pairs for training, but for test, they have many additional gallery images that describe different persons from the persons described by the probe images. The detail experimental configurations for training and test are given in Table 7. The sample images of the two datasets are demonstrated in Fig. 8.

We slightly and blindly crop the borders of the images in the PRID2011 dataset to reduce the background content, and then resize the images of both PRID2011 and GRID into 128 × 48 for feature extraction. The results for two datasets are averaged over 10 random partitions of training and test samples. In particular, the partitions of GRID have already been provided. The results in the form of CMC curves are displayed in Fig. 9b, c, and the typical top-n matching rates are listed in Tables 10 and 11.

Due to the challenging experimental protocol, the matching rates on the two datasets are generally lower than those on the VIPER dataset, while our method significantly outperforms the other methods. By comparing ESP, MSP and MESP, we find that the collaboration of multiple visual cues makes considerable improvement on all the range of CMC curves. Meanwhile, the exemplar-guided mixture model also steadily improves the performance especially at lower rank matching rates.

6.4.3 Results on 3DPES, CUHK03 and Market-1501

Datasets 3DPES (Baltieri et al. 2011), CUHK03 (Li et al. 2014) and Market-1501 (Zheng et al. 2015) provide more images for each person, describing the persons from multiple

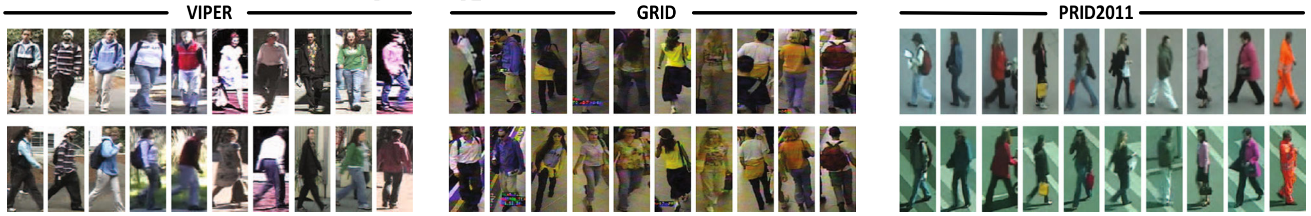

Fig. 8 Example images from VIPER, GRID and PRID2011. Images of each column are about a same person

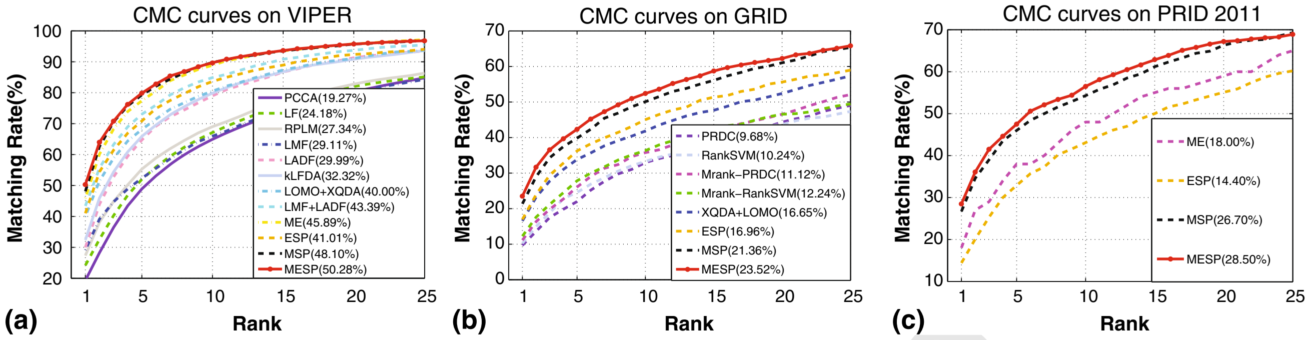

Fig. 9 CMC curves of our method and other state-of-the-art methods on (**a**) the VIPER dataset with 316 gallery images, (**b**) the GRID dataset with 900 gallery images, (**c**) the PRID2011 dataset with 649 gallery images

Table 8 Configurations of 3DPES, CUHK03, Market-1501

Dataset	Images	People	Camera	Training	Test
3DPES	1101	192	8	96	96
CUHK03	13164	1360	6	1260	100
Market-1501	32668	1501	6	751	700

Table 9 The rank-n matching rates (%) for comparison with PCCA (Mignon and Jurie 2012), LF (Pedagadi et al. 2013), RPLM (Hirzer et al. 2012), kLFDA (Xiong et al. 2014), LMF (Zhao et al. 2014), LADF (Li et al. 2013), NVWCM (Zhang et al. 2014), SCNCD (Yang et al. 2014), LOMO+XQDA (Liao et al. 2015), LMF+LADF (Zhao et al. 2014), and ME (Paisitkriangkrai et al. 2015) on the VIPER dataset

Method	r = 1	r = 5	r = 10	r = 20
PCCA	19.27	48.89	64.91	80.28
LF	24.18	52.00	67.12	82.00
RPLM	27.34	55.30	69.02	82.69
kLFDA	32.32	65.78	79.72	90.95
LMF	29.11	52.34	65.95	79.87
LADF	29.99	64.71	79.00	91.29
NVWCM	30.70	62.97	75.95	- -
SCNCD	37.80	68.50	81.20	90.40
LOMO+XQDA	40.00	68.13	80.51	91.08
LMF+LADF	43.39	73.04	84.87	93.07
ME	45.89	77.47	88.87	**95.81**
ESP	41.01	70.85	83.64	92.37
MSP	48.10	79.30	89.58	95.66
MESP	**50.28**	**79.87**	**89.62**	95.73

EPS, MSP and MESP are our methods with different configurations. The size of the gallery set is 316

Table 10 The rank-n matching rates (%) for comparison with PRDC (Zheng et al. 2013), RankSVM (Prosser et al. 2010), MRank-RankSVM, MRank-PRDC (Loy et al. 2013), MtMCML (Ma et al. 2014) and XQDA+LOMO (Liao et al. 2015) on the GRID dataset

Method	r = 1	r = 5	r = 10	r = 20
PRDC	9.68	22.00	32.96	44.32
RankSVM	10.24	24.56	33.28	43.68
MRank-PRDC	11.12	26.08	35.76	46.56
MRank-RankSVM	12.24	27.84	36.32	46.56
MtMCML	14.08	34.64	45.84	59.84
XQDA+LOMO	16.56	33.84	41.84	52.40
ESP	16.96	36.24	45.04	55.60
MSP	21.36	39.84	50.08	61.04
MESP	**23.52**	**42.32**	**52.40**	**62.24**

EPS, MSP and MESP are different configurations of our method. The size of the gallery set is 900

Table 11 The rank-n matching rates (%) for comparison with RPLM (Hirzer et al. 2012) and ME (Paisitkriangkrai et al. 2015) on the PRID dataset. EPS, MSP and MESP are different configurations of our method

Method	r = 1	r = 5	r = 10	r = 20
RPLM	15.00	- -	42.00	54.00
ME	17.90	38.30	48.90	59.00
ESP	14.40	32.90	43.10	55.10
MSP	26.70	46.00	54.30	66.30
MESP	**28.50**	**47.50**	**56.50**	**67.20**

The size of the gallery set is 649

cameras in campus scenes. Table 8 describes the experimental configuration for training and test. Sample images of the three datasets are demonstrated in Fig. 10.

For 3DPES, we follow the same protocol with the work of Xiong et al. (2014) and Paisitkriangkrai et al. (2015), where the images of 96 persons are used for training and the images of other 96 persons are used for testing. The experiments on CUHK03 are with the same protocol of Li et al. (2014). That is, the dataset was partitioned into a training set of 1160 persons and a test set of 100 persons. Market-1501 provides a different evaluation scheme. It consists of three parts: the training set, the query set and the test set. During testing, the query set is used as the probe set and the test set is used as gallery set. As the gallery set may has multiple images for a person, the top-n matching rate here indicates the expectation

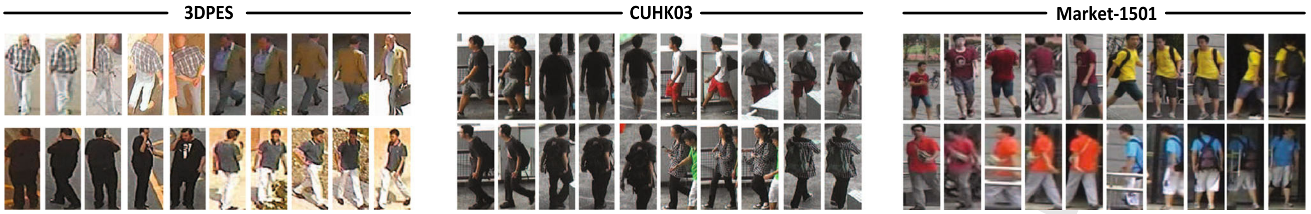

Fig. 10 Example images from 3DPES, CUHK03 and Market-1501. Each person has multiple images

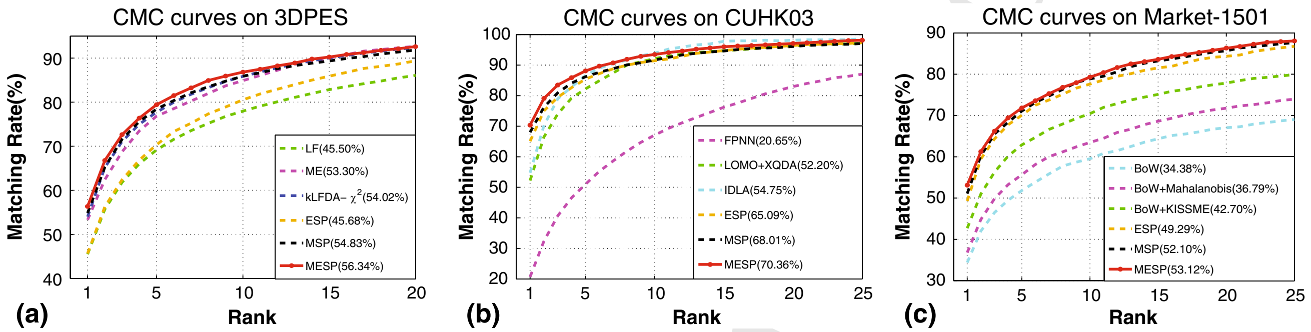

Fig. 11 CMC curves for comparison with our variants and other state-of-the-art methods on (**a**) the 3DPES dataset with 96 gallery images, **b** the CUHK03 dataset with 100 gallery images, **c** the Market-1501 dataset with 19732 gallery images of 750 persons

Table 12 The rank-n matching rates (%) for comparison with LF (Pedagadi et al. 2013), ME (Paisitkriangkrai et al. 2015) and kLFDA-χ^2(Xiong et al. 2014) on the 3DPES dataset, with FPNN (Li et al. 2014), LOMO+XQDA (Liao et al. 2015) and IDLA (Ahmed et al. 2015) on the CUHK03 dataset, with BOW based methods (Zheng et al. 2015) including the method using KISSME (Köstinger et al. 2012) on the Market-1501 dataset

3DPES	(p = 96)				CUHK03	(p = 100)				Market-1501 (19732 gallery images of p=750)			
Method	r = 1	r = 5	r = 10	r = 20	Method	r = 1	r = 5	r = 10	r = 20	Method	r = 1	r = 10	mAP
LF	45.50	69.18	77.98	86.06	FPNN	20.65	51.00	67.08	82.98	BOW	34.38	59.53	14.10
ME	53.30	76.81	85.11	92.78	LOMO+XQDA	52.20	82.23	92.14	96.25	+Mahalanobis	36.79	63.48	15.08
kLFDA-χ^2	54.02	77.74	85.86	92.38	IDLA	54.75	86.61	**93.90**	**98.14**	+KISSME	42.70	70.46	19.55
ESP	45.68	70.36	80.54	89.38	ESP	65.09	85.42	91.45	96.41	ESP	49.29	77.61	23.15
MSP	54.83	78.45	85.90	91.84	MSP	68.01	86.19	91.68	96.11	MSP	51.10	78.86	25.47
MESP	**56.34**	**79.44**	**86.83**	**92.62**	MESP	**70.36**	**88.09**	93.49	97.11	MESP	**53.12**	**79.31**	**26.69**

EPS, MSP and MESP are different configurations of our method

of finding any one of the correct matched images. Besides, the dataset also utilizes mAP criterion to evaluate the overall performance.

Because each person contains multiple images, the training will become very expensive if we take every image as the probe image in turn and formulate multiple hinge losses for one person. To reduce the computational burden, we take the mean image descriptors of a person for $\mathbf{x}_n$. In this way, the number of hinge loss in $L(\mathcal{W}, \mathcal{E})$ equals the identity number in the training set, making the training of large datasets such as CUHK03 and Market-1501 be feasible. Besides, the exemplars are also selected from the mean descriptors, which is slightly different from the datasets where each person has two images for training.

The results are shown in Fig. 11 and Table 12. Again, our approach is comparable to or outperforms current state-of-the-art methods. By comparing the results of MESP and ESP, we find that the effectiveness of multi-cue collaboration on large dataset (CUHK03 and Market-1501) is not as significant as that on small dataset(3DPES). We attribute this phenomenon to the improvement of the PCA-reduced descriptors. As much more images are available for the training of PCA, the reduced descriptors rely less on the way to extract the raw features but can summarize more inherent variability of the data in an unsupervised manner.

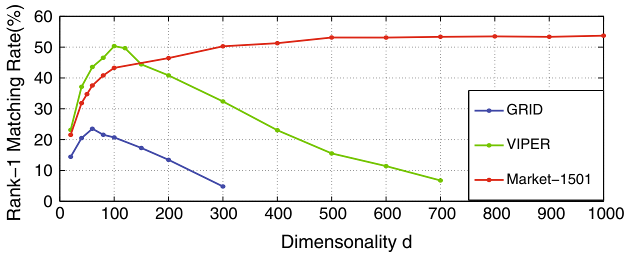

Fig. 12 Comparison of the performance change w.r.t. the PCA-reduced dimensionality d on the datasets with different amounts of training data. Among them, GRID, VIPER and Market-1501 have 250 images of 125 persons, 632 images of 316 persons and 12,936 images of 751 persons for training respectively

6.5 Other Properties

6.5.1 Relationship Between Training Data Size and the Optimal Dimensionality d.

The size of training data significantly influences performance of our method. Generally, more training data not only can improve the quality of PCA-reduced descriptors but also alleviates the overfitting of discriminative similarity learning. The optimal dimension d heavily depends on the training set size. In Fig. 12, we demonstrate the rank-1 matching rate with respect to d on GRID, VIPER and Market-1501, which have 250 images of 125 persons, 632 images of 316 persons and 12,936 images of 715 persons for training, respectively. It can be seen that more training data generally prefers a higher dimensionality d.

6.5.2 Cross-Datasets Experiments

To verify the generalization ability of our method, we consider the situation where the training samples and test samples are selected from different datasets. The experiments are conducted over three datasets including VIPER, GRID and 3DPES, where the images are with different resolutions, backgrounds and illumination conditions. We choose 632, 250 and 192 image pairs from these datasets and propose three groups of evaluations. For each group, two datasets are used for training, and the remaining one is for testing. At the testing stage, we select half of the test set for one evaluation and run 10 trails to obtain the average results, thus the gallery set of the three group are 316, 125 and 96, respectively.

We first compare our method with degraded variants. Among them, F1, F2, Semi-0 and L2R are the same as the variants introduced in Sect. 6.2 and are trained with visual cue Cue_1. The results in Table 13 verify the effectiveness of the polynomial kernel feature map (by comparing F1, F2 and ESP) and the effectiveness of the regularization strategies (by comparing Semi-0, L2R and ESP). Besides, from the comparison among ESP, MSP and MESP, we can see that the

Table 13 The rank-n matching rates (%) for the cross-dataset evaluations

VIPER	($p = 316$)				GRID	($p = 125$)				3DPES	($p = 96$)			
Trained on GRID and 3DPES with 442 pairs					Trained on VIPER and 3DPES with 824 pairs					Trained on VIPER and GRID with 882 pairs				
Method	r = 1	r = 5	r = 10	r = 20	Method	r = 1	r = 5	r = 10	r = 20	Method	r = 1	r = 5	r = 10	r = 20
F1	12.47	25.66	32.44	42.03	F1	27.60	52.32	61.20	73.12	F1	61.87	80.21	86.04	92.60
F2	11.61	24.78	33.03	41.80	F2	26.16	45.44	54.64	69.76	F2	47.29	73.96	81.25	89.48
Semi-0	13.96	26.33	33.20	43.64	Semi-0	28.72	50.88	62.48	74.72	Semi-0	54.79	76.67	84.38	92.81
L2R	15.19	27.18	33.64	44.27	L2R	29.04	52.40	62.08	76.00	L2R	61.25	79.79	88.33	94.69
KISSME	10.79	22.53	30.98	41.42	KISSME	25.92	50.48	62.96	77.92	KISSME	31.46	63.33	75.83	88.54
MFA-χ^2	11.30	26.71	36.42	48.16	MFA-χ^2	25.12	55.84	70.40	82.32	MFA-χ^2	61.25	81.04	87.40	91.15
XQDA	14.05	28.80	37.63	50.35	XQDA	30.08	54.80	63.92	74.16	XQDA	67.08	87.50	91.46	95.00
M-KISSME	15.98	31.84	41.30	51.93	M-KISSME	32.64	57.36	68.88	81.76	M-KISSME	53.85	79.58	91.77	**96.98**
M-MFA-χ^2	12.78	28.48	38.16	50.06	M-MFA-χ^2	29.12	55.68	70.40	82.48	M-MFA-χ^2	65.31	82.92	88.96	93.44
M-XQDA	14.11	29.53	39.72	51.27	M-XQDA	29.04	50.48	64.00	76.40	M-XQDA	69.58	86.46	89.79	94.79
ESP	15.95	29.72	36.80	45.70	ESP	31.28	56.00	65.20	78.24	ESP	65.42	80.31	86.98	91.98
MSP	19.81	36.14	44.08	53.07	MSP	40.46	66.16	75.36	84.72	MSP	73.33	91.77	**95.00**	96.46
MESP	**20.98**	**36.36**	**44.59**	**54.53**	MESP	**42.96**	**66.72**	**76.32**	**85.04**	MESP	**75.10**	**92.08**	94.58	96.87

multiple cue collaboration still plays an important role. The exemplar-guided mixture model can further improve the final accuracy.

We also compare with other metric learning methods including KISSME (Köstinger et al. 2012), MFA-χ^2 (Xiong et al. 2014) and XQDA (Liao et al. 2015) using a single visual cue and multiple visual cues. On the single cue situation, ESP performs quite well on rank-1 and rank-5 matching rate but less better than MFA-χ^2 and XQDA on the rank-10 and rank-20 matching rate. Such insufficiency can be compensated by utilizing multi-cue collaboration. The final results indicate that our method can better exploit the complementary properties of different visual cues.

6.5.3 Runtime

Our method is implemented in MATLAB/MEX with a 3.07Ghz, 2 Cores CPU. As the runtime varies according to different datasets, we take VIPER as an example. It takes about 0.02 second(s) to extract the raw features from a 128 × 48 person image. At the training stage, our method takes about 55.84 s to learn 4 PCA projection matrices from 632 training images (each for one visual cue), 20.68 s to generate both positive and negative polynomial feature maps for 316 persons, 25.88 s for the ADMM iterations and 2 s for exemplar estimation procedure. At the testing stage, it requires 0.016 s to rank 316 gallery images for a probe image. Note that we don't need to explicitly generate polynomial kernel feature map for testing, because the similarity function can be decomposed into basic similarity measurements according to Eq. (9). By pre-computing the coefficients $\sum_{k=0}^{K} w_n^{c,k}\mathbf{W}^{c,k}$ for $\mathbf{x}_n$, the testing cost is only linear with respect to C.

7 Conclusion

We have presented a novel similarity learning approach for person re-identification. The success of our approach stems from a novel organization of explicit polynomial kernel feature map, which takes the advantages of both Mahalanobis distance and bilinear similarity. It also benefits from the exemplar-guided similarity function, which not only employs a set of sub-functions to breakdown the variability of positive image pairs but also makes use of multiple visual cues. Besides, two regularization strategies also increase the generalization ability of our approach.

Results are promising. We have thoroughly studied the gain of each component and analyzed the influences of the parameters. The proposed method performed state-of-the-art results over six datasets, and showed its effectiveness during the cross-dataset experiments where the training samples and test samples are from different datasets.

Acknowledgements This work was supported by the National Key Research and Development Program of China (No. 2016YFB1001001), the National Basic Research Program of China (No. 2015CB351703, No. 2012CB316400), the National Natural Science Foundation of China (No. 61573280, No. 91648121, No. 61603022, No. 61573273).

8 Dual Form Deviation for Updating $\mathbf{U}_1$

In this section, we will give the deviation from Eqs. (31) to (35). As $g_1(\mathbf{U}_1) = L(\mathbf{U}_1) = \frac{1}{N}\sum_{n=1}^{N}[1 - \langle \Omega(\mathbf{x}_n), \mathbf{U}_1\rangle]_+$, the problem of Eq. (31) equals:

$$\begin{aligned} &\min_{\mathbf{U}_1,\xi} \ \frac{\rho}{2}\|\mathbf{U}_1 - \left(\mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l\right)\|_F^2 + \frac{1}{N}\sum_{n=1}^{N}\xi_n \\ &\text{s.t.} \quad 1 - \langle \Omega(\mathbf{x}_n), \mathbf{U}_1\rangle \le \xi_n, \quad \xi_n \ge 0, \quad \forall n. \end{aligned} \tag{40}$$

The Lagrangian associated with Eq. (40) is:

$$\begin{aligned} La_1(\mathbf{U}_1, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{\rho}{2}\|\mathbf{U}_1 - \left(\mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l\right)\|_F^2 + \frac{1}{N}\mathbf{1}^\top\boldsymbol{\xi} \\ &+ \sum_{n=1}^{N}(1 - \langle \Omega(\mathbf{x}_n), \mathbf{U}_1\rangle_F)\alpha_n - \boldsymbol{\alpha}^\top\boldsymbol{\xi} - \boldsymbol{\beta}^\top\boldsymbol{\xi}, \end{aligned} \tag{41}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are dual variable vectors with non-negative elements, and α_n is the n-th element of $\boldsymbol{\alpha}$. The Lagrangian dual is $h_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{U}_1,\boldsymbol{\xi}} La_1(\mathbf{U}_1, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, with $\boldsymbol{\alpha} \succeq 0$ and $\boldsymbol{\beta} \succeq 0$ being the dual feasible. We solve $\mathbf{U}_1$ and $\boldsymbol{\xi}$ from the optimality condition:

$$\begin{aligned} \frac{\partial La_1}{\partial \mathbf{U}_1} &= \rho\left(\mathbf{U}_1 - \mathbf{U}_3^l + \boldsymbol{\Lambda}_1^l\right) - \sum_{n=1}^{N}\Omega(\mathbf{x}_n)\alpha_n = \mathbf{0} \\ \frac{\partial La_1}{\partial \boldsymbol{\xi}} &= \frac{1}{N}\mathbf{1}^\top - \boldsymbol{\alpha}^\top - \boldsymbol{\beta}^\top = \mathbf{0}. \end{aligned} \tag{42}$$

Taking Eq. (42) into Eq. (41), $h_1(\boldsymbol{\alpha}, \boldsymbol{\beta})$ becomes:

$$\begin{aligned} h_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = &-\frac{1}{2\rho}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\langle \Omega(\mathbf{x}_i), \Omega(\mathbf{x}_j)\rangle_F \\ &+ \mathbf{1}^\top\boldsymbol{\alpha} - \sum_{n=1}^{N}\left\langle \Omega(\mathbf{x}_n), \mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l\right\rangle_F \alpha_n. \end{aligned} \tag{43}$$

The dual variable vector $\boldsymbol{\beta}$ is eliminated. We define the kernel matrix $\mathbf{H}$ with each element $H_{i,j} = \langle \Omega(\mathbf{x}_i), \Omega(\mathbf{x}_j)\rangle_F$ and define a vector $\mathbf{b}$ with $b_n = \langle \Omega(\mathbf{x}_n), \mathbf{U}_3^l - \boldsymbol{\Lambda}_1^l\rangle_F - 1$, the dual function becomes:

$$h(\boldsymbol{\alpha}) = -\frac{1}{2\rho}\boldsymbol{\alpha}^\top\mathbf{H}\boldsymbol{\alpha} - \mathbf{b}^\top\boldsymbol{\alpha} \tag{44}$$

Also because $\frac{1}{N}\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\alpha} \succeq 0$, α_n belongs to the domain $[0, \frac{1}{N}]$. We therefore obtain the standard quadratic programming problem of Eq. (35). With optimal $\boldsymbol{\alpha}^*$, Eq. (36) is obtained according to Eq. (42).

9 A Separated ADMM Algorithm for Updating $\mathbf{U}_3$

In this section, we present a separated ADMM algorithm for the update of $\mathbf{U}_3$. Directly projecting a non-symmetric matrix onto $\mathbb{S}_-^d$ is difficult, we therefore introduce an auxiliary set $\mathcal{C}' = \{\mathbf{U}|\mathbf{W}_M^{c,k} \in \mathbb{S}^d, \forall c, \forall k\}$, and define $g_3'(\mathbf{E}) = \infty\delta[\mathbf{E} \in \mathcal{C}']$. As $\mathcal{C} \subset \mathcal{C}'$, the sub-problem of Eq. (33) is equivalent to:

$$\begin{aligned} \min_{\mathbf{E},\mathbf{F}} \quad & g_3'(\mathbf{E}) + \rho\|\mathbf{E} - \mathbf{B}\|_F^2 + g_3(\mathbf{F}) \\ \text{s.t.} \quad & \mathbf{E} = \mathbf{F}, \end{aligned} \tag{45}$$

where $\mathbf{B} = \frac{1}{2}(\mathbf{U}_1^{l+1} + \mathbf{U}_2^{l+1} + \boldsymbol{\Lambda}_1^l + \boldsymbol{\Lambda}_2^l)$, and $\mathbf{U}_3^{l+1}$ equals the optimal $\mathbf{E}$ or $\mathbf{F}$. By introducing Lagrange multipliers $\mathbf{G}$, we obtain the augmented Lagrangian:

$$\begin{aligned} La_2(\mathbf{E}, \mathbf{F}, \mathbf{G}) = {} & g_3'(\mathbf{E}) + \rho\|\mathbf{E} - \mathbf{B}\|_F^2 + g_3(\mathbf{F}) \\ & + \rho'\langle \mathbf{G}, \mathbf{E} - \mathbf{F}\rangle_F + \frac{\rho'}{2}\|\mathbf{E} - \mathbf{F}\|_F^2, \end{aligned} \tag{46}$$

where ρ' is a scaling parameter for this ADMM. The ADMM algorithm consists of the following iterations.

$$\mathbf{E}^{k+1} = \arg\min_{\mathbf{E}} g_3'(\mathbf{E}) + \rho\|\mathbf{E} - \mathbf{B}\|_F^2 + \frac{\rho'}{2}\|\mathbf{E} - \mathbf{F}^k + \mathbf{G}^k\|_F^2, \tag{47}$$

$$\mathbf{F}^{k+1} = \arg\min_{\mathbf{F}} g_3(\mathbf{F}) + \frac{\rho'}{2}\|\mathbf{F} - \mathbf{E}^{k+1} - \mathbf{G}^k\|_F^2, \tag{48}$$

$$\mathbf{G}^{k+1} = \mathbf{G}^k + \mathbf{E}^{k+1} - \mathbf{F}^{k+1}. \tag{49}$$

The problem of Eq. (47) is equivalent to:

$$\begin{aligned} \mathbf{E}^{k+1} = {} & \arg\min_{\mathbf{E}} g_3'(\mathbf{E}) + \\ & \frac{2\rho + \rho'}{2}\|\mathbf{E} - \left(\frac{2\rho}{2\rho + \rho'}\mathbf{B} + \frac{\rho'}{2\rho + \rho'}(\mathbf{F}^k - \mathbf{G}^k)\right)\|_F^2. \end{aligned} \tag{50}$$

Equation (50) indicates projecting $\left(\frac{2\rho}{2\rho+\rho'}\mathbf{B} + \frac{\rho'}{2\rho+\rho'}(\mathbf{F}^k - \mathbf{G}^k)\right)$ onto the set $\mathcal{C}'$. More specifically, we project the sub-matrices corresponding to $\mathbf{W}_M^{c,k}$ to be symmetric by the function $f(\mathbf{W}) := \frac{1}{2}(\mathbf{W} + \mathbf{W}^\top)$. $\mathbf{G}^k$ is initialized as zero matrix and it stays in $\mathcal{C}'$ during the updating of Eq. (49), therefore, $(\mathbf{E}^{k+1} + \mathbf{G}^k) \in \mathcal{C}'$. Equation (48) indicates projecting $\left(\mathbf{E}^{k+1} + \mathbf{G}^k\right)$ onto set $\mathcal{C}$. It needs to project the corresponding sub-matrices from $\mathbb{S}^d$ to $\mathbb{S}_-^d$. The projection can be efficiently obtained by cropping the positive eigenvalues to be zero (Boyd and Vandenberghe 2004).

We operate $K = 10$ iterations for the separated ADMM, and $\mathbf{U}_3^{l+1}$ is obtained by $\mathbf{F}^K$.

References

Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In: *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Bak, S., Corvée, E., Brémond, F., & Thonnat, M. (2010). Person re-identification using spatial covariance regions of human body parts. In: *International conference on advanced video and signal based surveillance*.

Baltieri, D., Vezzani, R., & Cucchiara, R. (2011). Sarc3d: A new 3D body model for people tracking and re-identification. In: *International conference on image analysis and processing*.

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge university press.

Boyd, S. P., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Chen, D., Yuan, Z., Hua, G., Zheng, N., & Wang, J. (2015). Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: *Conference on computer vision and pattern recognition*.

Chen, D., Yuan, Z., Chen, B., & Zheng, N. (2016). Similarity learning with spatial constraints for person re-identification. In: *Conference on computer vision and pattern recognition*.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., & Murino, V. (2011). Custom pictorial structures for re-identification. In: *British machine vision conference*.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In: *Computer vision and pattern recognition*.

Ding, S., Lin, L., Wang, G., & Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, *48*(10), 2993–3003.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In: *Computer vision and pattern recognition*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*, 2007.

Gong, S., Cristani, M., Yan, S., & Loy, C. C. (Eds.). (2014). *Person re-identification. Advances in computer vision and pattern recognition*. Berlin: Springer.

Gray, D., Brennan, S., & Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In: *International Workshop on PETS, Rio de Janeiro*.

Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In: *Scandinavian conference on image analysis* (pp. 91–102).

Hirzer, M., Roth, P. M., Kostinger, M., & Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. In: *European conference on computer vision*.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321–377.

Jégou, H., & Chum, O. (2012). Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In: *European conference on computer vision*.

Kjems, U., Hansen, L.K., & Strother, S. C. (2000). Generalizable singular value decomposition for ill-posed datasets. In: *Neural information processing systems* (pp. 549–555).

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In: *Computer vision and pattern recognition*.

Li, W., & Wang, X. (2013). Locally aligned feature transforms across views. In: *Computer vision and pattern recognition*.

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Li, Y., Wu, Z., & Radke, R.J. (2015). Multi-shot re-identification with random-projection-based random forests. In *Winter conference on applications of computer vision*.

Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., & Smith, J.R. (2013). Learning locally-adaptive decision functions for person verification. In *Computer vision and pattern recognition*.

Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., & Li, S.Z. (2010). Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Conference on computer vision and pattern recognition* (pp. 1301–1306).

Liao, S., Hu, Y., Zhu, X., & Li, S.Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Loy, C.C., Liu, C., & Gong, S. (2013). Person re-identification by manifold ranking. In *International conference on image processing*.

Ma, B., Su, Y., & Jurie, F. (2012a). Bicov: A novel image representation for person re-identification and face verification. In *British machine vision conference*.

Ma, B., Su, Y., & Jurie, F. (2012b). Local descriptors encoded by fisher vectors for person re-identification. In *ECCV workshops and demonstrations*.

Ma, L., Yang, X., & Tao, D. (2014). Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, *23*(8), 3656–3670.

Mignon, A., & Jurie, F. (2012). PCCA: A new approach for distance learning from sparse pairwise constraints. In *Computer vision and pattern recognition*.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.

Paisitkriangkrai, S., Shen, C., & van den Hengel, A. (2015). Learning to rank in person re-identification with metric ensembles. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Pedagadi, S., Orwell, J., Velastin, S.A., & Boghossian, B.A. (2013). Local fisher discriminant analysis for pedestrian re-identification. In *Computer vision and pattern recognition*.

Prosser, B., Zheng, W., Gong, S., & Xiang, T. (2010). Person re-identification by support vector ranking. In *British machine vision conference*.

Schwartz, W.R., & Davis, L.S. (2009). Learning discriminative appearance-based models using partial least squares. In *XXII Brazilian symposium on computer graphics and image processing*.

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *European conference on computer vision*.

Wu, Z., Li, Y., & Radke, R. J. (2015). Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(5), 1095–1108.

Xiong, F., Gou, M., Camps, O.I., & Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*.

Xu, Y., Lin, L., Zheng, W., & Liu, X. (2013). Human re-identification by matching compositional template with cluster sampling. In *International conference on computer vision* (pp. 3152–3159).

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., & Li, S.Z. (2014). Salient color names for person re-identification. In *European conference on computer vision*.

Zhang, Z., Chen, Y., & Saligrama, V. (2014). A novel visual word co-occurrence model for person re-identification. In: *ECCV workshops*.

Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Computer vision and pattern recognition*.

Zhao, R., Ouyang, W., & Wang, X. (2014). Learning mid-level filters for person re-identification. In *Conference on computer vision and pattern recognition*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Bu, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *International conference on computer vision*.

Zheng, W., Gong, S., & Xiang, T. (2009). Associating groups of people. In *British machine vision conference*.

Zheng, W., Gong, S., & Xiang, T. (2013). Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(3), 653–668.

Journal: 11263
Article: 991

Springer
the language of science

Author Query Form

Please ensure you fill out your response to the queries raised below and return this form along with your corrections

Dear Author

During the process of typesetting your article, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query	Details required	Author's response
1.	Please confirm if the inserted city name is correct of the affiliations 1 and 2. Amend if necessary.	
2.	Please confirm if the corresponding author and affiliation is correctly identified. Amend if necessary.	
3.	Please check and confirm if the inserted citation of Figs 1, 2, 3 is correct. If not, please suggest an alternate citation. Please note that Figures should be cited sequentially in the text.	
4.	Please note that part figure label (f) present in artwork, but there is no such description in Fig. 4 caption. Kindly confirm.	
5.	Please provide a definition for the significance of bold in the Tables 5, 9, 10, 11, 12, 13	
6.	Please check and confirm if the inserted citation of Tables 7 and 8 is correct. If not, please suggest an alternate citation. Please note that Tables should be cited sequentially in the text.	
7.	Please check that the inserted publisher location is correctly identified for the reference Gong et al. (2014).	