Journal of Computers Vol. 31 No. 2, 2020, pp. 114-126
doi:10.3966/199115992020043102011

Novel Feature Representation and Enhanced Metric Learning for Person Re-identification

Zhuochen Lei[1,2*], Wanggen Wan[1,2], Xiaoqing Yu[1,2]

[1] School of Communication & Information Engineering, Shanghai University, Shanghai, China

[2] Institute of Smart City, Shanghai University, Shanghai, China
leozoson@yeah.net, wanwg@staff.shu.edu.cn, yxq@staff.shu.edu.cn

Received 19 June 2018; Revised 19 October 2018; Accepted 18 December 2018

Abstract. Person re-identification (re-ID) includes two vital parts: feature representation and metric learning. However, person re-ID is affected by cross-camera factors such as illumination, pose and viewpoint variations. In this paper, we propose a novel feature called Global Maximal Occurrence (GOMO) based on Local Maximal Occurrence (LOMO), and an enhanced metric learning method. Our proposed feature not only maximizes horizontal occurrence of local features but also combines the maximum occurrence of vertical and horizontal features. To handle vertical feature misalignment between the cross-cameras, we maximize the maximum occurrence of the vertical direction and then cascade features in two directions as a fusion feature. Besides, histogram equalization is applied to enhance images for overcoming cross-camera variations. To reduce the errors between camera views caused by the outliers in final distance matrix of traditional metric learning, we propose a re-ranking method using the max-min criteria that linearly map all values in final distance matrix to a specified subspace to enhance the Cross-view Quadratic Discriminant Analysis (XQDA) method. We conducted experiments on two challenging person re-ID datasets, VIPeR and CUHK Campus, and the experimental results demonstrate the effectiveness of our proposed feature and the superiority of our enhanced metric learning method over the state-of-the-art methods.

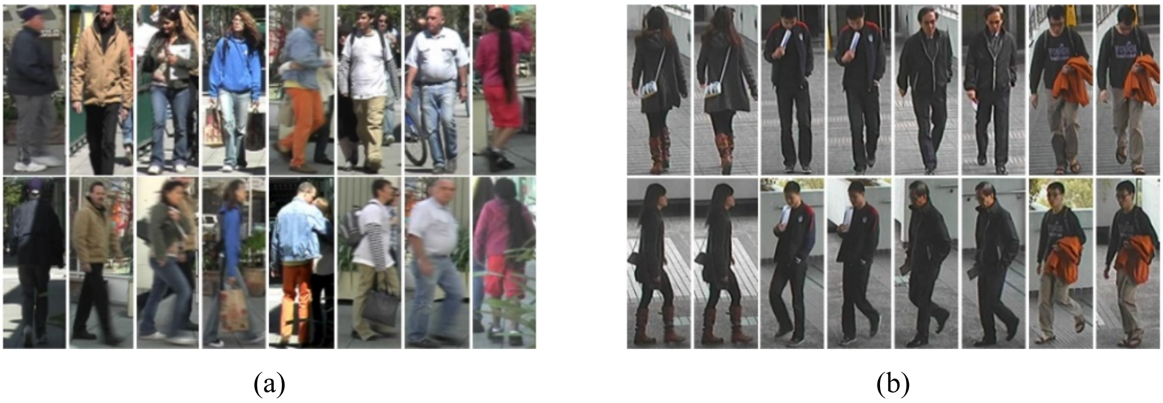
Keywords: enhanced metric, max-min criteria, novel feature, person re-ID

1 Introduction

Person re-ID involves the situation in which the pedestrian appearance does not change too much and is devoted to determining whether pedestrian image pairs in different camera views are the same or not based on appearance features. Person re-ID is widely used in video surveillance, such as criminal investigation and person retrieval etc. However, there are illumination, pose, viewpoint variations and occlusions under cross-camera views. For example, Fig. 1 shows some example images from VIPeR dataset [1] and CUHK Campus dataset [2]. Images in the same column represent the same pedestrian, and images at the first and second row respectively represent pedestrians in camera *A* and camera *B*. It can be clearly seen that pedestrian images have variations on illumination, pose and viewpoint under cross-camera views. In order to handle these issue, the existing methods have been proposed in varying degrees [3-4], and have continuously promoted the rapid and high-precision development of person re-ID. To this end, person re-ID is still an open issue so far.

With the continuous development in recent years, person re-ID mainly focuses on the following aspects: (1) there are always many variations between pedestrian images under cross-camera views, such as illumination, pose and viewpoint etc. Thereby, how to represent a robust feature that can overcome the variations between pedestrians under cross-camera views? (2) Because there are many low resolution images captured by cameras in the current video surveillance field, many mature technologies such as the face, gait and other biological features cannot be used. Therefore, image pre-processing plays an

[*] Corresponding Author



(a) (b)

Fig. 1. (a) and (b) respectively represent some example images from VIPeR dataset and CUHK Campus dataset

important role in the process of feature representation and has been widely applied by person re-ID researchers. (3) Besides feature extraction, metric learning is also a crucial part of person re-ID, and is used to determine whether the pedestrian images in different camera views are the same or not. An ideal metric learning greatly reduces the differences between classes and improves the matching accuracy effectively. In summary, the research of person re-ID related fields still focuses on two categories: feature representation and metric learning. In this paper, we thus propose a novel feature and an enhanced metric to handle variations issue in person re-ID.

The main purpose of feature representation is to represent an ideal feature that overcomes illumination, pose, viewpoint variations and occlusions under cross-camera views, and is invariant with the significant variations of camera viewpoints and settings. Many effective feature representation methods [5-9] have been proposed for person re-ID, which learn invariant features such as visual salience [6], texture features [8] to achieve the maximum similarity discrimination between pairs of pedestrian under cross-camera views. Some excellent feature representation methods include novel feature representation [7-9], local feature representation [6-7, 10], scale invariant local ternary patterns [8, 11], color histograms [6, 8, 10] and so on. Many deep learning methods [12-14] have also been used to extract features with the popularity of machine learning in recent years. These works proposed for person re-ID are local-based features mostly, so there will be information lose more or less. Different from traditional feature representation, deep learning-based features are favored by more and more person re-ID researchers for automatic and ideal features. However, complicated feature model training and limited large-scale datasets make feature representation based on deep learning difficult. In summary, although these proposed features have greatly advanced person re-ID actually, they are not designed to address changes between pairs of images in the global view and image pre-processing is ignored. Motivated by these issues, we intend to design a robust global feature which contains more information than local-based feature and combines image pre-processing process previously.

Without being significantly affected by what kind of feature representation method is used, metric learning methods learn a suitable and non-Euclidean distance/similarity metric that can optimize the matching results for person re-ID. Of course, a metric learning method will have better performance if it is combined with a more robust feature. A feature transformation function in specific feature space is learned through the training process, which makes the features from same class closer, and further separates different classes. To this end, it is necessary to learn a function that can be used to summarize the invariant transition patterns between pedestrian pairs that have significant variations under cross-camera views. Many excellent metric learning methods have been proposed with long-term development in person re-ID, such as PRDC [15], LADF [16], KISSME [17], XQDA [8], and DNS [18]. Even the accuracy of matching results is increasing with the development of person re-ID methods, it is also a long-term research for actual application. In this paper, we have found outliers that affect match accuracy in distance matrix learned by metric learning method. Therefore, we are committed to use related mathematical theory to tackle this problem and improve the accuracy of metric learning.

The main contribution of this paper can be summarized as follows.

We propose a novel global-based feature called GOMO that is based on LOMO [8]. The LOMO feature is widely used by some latest person re-ID methods, which divides an images into small pieces and then extracts features for each small piece. Next, it maximizes the features of all pieces in the horizontal direction, so that the most stable horizontal features of an image can be represented. As an improved feature representation, the GOMO extracts the most stable vertical features, and specially maximizes those features in the horizontal direction to solve the problem that the original stable vertical features will shift with the various pedestrian positions under cross-camera views. Then it combines vertical and horizontal features as a global fusion feature that effectively handles illumination, pose and viewpoint variations under cross-camera views. To the best of our knowledge, this is a pioneer work to address cross-camera variations issue by employing a global-based feature if it is compared with widely used local-based features in related works.

(1) Besides, we consider more about image pre-processing in feature extraction, and apply the histogram equalization to enhance the images for effectively solving the illumination variations in person re-ID [19-21]. The histogram equalization technique proposed in this paper is an effective image pre-processing method, which can solve the problems in person re-ID datasets such as low pixel, contrast difference and so on, and it can provide ideas for follow-up researchers.

(2) In order to reduce the cross-camera errors caused by the outliers in final distance matrix of traditional metric learning, we propose a re-ranking method using max-min criteria that linearly map all values in final distance matrix to the specified subspace. Especially, the re-ranking method is combined with XQDA [8] as an enhanced metric learning method in this paper. The XQDA metric applies KISSME metric in a learned discriminant subspace, which can effectively overcome cross-camera variations.

(3) We conducted experiments on two challenging person re-ID datasets, VIPeR and CUHK Campus, and the experimental results demonstrate the effectiveness of our proposed feature and the superiority of our enhanced metric learning method over the state-of-the-art methods.

2 Related Work

In this section, we briefly review two categories of works that are related to our method: (1) Feature representation methods. (2) Metric learning methods.

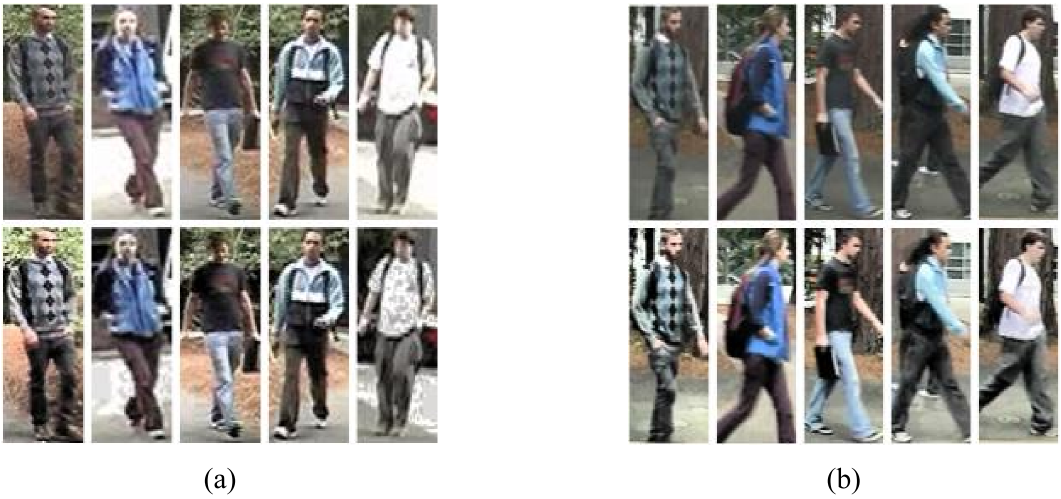
An ideal feature representation should be invariant with variations on illumination, pose and viewpoint from one camera view to the other. Therefore, the features based on hand-crafted appearance pattern are usually used to describe person re-ID images, which are designed and computed as human domain knowledge [5-10, 22-23]. In order to tackle cross-camera variations, many robust features have been proposed [5, 7-8, 10, 13, 22-23]. These features are different types of descriptors such as color histograms [8, 10], textures [8, 10], patch-based features [8, 10, 22], novel features [7, 22] and deep features [13]. Farenzena et al. [22] proposed the SDALF feature, which uses symmetry and asymmetry perceptual principles to extract features from different human body parts. Zhao et al. [10] proposed a salience learning method based on extracted distinctive features in an unsupervised manner. Liao et al. [8] proposed the LOMO feature, which maximizes the horizontal occurrence of local features to obtain a robust feature. Matsukawa et al. [7] proposed a hierarchical Gaussian descriptor, which improves the hierarchical covariance descriptor with modeling both the mean and the covariance information of pixel features properly. Recently, Chen et al. [5] proposed the CRAFT framework, which learns a new adaptive space to transform the original features. Moreover, the convolutional network itself has high discriminative ability without explicit patch-matching, which is also used for feature extraction [14, 26-27]. Without large enough datasets to learn a robust feature representation, Xiao et al. [26] proposed Domain Guided Dropout algorithm, which uses Convolutional Neural Networks (CNNs) to learn deep feature representation with data from multiple domains. The major differences between our approach and descriptors mentioned above are as follows: (1) compared with traditional descriptors, we propose a global-based feature that considers more about how to design a novel feature with as more feature information as possible. Besides, image pre-processing is also considered as a novel idea for feature design. (2) Compared with deep features, our proposed novel feature is not limited by specific datasets and does not require complicated training models.

Metric learning is another crucial part of person re-ID and an ideal it can match correct pedestrian pairs in large samples with only few samples training. Furthermore, metric learning has the same problems that feature representation faced with, such as variations on viewpoint and pose etc. So it is also a challenging task for metric learning to improve matching accuracy between pedestrian pairs under cross-camera views. However, the matching accuracy of metric learning has been improved gradually by recent methods [15-18, 24-25]. Based on decision-making way, Zheng et al. [15] proposed the PRDC algorithm, which uses a relative distance comparison model to make the distance between true matches closer. Li et al. [16] proposed the LADF algorithm, which learns a decision function for verification that can be viewed as a joint model of a distance metric and a locally adaptive thresholding rule. Based on a statistical inference perspective, Kostinger et al. [17] proposed the KISSME metric, which introduces a simple but effective strategy to learn a distance metric from equivalence constraints. Further, Liao et al. [8] proposed the XQDA algorithm, which uses the KISSME metric in a learned discriminant subspace to overcome variations under cross-camera views. Aside from the methods using Mahalanobis distance, Zhang et al. [18] proposed the learning a Discriminative Null Space (DNS) for person re-ID, which matches people in a discriminative null space of the training data to overcome the small sample size problem in re-ID distance metric learning. The re-ranking method can improve retrieval accuracy [28-30], which is further improving the accuracy of person re-ID based on existing algorithms. Zhong et al. [31] proposed a k-reciprocal encoding method to re-rank the re-ID results. In addition to the traditional metric learning methods, deep learning is also a popular field of metric learning in recent years, which requires further considered by researchers in the future. Arulkumar et al. [12] proposed two CNN-based architectures for person re-ID, one used multiple stages of convolution and pooling for feature extraction, and the other matched pixels across a wider region for verification. The major differences between our approach and metric methods mentioned above are as follows: (1) In this paper, our proposed enhanced metric is a pioneer work using the re-ranking technique to reduce the cross-camera errors caused by the outliers in final distance matrix of traditional metric learning. (2) Different from the deep learning methods, our proposed enhanced metric is not limited by what kind of dataset is used and it can quickly train the evaluation model.

3 Novel Feature Representation

3.1 Histogram Equalization for Illumination Variations

Illumination variations between cross-camera views are mainly due to the different settings and viewpoints of non-overlapping cameras. Based on experience, the different visual colors that we feel in our daily life are affected by illumination variations. In order to better illustrate this issue, we show some example images from VIPeR dataset in Fig. 2(a) and Fig. 2(b), where Fig. 2(a) and Fig. 2(b) respectively show images in camera *A* and camera *B*. It can be seen that there are significant illumination variations in different camera views.



(a) (b)

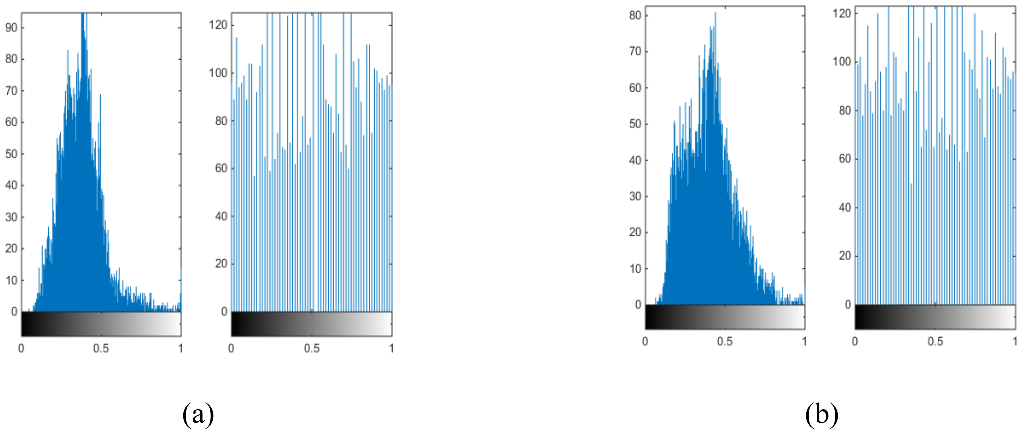
Fig. 2. Illumination variations show at the first row and the histogram equalization processed images show at the second row

In this paper, in order to tackle illumination variations, we apply histogram equalization to improve image contrast [21], so that the illumination intensity of images in different camera views is closer and the illumination variations have the least influence on the matching results. By processing all images in different camera views with histogram equalization method, it can clearly see that images in the second row of Fig. 2(a) and Fig. 2(b) are significantly enhanced with respect to the first row, and illumination variations are more similar under non-overlapping camera views.

Considering more about the visual characteristics of human eyes, the HSV color histogram is used for feature representation, which includes three channels: Hue, Saturation and Value. In order to improve the illumination intensity and maintain the original features of images as much as possible, we specially applied histogram equalization method in Value channel. Wang et al. [20] proposed that if the histogram of an image is uniformly distributed, the image contrast is the largest. So ideally, it is possible to find a transformation function $f(x)$ that makes the histogram uniformly distributed, and the ideal $f(x)$ is defined as

$$f(x)=(L-1)\sum_{0}^{x_i}\frac{h(x_i)}{w\times h}. \tag{1}$$

Where L represents the gray level, $h(x_i)$ represents the number of pixels in each gray level, and w and h are width and height of an image respectively. We apply the histogram equalization technique to process images at the first row of Fig. 2(a) and Fig. 2(b), and the corresponding results are displayed at the second row. Fig. 3(a) and Fig. 3(b) respectively show the pairwise histogram distribution results for images at the first column in Fig. 2(a) and Fig. 2(b). Comparing the left figures of Fig. 3(a) and Fig. 3(b), it clearly shows that the distribution of histograms is quite different in the original Value channel. Furthermore, the results of the corresponding uniform distribution are shown in the right figures, which are processed by histogram equalization.



(a)

(b)

Fig. 3. (a) and (b) show the histogram contrast results of first column in Fig. 2(a) and Fig. 2(b) respectively

In addition to HSV color histogram, the SILTP feature is also used to handle illumination variations that are caused by multiple scales under cross-camera views [11], which is an improved feature over Local Binary Pattern (LBP) [32]. The LBP is a kind of texture feature with significant advantages such as gray and rotation invariance, which is invariant to illumination variations but are susceptible to image noises. Compared with the LBP, the SILTP applies scale-invariant factors to enhance the LBP and overcomes illumination variations at multiple scales.

3.2 GOMO Feature

By comparing the images in different camera views in Fig. 2(a) and Fig. 2(b), it shows that viewpoint changes greatly under different camera views. Therefore, how to represent relatively invariant features from the changed viewpoint under different camera views is the key technique to address illumination, pose and viewpoint changes. The LOMO feature divides an image into sub-windows size of 10×10 with overlapping step of 5 pixels, and each sub-window extracts two scales of SILTP histograms ($\mathrm{SILTP}^{0.3}_{4,3}$ and $\mathrm{SILTP}^{0.3}_{4,5}$) and an 8×8×-bin joint HSV histogram. Then it extracts the maximum histogram of SILTP and HSV features from all the horizontal sub-windows, so that the most stable horizontal features can be represented. However, The LOMO feature has lost much feature information in the process of representing the maximum stable features. Even it seems to be stable in the horizontal direction, the LOMO feature is still not a stable feature representation in the global view.

In this paper, we propose a global feature that combines horizontal and vertical features. Following the LOMO feature representation, we also divide an image into $M \times N$ sub-windows size of 10×10 and overlapping step of 5 pixels, and extract SILTP and HSV histogram for each sub-window. As shown in Fig. 4, the horizontal features of an image are represented as $X^r = \{\chi^r_{p_i} \mid p_i = 1, \dots, N\}$, where (r, p) denotes the p^{th} sub-window in r row, and $\chi^r_{p_i}$ is the corresponding feature representation of each horizontal sub-window. As the same, we use X^c to represent the features in all columns. We extract the maximum feature for each sub-window in horizontal and vertical directions, respectively. Therefore, the most stable feature representation in the horizontal direction is defined as

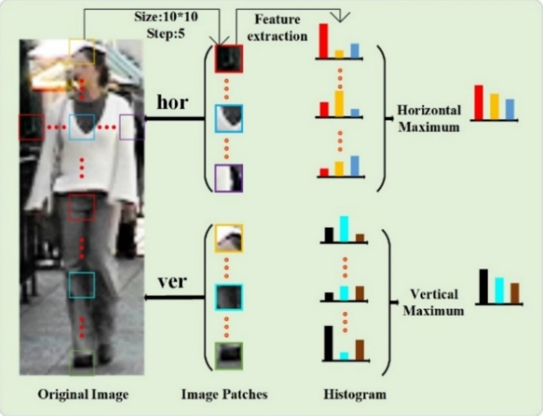


Fig. 4. Illustration of the maximum feature extraction in horizontal and vertical direction

$$X^r_{p_{\max}} = \{\max(\chi^{r_m}_{p_i} \mid p_i = 1, \dots, N) \mid r_m = 1, \dots, M\}. \tag{2}$$

Where M represents the total number of rows, $X^r_{p_{\max}}$ represents the maximum features set of all rows, and $X^c_{p_{\max}}$ is used to represent the maximum features set of all columns. In Fig. 5, it shows that the height of a pedestrian is invariant under different camera views such as original image (a) and (b), so it is easy to extract stable features at the same height of pairwise pedestrian images in different camera views. For example, the maximum features extracted from the same row at height h in different camera views in Fig. 5 are more similar. However, without pedestrian height limitations, there is a misaligned problem in the process of maximum vertical feature representation. At top of the original image (a) and (b) in Fig. 5, it clearly shows that the center feature distribution of the original image (a) but the left feature distribution of the original image (b). In order to address this issue, we once again extract horizontal maximum feature based on the results of vertical maximum feature representation, and it is shown in Fig. 5, and the formula is defined as

$$X^{c_{\max}}_{p_{\max}} = \max(X^{c_n}_{p_{\max}} \mid c_n = 1, \dots, N). \tag{3}$$

Where $X^{c_{\max}}_{p_{\max}}$ represents the most stable feature in the vertical direction actually. Finally, a global fusion feature is represented by cascading the features in two directions.

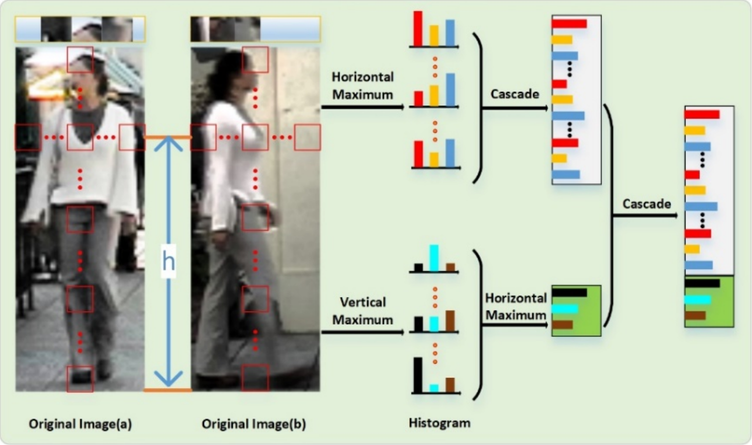


Fig. 5. Illustration of the problem vertical feature extraction faced with and the fused GOMO feature

4 Enhanced Metric Learning

4.1 XQDA

As an extended algorithm of KISSME [17], the XQDA applies KISSME metric in a learned *r* dimensional subspace $W = (w_1, w_2, \ldots, w_r) \in \mathbb{R}^{d \times r}$, where *d* is the original space dimension. Let's overview KISSME metric briefly. The KISSME metric determines whether a pair of images $\hat{x}^A$ and $\hat{x}^B$ in cross-camera views are similar or not in statistical inference point of view. And h_0 and h_1 respectively represent the hypothesis that images pairs are dissimilar $((A, B) = 0)$ and similar $((A, B) = 1)$. Thus, the likelihood ratio can be defined by statistical inference theory as

$$\delta^{(A,B)} = \log\left(\frac{p(\hat{x}^A - \hat{x}^B \mid h_0)}{p(\hat{x}^A - \hat{x}^B \mid h_1)}\right). \tag{4}$$

Where the pairwise feature differences $\hat{x}^{A,B} = -\hat{x}^A - \hat{x}^B$ are a normal distribution with zero mean. Then, the eq. (4) can be rewritten as

$$\delta^{(A,B)} = \log\left(\frac{N(\hat{x}^{A,B}, 0, \Sigma_{(A,B)=0})}{N(\hat{x}^{A,B}, 0, \Sigma_{(A,B)=1})}\right). \tag{5}$$

where

$$\Sigma_{(A,B)=0} = \sum_{(A,B)=0} (\hat{x}^{A,B})(\hat{x}^{A,B})^T, \tag{6}$$

$$\Sigma_{(A,B)=1} = \sum_{(A,B)=1} (\hat{x}^{A,B})(\hat{x}^{A,B})^T. \tag{7}$$

By taking the log and discarding the constant terms, the eq. (5) can be rewritten as

$$\delta^{(A,B)} = (\hat{x}^{A,B})^T (\Sigma^{-1}_{(A,B)=1} - \Sigma^{-1}_{(A,B)=0})(\hat{x}^{A,B}).. \tag{8}$$

Assuming that $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^{d \times r}$ is a training set from one camera view and $Z = (Z_1, Z_2, \ldots, Z_m) \in \mathbb{R}^{d \times r}$ is from the other view. Where *m* and *n* represent the number of samples in a *d* dimensional space. Considering a *r* dimensional subspace *W*, the eq. (8) can be computed as

$$\delta^{(X,Z)} = (\hat{x}^{X,Z})^T W (\Sigma'^{-1}_{(X,Z)=1} - \Sigma'^{-1}_{(X,Z)=0}) W^T (\hat{x}^{X,Z}), \tag{9}$$

where $\Sigma'^{-1}_{(X,Z)=1} = W^T \Sigma_{(X,Z)=1} W$ and $\Sigma'^{-1}_{(X,Z)=0} = W^T \Sigma_{(X,Z)=0} W$. But it is difficult to optimize the distance $\delta^{(X,Z)}$ with two inverse matrices in eq. (9). Given a basis *w*, the projected samples of the differences $\hat{x}_{(X,Z)=1}$ and $\hat{x}_{(X,Z)=0}$ are still zero mean but different variances. However, the variances $\hat{\varphi}_{(X,Z)=1}$ and $\hat{\varphi}_{(X,Z)=0}$ can also be used to determine whether pairwise pedestrian images are similar or not. Thus, the problem is how to maximize $J(w) = \hat{\varphi}_{(X,Z)=0}(w) / \hat{\varphi}_{(X,Z)=1}(w)$ by optimizing *w*, where $\hat{\varphi}_{(X,Z)=1}(w) = w^T \Sigma_{(X,Z)=1} w$ and $\hat{\varphi}_{(X,Z)=0}(w) = w^T \Sigma_{(X,Z)=0} w$. It is a generalized Rayleigh quotient whose maximum solution is equivalent to

$$\max_{w} w^T \Sigma_{(X,Z)=0} w, s.t. w^T \Sigma_{(X,Z)=1} w = 1 . \qquad \textbf{(10)}$$

Therefore, the corresponding eigenvector w_1 in the process of solving the largest eigenvalue of $\Sigma'_{(X,Z)=1} \Sigma_{(X,Z)=0}$ is the optimal solution that maximizes $J(w)$. Besides, the corresponding eigenvector w_2 in the process of solving the second largest eigenvalue of $\Sigma'_{(X,Z)=1} \Sigma_{(X,Z)=0}$ is also the second largest value of $J(w)$, which is the solution orthogonal to w_1, and so on. In this way, the subspace $W = (w_1, w_2, \ldots, w_r)$ can be learned.

4.2 Proposed Approach

Assuming that there are *M* images in the probe and *N* images in the gallery, and the u^{th} image in the probe is represented as $X^{P,u}$. With the learned discriminant subspace *W*, the distances between images in the probe and the gallery can be computed by KISSME metric. That is, the distance matrix is defined as

$$D = (d_{ij} \mid i = 1, \ldots, N; j = 1, \ldots, M\}, , \qquad \textbf{(11)}$$

and is shown in Fig. 6. It is widely used principle in person re-ID that ranks the distances between each image $X^{P,i} = \{x^{P,i} \mid 1 \le i \le M\}$ in the probe and every image $X^{G,j} = \{x^{G,i} \mid 1 \le i \le N\}$ in the gallery to obtain final matching results. According to common sense, the visual field monitored by each camera will not change too much. Therefore, the ideal variation between images in non-overlapping camera views only varies from person to person. The distances between the k^{th} image in the gallery $X^{G,k} = \{x^{G,k} \mid 1 \le k \le N\}$ and every image in the probe is defined as

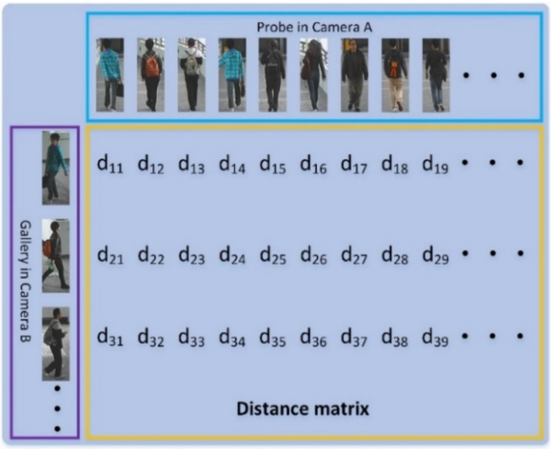

Probe in Camera A
Gallery in Camera B
d11 d12 d13 d14 d15 d16 d17 d18 d19 • • •
d21 d22 d23 d24 d25 d26 d27 d28 d29 • • •
d31 d32 d33 d34 d35 d36 d37 d38 d39 • • •
Distance matrix

Fig. 6. The distances between images in the probe and the gallery are represented as a distance matrix

$$D_k = \{d_{kj} \mid j = 1, ..., M\} . \qquad \textbf{(12)}$$

However, the reason why the final ranking results obtained by metric learning are not accurate that is the range of distances D_k changes greatly with different images in the gallery.

In order to solve this problem and improve the accuracy of final sorting results, the max-min criteria for linear space transformation is used to handle the outliers in distance matrix shown in Fig. 6, which is defined as

$$D_{k,final} = \frac{(D_{l,\max} - D_{l,\min}) \times (D_k - D_{k,\min})}{(D_{k,\max} - D_{k,\min})}. \quad \mathbf{(13)}$$

Where $D_{l,\max}$ and $D_{l,\min}$ respectively represent the maximum and minimum value of the distances between each image in the probe and every image in the gallery, and $D_{k,\max}$ and $D_{k,\min}$ respectively represent the maximum and minimum value of the distances between each image in the gallery and every image in the probe. According to linearly map the distance matrix to the interval (0, 1), the eq. (13) can be rewritten as

$$D_{k,final} = \frac{(D_k - D_{k,\min})}{(D_{k,\max} - D_{k,\min})}. \quad \mathbf{(14)}$$

Finally, we re-rank the new distances $D_{k,final}$ with the widely used principle in person re-ID as the final matching results, and the experiments show that it is effective to improve the accuracy of re-ID.

5 Experiments

5.1 Datasets and Evaluation Protocol

We evaluated our methods on two challenging datasets, VIPeR [1] and CUHK Campus [2]. There are significant occlusions, background clutters and cross-camera variations in these datasets.

VIPeR. It is the most widely used dataset in person re-ID, which is captured in academy environment and contains 1264 images of 632 pedestrians in two camera views, and all the images are cropped to 128×48 pixels. The dataset is the most challenging dataset in person re-ID for significant differences between pedestrian pairs, which contains variations on illumination, pose and viewpoint under cross-camera views. Some example images from VIPeR are shown in Fig. 1(a).

CUHK Campus. It is captured in a campus environment, which contains 971 pedestrians and 3884 images. And the size of images in this dataset is cropped to 160×60 pixels. There are two cameras and each camera captures two images for one person. The camera *A* mainly captures the front and back of the pedestrian while the camera *B* captures the side view. Some example images from CUHK Campus are shown in Fig. 1(b).

Evaluation protocol. We followed on two datasets is introduced in [23], the training set is randomly selected half of the images from the dataset and the remaining images as the testing set. Images in camera *A* and camera *B* are probe and gallery images. Matching each image in the probe with every image in the gallery and ranking the results are a trial in experiments. We also report the results with standard Accumulative Matching Characteristic (CMC) curves in person re-ID and the results of rank-*k* are obtained by getting an average of 10 trials.

5.2 Results and Analysis

In this paper, we propose a novel feature GOMO and an enhanced metric. In order to show the effective performance of our methods, the experiments are conducted as two parts. In the first part, we compare the proposed feature and metric with the latest methods used LOMO and XQDA respectively. The other part is used to compare our proposed methods with state-of-the-art methods. There are some methods have been proposed recently, which combine LOMO or XQDA as a part of the complete method to achieve a good performance. Our proposed methods in this paper are used to compare with those methods, and it is effective to highlight our contributions. Combining our proposed re-ranking method with XQDA metric as a new method is the biggest contribution in this paper, which outperforms state-of-the-art methods in person re-ID.

We compare the GOMO feature with recent methods using the LOMO feature, such as DNS [18] and XQDA [8], the results report in Fig. 7 and Table 1. The results show that our proposed feature GOMO has an effective improvement at each rank of CMC curves. And we compare our proposed enhanced metric with recent methods using the XQDA metric, such as $\text{GOG}_{\text{Fusion}}$ [7] and LOMO [8] feature, the results show in Fig. 8 and Table 2. Eight methods, the CRAFT [5], $\text{GOG}_{\text{Fusion}}$ [7], DNS [18], MetricEnsemble [24], MLAPG [33], XQDA [8], SalMatch [34], Mid-level Filter [35] are the best performances in the state-of-the-art, which are used to compare with our proposed method $\text{XQDA}_{\text{Ours}}$(+$\text{GOG}_{\text{Fusion}}$). The results are shown in Fig. 9 and Table 3 and our proposed method achieves the highest accuracy over state-of-the-art methods, 51.0% at rank 1, and exceeds 0.7% of the second best CRAFT method.

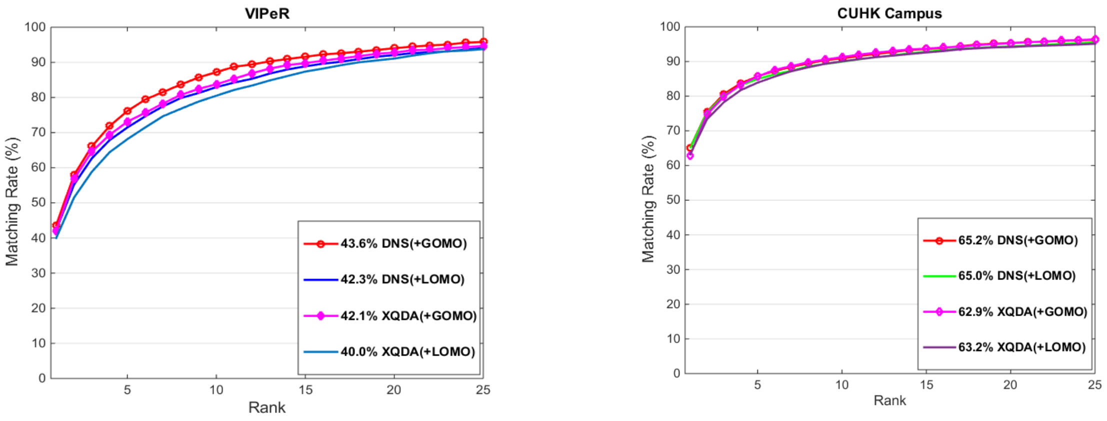


Fig. 7. The CMC curves of comparing the GOMO feature with recent methods using the LOMO feature

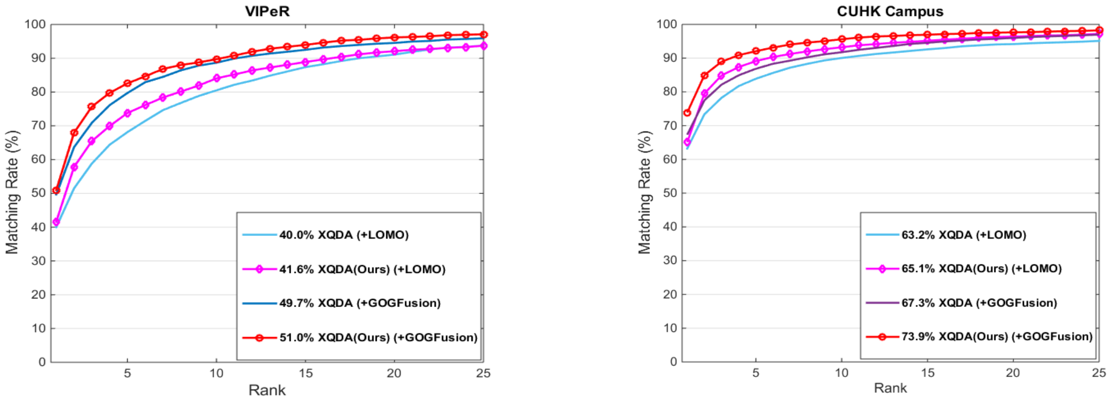


Fig. 8. The CMC curves of comparing the enhanced metric with recent methods using the XQDA metric

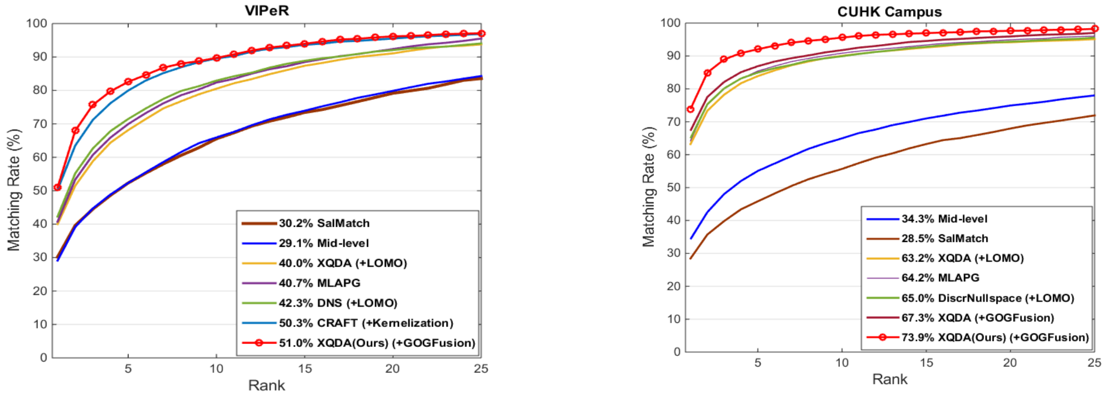


Fig. 9. The CMC curves of comparing the proposed method $\text{XQDA}_{\text{Ours}}$(+$\text{GOG}_{\text{Fusion}}$) with the state-of-the-art

Table 1. The list results of comparing the GOMO feature with recent methods using the LOMO feature

Methods	Reference	VIPeR				CUHK Campus			
		rank=1	*rank=5*	*rank=10*	*rank=20*	*rank=1*	*rank=5*	*rank=10*	*rank=20*
DNS (+LOMO)	**CVPR (2016) [18]**	42.3	71.5	82.9	92.1	65.0	85.0	89.9	94.4
DNS (+GOMO)	**Ours**	**43.6**	**76.1**	**87.2**	**94.0**	**65.2**	**85.7**	**90.9**	**95.3**
XQDA (+LOMO)	**CVPR (2015) [8]**	40.0	68.1	80.5	91.1	**63.2**	83.9	90.0	94.2
XQDA (+GOMO)	**Ours**	**42.1**	**73.1**	**83.8**	**92.8**	62.9	**85.7**	**91.3**	**95.3**

Table 2. The list results of comparing the enhanced metric with recent methods using the XQDA metric

Methods	Reference	VIPeR				CUHK Campus			
		rank=1	*rank=5*	*rank=10*	*rank=20*	*rank=1*	*rank=5*	*rank=10*	*rank=20*
XQDA(+GOG$_{\text{Fusion}}$)	**CVPR (2016) [7]**	49.7	79.7	88.7	94.5	67.3	86.9	91.8	96.0
XQDA$_{\text{Ours}}$(+GOG$_{\text{Fusion}}$)	**Ours**	**51.0**	**82.6**	**89.7**	**96.2**	**73.9**	**92.1**	**95.6**	**97.6**
XQDA(+LOMO)	**CVPR (2015) [8]**	40.0	68.1	80.5	91.1	63.2	83.9	90.0	94.2
XQDA$_{\text{Ours}}$(+LOMO)	**Ours**	**41.6**	**73.8**	**84.1**	**92.1**	**65.1**	**89.1**	**93.2**	**96.4**

Table 3. The list results of comparing the proposed method XQDA$_{\text{Ours}}$(+GOG$_{\text{Fusion}}$) with the state-of-the-art

Methods	Reference	VIPeR				CUHK Campus			
		rank=1	*rank=5*	*rank=10*	*rank=20*	*rank=1*	*rank=5*	*rank=10*	*rank=20*
XQDA$_{\text{Ours}}$ (+GOG$_{\text{Fusion}}$)	**Ours**	**51.0**	**82.6**	**89.7**	**96.2**	73.9	**92.1**	**95.6**	**97.6**
CRAFT (+Kernelization)	**TPAMI (2018) [5]**	50.3	80.0	89.6	95.5	**74.5**	**92.1**	94.8	97.1
XQDA (+GOG$_{\text{Fusion}}$)	**CVPR (2016) [7]**	49.7	79.7	88.7	94.5	67.3	86.9	91.8	96.0
DNS (+LOMO)	**CVPR (2016) [18]**	42.3	71.5	82.9	92.1	65.0	85.0	89.9	94.4
MetricEnsemble	**CVPR (2015) [24]**	45.9	77.5	88.9	95.8	-	-	-	-
MLAPG	**ICCV (2015) [33]**	40.7	69.9	82.3	92.4	64.2	85.4	90.8	94.9
XQDA (+LOMO)	**CVPR (2015) [8]**	40.0	68.1	80.5	91.1	63.2	83.9	90.0	94.2
Mid-level Filter	**CVPR (2014) [35]**	29.1	52.3	65.9	79.9	34.3	55.1	65.0	74.9
SalMatch	**ICCV (2013) [34]**	30.2	52.3	65.5	79.1	28.5	45.9	55.7	68.0

6 Conclusion and Future Work

In this work, we have proposed a novel feature representation and an enhanced metric learning method. The new feature GOMO is a maximum global feature, which describes image feature in two directions. First, a maximum feature representation in the horizontal direction is attracted. Then we solve the misalignment problem that appears in maximum feature representation of vertical direction, and combine features in two directions as a fusion future GOMO. The re-ranking method we proposed in this paper is combined with XQDA as a new method, which is devoted to reduce the errors between camera views caused by the outliers in final distance matrix of the XQDA method. We use the max-min criterion to linearly map the distance matrix obtained by the XQDA method, and re-rank the new distance matrix to achieve the final ranks. We conducted experiments on two challenging person re-ID datasets, VIPeR and CUHK Campus, and the experimental results demonstrate the effectiveness of our proposed feature and the superiority of our enhanced metric learning method over the state-of-the-art methods.

Finally, we have proposed two promising methods, the GOMO feature and the XQDA$_{\text{Ours}}$(+GOG$_{\text{Fusion}}$) metric learning. Our GOMO feature is designed to use HSV visual descriptor and SILTP texture descriptor. In practice, the only HSV visual descriptor may not enough to support high-precision feature extraction. Therefore, if the limited descriptors can be well handled. The more robust feature may be represented by our method. Besides, the re-ranking method is only combined with XQDA method as an enhanced metric. Considering the fact that metric learning has developed rapidly. We are interested in combining the new metric learning technique with our re-ranking method in future work.

Acknowledgments

This work is supported by key projects of Shanghai Science and Technology Commission (No. 17511106802).

References

[1] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, 2007.

[2] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning, in: Proc. Computer Vision – ACCV 2012, 2012.

[3] A. Bedagkar-Gala, S.K. Shah, A survey of approaches and trends in person re-ID, Image and Vision Computing 32(4)(2014) 270-286.

[4] X. Wang, R. Zhao, Person re-ID: system design and evaluation overview, in S. Gong, M. Cristani, S. Yan, C.C. Loy (Eds.), Person Re-ID, Springer, London, 2014, pp. 351-370.

[5] Y.C. Chen, X. Zhu, W.S. Zheng, J.H. Lai, Person re-ID by camera correlation aware feature augmentation, IEEE Transactions on Pattern Analysis Machine Intelligence 40(2)(2018) 392-408.

[6] R. Zhao, W. Oyang, X. Wang, Person re-ID by saliency learning, IEEE Transactions on Pattern Analysis Machine Intelligence 39(2)(2017) 356-370.

[7] T. Matsukawa, T. Okabe, E. Suzuki, Y. Sato, Hierarchical Gaussian descriptor for person re-ID, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-ID by local maximal occurrence representation and metric learning, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[9] R. R. Varior, G. Wang, J. Lu, T. Liu, Learning invariant color features for person reidentification, IEEE Transactions on Image Processing 25(7)(2016) 3395-3410.

[10] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-ID, in: Proc. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[11] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, S.Z. Li, Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes, in: Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.

[12] A. Subramaniam, M. Chatterjee, A. Mittal, Deep neural networks with inexact matching for person re-ID, in: Proc. Advances in Neural Information Processing Systems, 2016.

[13] W. Li, R. Zhao, T. Xiao, X. Wang, DeepReID: Deep filter pairing neural network for person re-ID, in: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[14] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-ID, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] W.S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Transactions on Pattern Analysis Machine Intelligence 35(3)(2013) 653-668.

[16] Z. Li, S. Chang, F. Liang, T.S. Huang, L. Cao, J.R. Smith, Learning locally-adaptive decision functions for person verification, in: Proc. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[17] M. Köstinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[18] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-ID, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[19] A. Khellaf, A. Beghdadi, H. Dupoisot, Entropic contrast enhancement, IEEE Transactions on Medical Imaging 10(4)(1991) 589-592.

[20] C. Wang, Z. Ye, Brightness preserving histogram equalization with maximum entropy: a variational perspective, IEEE Transactions on Consumer Electronics 51(4)(2005) 1326-1334.

[21] D. Sheet, H. Garud, A. Suveer, M. Mahadevappa, J. Chatterjee, Brightness preserving dynamic fuzzy histogram equalization, IEEE Transactions on Consumer Electronics 56(4)(2010) 2475-2480.

[22] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-ID by symmetry-driven accumulation of local features, in: Proc. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.

[23] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: Proc. Computer Vision – ECCV 2008, 2008.

[24] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-ID with metric ensembles, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[25] A. Mignon, F. Jurie, PCCA: a new approach for distance learning from sparse pairwise constraints, in: Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[26] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-ID, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[27] S. Wu, Y.C. Chen, X. Li, A.C. Wu, J.J. You, W.S. Zheng, An enhanced deep feature representation for person re-ID, in: Proc. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016.

[28] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: automatic query expansion with a generative feature model for object retrieval, in: Proc. 2007 IEEE 11th International Conference on Computer Vision, 2007.

[29] H. Jegou, H. Harzallah, C. Schmid, A contextual dissimilarity measure for accurate and efficient image search, in: Proc. 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[30] D. Qin, S. Gammeter, L. Bossard, T. Quack, L. van Gool, Hello neighbor: accurate object retrieval with k-reciprocal nearest neighbors, in: Proc. CVPR 2011, 2011.

[31] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-ID with k-reciprocal encoding, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[32] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition 29(1)(1996) 51-59.

[33] S. Liao, S. Z. Li, Efficient PSD constrained asymmetric metric learning for person re-ID, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

[34] R. Zhao, W. Ouyang, X. Wang, Person re-ID by salience matching, in: Proc. 2013 IEEE International Conference on Computer Vision, 2013.

[35] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-ID, in: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.