doi:10.3772/j.issn.1006-6748.2022.03.002

Feature mapping space and sample determination for person re-identification①

HOU Wei(侯 巍), HU Zhentao②, LIU Xianxing, SHI Changsen
(School of Artificial Intelligence, Henan University, Zhengzhou 450046, P. R. China)

Abstract

Person re-identification (Re-ID) is integral to intelligent monitoring systems. However, due to the variability in viewing angles and illumination, it is easy to cause visual ambiguities, affecting the accuracy of person re-identification. An approach for person re-identification based on feature mapping space and sample determination is proposed. At first, a weight fusion model, including mean and maximum value of the horizontal occurrence in local features, is introduced into the mapping space to optimize local features. Then, the Gaussian distribution model with hierarchical mean and covariance of pixel features is introduced to enhance feature expression. Finally, considering the influence of the size of samples on metric learning performance, the appropriate metric learning is selected by sample determination method to further improve the performance of person re-identification. Experimental results on the VIPeR, PRID450S and CUHK01 datasets demonstrate that the proposed method is better than the traditional methods.

Key words: person re-identification (Re-ID), mapping space, feature optimization, sample determination

0 Introduction

The purpose of person re-identification (Re-ID) is to match the same person from different camera views[1]. Person Re-ID is a key component of video surveillance, which is of great significance in security monitoring, person search and criminal investigation. Although great progress has been made in person Re-ID, there are still many problems to be solved due to the existence of visual ambiguities.

The visual ambiguities brought by changes in viewpoint and illumination are manifested in the person images like large changes in scale and background of the same person, which can significantly degrade the performance of the person Re-ID system. To overcome this limitation, there have been studies that try to use local information and information discrimination[2-3]. Properly utilizing the information in person images and better discriminating them can effectively improve the performance of person Re-ID. The related studies that have emerged in person Re-ID can be generally classified into two types: feature extraction and metric learning.

Some researchers construct features of person images based on color, texture and other appearance attributes[4-5]. The basic idea is that the person image is divided into multiple overlapping or non-overlapping local image blocks, and then color or texture features are extracted from them separately, thus adding spatial region information into person image features. When calculating the similarity of two person images, the features within the corresponding image blocks will be compared separately, and then the comparison results of each image block will be fused as the final recognition result. Nevertheless, the features constructed by the above method are weak and the feature representation for person Re-ID is abated.

On the other hand, there are many work that use a given set of training samples to obtain a metric matrix that effectively reflects the similarity between data samples, increasing the distance between non-similar samples while reducing the distance between similar samples[6]. However, these methods do not consider the effect of sample size on the metric learning performance, making the person Re-ID results less reliable.

Color features are robust to pose and viewpoint

① Supported by the National Natural Science Foundation of China (No. 61976080), the Science and Technology Key Project of Science and Technology Department of Henan Province (No. 212102310298), the Innovation and Quality Improvement Project for Graduate Education of Henan University (No. SYL20010101), and the Academic Degress & Graduate Education Reform Project of Henan Province (2021SJLX195Y).

② To whom correspondence should be addressed. E-mail: huzhentao2011@126.com.

Received on Sep. 19, 2021

changes, but are susceptible to illumination and obstructions. It is difficult to effectively distinguish large-scale person images using only color features due to the similarity of dressing problem. The clothing often contains texture information, and texture features involve comparison of neighboring pixels and are robust to illumination, so making full use of color and texture features is very effective for person Re-ID. However, traditional methods apply single color and texture features to the person Re-ID task, and they are insufficient to handle the differences between different person images. In addition, the completeness and richness of feature representations also affect the results of similarity metrics, and traditional methods do not fully utilize the richness of samples when dealing with such metrics, resulting in lower overall performance of the methods.

To address the above problems, this paper proposes a person Re-ID method based on feature mapping space and sample determination metric learning. The method combines an improved weighted local maximal occurrence (wLOMO) feature that modifies the original LOMO[7] feature with the Gaussian of Gaussian (GOG)[8] feature, and uses a sample determination method to select a suitable metric learning method to rank the similarity of person images. The method in this paper performs simulation experiments on each of the three typical datasets and is compared with other methods. The main contributions are summarized as follows.

(1) A fused feature mapping space is proposed to enhance the person images features. The mean information of the horizontal direction of person image is introduced into LOMO feature, and the weighted mean and max are fused to obtain the proposed wLOMO feature. To enhance the feature expression of each person image, wLOMO feature is combined with GOG feature. On this basis, in order to simplify the complexity of feature extraction model, the feature transformation processes of wLOMO and GOG are integrated into one feature mapping space.

(2) A sample determination method is proposed to accommodate different sample sizes. In the dataset, the sample determination method selects the appropriate metric learning to accomplish the similarity ranking of person images according to the demand of different sample sizes. In addition, the selected sample size is dynamically tuned according to the matching rate of different metric learning outputs.

(3) Extended experiments on three publicly available datasets are designed to evaluate the performance of the proposed method and the comparison method, and to demonstrate the effectiveness and applicability of the proposed method in person Re-ID.

1 Related work

The research on person Re-ID can be divided into two groups: feature extraction and metric learning. Person Re-ID based on feature extraction is usually constructed by basic color, texture and other appearance attributes. Ref. [2] proposed the symmetry driven accumulation of local feature (SDALF) based on the symmetrical and asymmetric characteristics of person body structure, which fused three kinds of color feature in person image to complete the discrimination of person image. Ref. [4] proposed an ensemble of localized features (ELF) method. The method adopted AdaBoost algorithm to select the appropriate feature combination from a group of color and texture features, which improved the experimental accuracy. Refs[5,9,10] introduced biologically inspired features (BIF) in person images. By calculating the characteristics of BIF on adjacent scales, a feature called Bicov was proposed. On this basis, Gabor filter and covariance feature were introduced to deal with the problems caused by illumination change and background transformation in person images. Ref. [11] proposed a feature transformation method based on Zero-Padding Augmentation, which could align the features distributed across the disjoint person images to improve the performance of the matching model. Ref. [12] constructed the feature fusion network (FNN) by combining the manually extracted features and deep learning features, and realized the fusion of deep learning features and artificial features by constantly adjusting the parameters of the deep neural network. Ref. [13] proposed a deep convolution model, which highlights the discriminative part by giving the features in each part of the person a different weight to realize the person Re-ID task. The person Re-ID method based on deep learning needs to consider using a large number of labeled samples to train a complex model, and the training process is very time-consuming.

Person Re-ID methods based on metric learning minimizes the distance between similar person by learning appropriate similarity. Ref. [3] introduced the concept of large margin in Mahalanobis distance and proposed a metric learning method called large margin nearest neighbor (LMNN). LMNN assumed that the sample features of the same class were adjacent, so there was a big gap between the feature samples of different classes. Thus, when calculating the distance, the features of the same kind of samples were gathered, and the different types of samples were pushed. Ref. [6] proposed a local fisher discriminative analysis (LFDA) method, which introduced a matrix based on subspace

learning, allocated different scale factors for the same classes and different classes, and used the local invariance principle to calculate the distance. Ref. [14] proposed a Mahalanobis distance metric called keep it simple and straightforward metric (KISSME) by calculating the difference between the intra class and inter class covariance matrices of sample features. The method did not need to calculate the metric matrix through complex iterative algorithm, so it was more efficient. Ref. [15] used a new multi-scale metric learning method based on strip descriptors for person Re-ID. According to this method, the internal structure of different person images can be effectively extracted, improving the recognition rate. However, due to the non-linearity of the person image in the cross field of view, the linear transformation generated by the general metric learning method effects commonly general. Therefore, the kernel correlation based metric learning method was introduced to solve the nonlinear problem in person Re-ID[16-17]. However, the above-mentioned methods adopt a single strategy to deal with the change of sample size, without considering the accuracy impact of the method itself.

2 Problem description

It considers that the general process of person re-recognition is to extract features first and then rank them by metric learning. The performance of the method depends strongly on the expression ability of features and metric learning, and the existence of visual ambiguities will inevitably affect the ability. To solve this problem, a new method is proposed to improve the matching rate of person re-recognition.

The framework of the proposed method is divided into three parts in Fig. 1. The first part is the extraction of basic color, texture and spatial features, the second part is the mapping process of basic features, and the third part is the metric learning method based on sample determination.

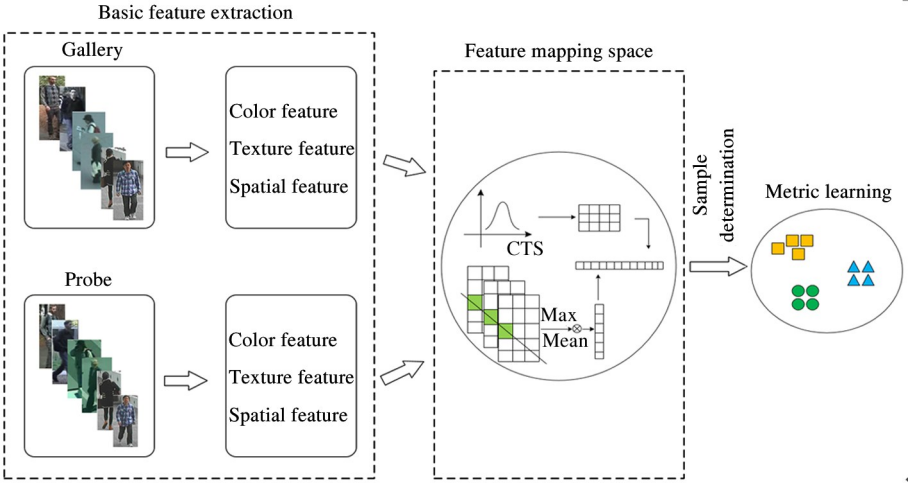

Fig. 1 The person re-identification framework

3 Methodology

Based on the wLOMO in subsection 3.1 and the proposed sample determination in subsection 3.2, the proposed method flowchart is shown in Fig. 2.

3.1 Feature mapping space

When designing the feature mapping space, two state-of-the-art feature transformation processes are merged into one feature mapping space by cascading, which simplifies the feature extraction.

3.1.1 LOMO

When extracting LOMO features, a 10×10 sliding subwindow is used to represent the local area of a person image, and an $8 \times 8 \times 8$ bin combined color histogram of the hue, saturati, value (HSV) and two scale the scale invariant local ternary pattern (SILTP) texture histogram $\boldsymbol{F}_{\mathrm{SILTP}}$ are extracted from each subwindow. Then the maximum value of pixel features occurrence of all subwindows at the same horizontal position is calculated as

$$\boldsymbol{F}^{1}_{\mathrm{HSV}} = \max(\rho_{\mathrm{hsv}}) \quad (1)$$

$$\boldsymbol{F}_{\mathrm{SILTP}} = \max(\rho_{\mathrm{SILTP}}) \quad (2)$$

where $\rho_{(\bullet)}$ is the pixel feature occurrence in all subwindows.

3.1.2 The proposed wLOMO

Considering that the maximization of pixel features leads to the loss of some person features, and the clothes worn by person are often composed of a small number

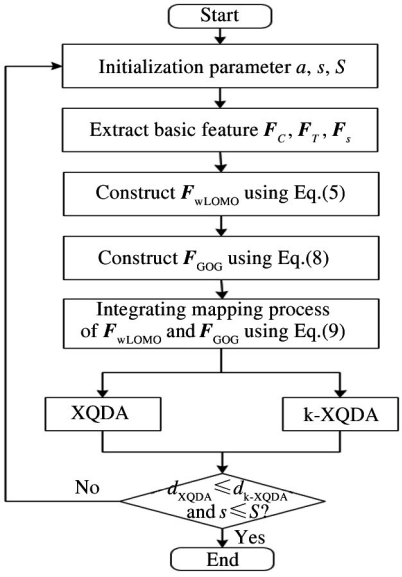

Fig. 2 Flowchart of the proposed method

of colors in each part, the mean information can enhance the feature expression of person images when the person background changes little. Therefore, the mean information of pixel feature distribution is introduced into the feature expression, expressed as

$$\boldsymbol{F}_{\mathrm{HSV}}^{2} = mean(\rho_{\mathrm{hsv}}) \tag{3}$$

Then, the maximum and mean of feature occurrence are weighted with parameter a, and the added feature is

$$\boldsymbol{F}_{\mathrm{HSV}} = (1-a)\boldsymbol{F}_{\mathrm{HSV}}^{1} + a\boldsymbol{F}_{\mathrm{HSV}}^{2} \tag{4}$$

Next, through 2×2 pooling twice, the original person images are down sampled to two smaller scales, and then the image features are extracted again by the above extraction method. Finally, the features of all scales are combined to form feature $\boldsymbol{F}_{\mathrm{wLOMO}}$.

$$\boldsymbol{F}_{\mathrm{wLOMO}} = [\boldsymbol{F}_{\mathrm{HSV}}, \boldsymbol{F}_{\mathrm{SILTP}_{4,3}^{0.3}}, \boldsymbol{F}_{\mathrm{SILTP}_{4,5}^{0.3}}] \tag{5}$$

where $\boldsymbol{F}_{\mathrm{wLOMO}}$ is the improved LOMO feature, $\boldsymbol{F}_{\mathrm{SILTP}_{4,3}^{0.3}}$ and $\boldsymbol{F}_{\mathrm{SILTP}_{4,5}^{0.3}}$ are the feature representation when the SILTP algorithm takes the radius 3 and 5 respectively. Additionally, 0.3 is a scale factor indicating the comparing range, and 4 is the number of neighboring pixels.

3.1.3 GOG

Considering that color features are more sensitive to illumination changes in cross view person images, and the impact of spatial information loss on person ReID, this paper further extracts GOG features from the same person image to enhance the feature expression. Firstly, the pixel level feature $\boldsymbol{f}$ is extracted as

$$\boldsymbol{f} = [\boldsymbol{y}, \boldsymbol{F}_{M_\theta}, \boldsymbol{F}_{\mathrm{RGB}}, \boldsymbol{F}_{\mathrm{HSV}}, \boldsymbol{F}_{\mathrm{LAB}}, \boldsymbol{F}_{\mathrm{RG}}]^{\mathrm{T}} \tag{6}$$

where $\boldsymbol{F}_{\mathrm{RGB}}$, $\boldsymbol{F}_{\mathrm{HSV}}$, $\boldsymbol{F}_{\mathrm{LAB}}$, $\boldsymbol{F}_{\mathrm{RG}}$ are the color features, $\boldsymbol{F}_{M_\theta}$ is the texture feature, $\boldsymbol{y}$ is the space feature. The color features are channel values of person image, M_θ consists of the values of pixel intensity gradients in the four standard directions of the two-dimensional coordinate system. $\boldsymbol{y}$ is the position of the pixel in the vertical direction of image. After that, block level features are extracted. Each person image is divided into G partially overlapped horizontal regions, and each region is divided into $k \times k$ local blocks. The pixel features in each local block s are represented by Gaussian distribution to form a Gaussian block $\boldsymbol{z}_i$

$$\boldsymbol{z}_i = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_s|} \exp\left(-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu}_s)^{\mathrm{T}}\boldsymbol{\Sigma}_s^{-1}(\boldsymbol{f}-\boldsymbol{\mu}_s)\right) \tag{7}$$

where $\boldsymbol{\mu}_s$ is the mean vector, $\boldsymbol{\Sigma}_s$ is the covariance matrix of block s.

Then, the Gauss block $\boldsymbol{z}_i$ is mapped to symmetric positive definite matrix to complete block level feature extraction. Finally, the region level features are extracted. The Gaussian blocks are modeled as a Gaussian region by Gaussian distribution. Meanwhile, Gaussian region is embedded into symmetric positive definite matrix. These vectors are finally aggregated to form the GOG feature $\boldsymbol{F}_{\mathrm{GOG}}$ of a person image.

$$\boldsymbol{F}_{\mathrm{GOG}} = [\boldsymbol{z}_1^{\mathrm{T}}, \cdots, \boldsymbol{z}_G^{\mathrm{T}}]^{\mathrm{T}} \tag{8}$$

where $\boldsymbol{z}_G$ is the G-th horizontal region feature of a person image.

3.1.4 Feature mapping space

The proposed wLOMO describes only maximum occurrence and mean occurrence of pixel features, moreover, GOG can provide covariance information.

To comprehensively consider the maximum occurrence, mean occurrence and covariance information of pixel features, Eq. (5) and Eq. (8) are combined. It means that wLOMO feature and GOG feature are aligned according to the person's identity, and their feature mapping process is simplified to one feature mapping space by cascading.

$$\boldsymbol{F} = [\boldsymbol{F}_{\mathrm{wLOMO}}, \boldsymbol{F}_{\mathrm{GOG}}] \tag{9}$$

where $\boldsymbol{F}$ is the feature of the output of the mapping space.

3.2 Sample determination

Cross-view quadratic discriminant analysis (XQDA)[7] and kernel cross-view quadratic discriminant analysis (k-XQDA)[18] are state-of-the-art methods in depending on feature dimension and samples size respectively. Based on the two methods, a sample determination method is proposed to synthesize the advantages of the two methods.

3.2.1 XQDA

Before summarizing the XQDA method, a brief in-

troduction is given to the distance measurement of person Re-ID. For a dataset X, it contains C classes person $c_i (1 \leqslant i \leqslant C) \in R^n$. The classical Mahalanobis distance metric learns the distance $d(\boldsymbol{x}_i, \boldsymbol{z}_j)$ between person $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{in}]$ in camera a and person $\boldsymbol{z}_j = [\boldsymbol{z}_{j1}, \boldsymbol{z}_{j2}, \cdots, \boldsymbol{z}_{jm}]$ in camera b.

$$d(\boldsymbol{x}_i, \boldsymbol{z}_j) = (\boldsymbol{x}_i - \boldsymbol{z}_j)^{\mathrm{T}} \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{z}_j) \tag{10}$$

In fact, the XQDA method estimates the covariance matrix $\boldsymbol{M}$ in Mahalanobis distance. Since calculating the internal aggregation relationship only in Mahalanobis distance is insufficient, it is also necessary to add the relationship between classes. Therefore, the covariance matrix $\boldsymbol{M}$ can be redefined as

$$\boldsymbol{M} = \boldsymbol{\Sigma}_I^{-1} - \boldsymbol{\Sigma}_E^{-1} \tag{11}$$

where $\boldsymbol{\Sigma}_I^{-1}$ is the intra class covariance matrix, $\boldsymbol{\Sigma}_E^{-1}$ is the inter class covariance matrix. XQDA combines the idea of dimensionality reduction and metric learning. By learning the mapping matrix $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_r) \in \mathbf{R}^{n \times r} (n > r)$, XQDA maps the original high dimension space to a low dimension space. Therefore, Eq. (10) can be rewritten as

$$d(\boldsymbol{x}_i, \boldsymbol{z}_j) = (\boldsymbol{x}_i - \boldsymbol{z}_j)^{\mathrm{T}} \boldsymbol{W} (\Sigma_I^{'-1} - \Sigma_E^{'-1}) \boldsymbol{W}^{\mathrm{T}} (\boldsymbol{x}_i - \boldsymbol{z}_j) \tag{12}$$

where $\Sigma_I^{'} = \boldsymbol{W}^{\mathrm{T}} \Sigma_I \boldsymbol{W}$, $\Sigma_E^{'} = \boldsymbol{W}^{\mathrm{T}} \Sigma_E \boldsymbol{W}$.

However, directly calculating d is difficult because $\boldsymbol{W}$ is contained in two inverse matrices, and can be converted to solve the generalized Rayleigh quotient problem of $J(\boldsymbol{w}_k)$.

$$J(\boldsymbol{w}_k) = \frac{\boldsymbol{w}_k^{\mathrm{T}} \Sigma_E \boldsymbol{w}_k}{\boldsymbol{w}_k^{\mathrm{T}} \Sigma_I \boldsymbol{w}_k} = \frac{\boldsymbol{w}_k^{\mathrm{T}} \Sigma_I^{-1} \Sigma_E \boldsymbol{w}_k}{\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{w}_k} \tag{13}$$

where the optimal solution of $\boldsymbol{w}_k$ in the mapping space $\boldsymbol{W}$ corresponds to the eigenvectors of the first r eigenvalues in $\Sigma_I^{-1} \Sigma_E$, $k \in \{1, \cdots, r\}$.

3.2.2 k-XQDA

XQDA metric learning method is directly trained in the original linear feature space, and the similarity and difference among samples are not well expressed. k-XQDA uses a kernel function to map the original samples into the easily distinguishable nonlinear space, and then distinguishes the differences of samples in the nonlinear space. The derivation of k-XQDA method involves mainly the distance metric function $d(\boldsymbol{x}_i, \boldsymbol{z}_j)$ in XQDA and the kernelization of the cost function $J(\boldsymbol{w}_k)$.

In the kernel space, two kinds of expansion coefficients α and β corresponding to person in camera a and b are used, respectively. Mapping matrix $\boldsymbol{w}_k$ can be expressed as

$$\boldsymbol{w}_k = \sum_{i=1}^{n} \alpha_i^{(k)} \phi(\boldsymbol{x}_i) + \sum_{j=1}^{m} \beta_j^{(k)} \phi(\boldsymbol{z}_j) \tag{14}$$

Let $\boldsymbol{\Phi}_x = [\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_n)]$, $\boldsymbol{\Phi}_z = [\phi(\boldsymbol{z}_1), \cdots, \phi(\boldsymbol{z}_m)]$, Eq. (14) can be rewritten as

$$\boldsymbol{w}_k = \boldsymbol{\Phi}_x \boldsymbol{\alpha}_k + \boldsymbol{\Phi} \boldsymbol{\beta}_k = \boldsymbol{\Phi} \boldsymbol{\theta}_k \tag{15}$$

where $\boldsymbol{\alpha}_k = [\alpha_1^{(k)}, \alpha_2^{(k)}, \cdots, \alpha_n^{(k)}]^{\mathrm{T}}$, $\boldsymbol{\beta}_k = [\beta_1^{(k)}, \beta_2^{(k)}, \cdots, \beta_m^{(k)}]^{\mathrm{T}}$, $\boldsymbol{\theta}_k = [\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k]^{\mathrm{T}}$, $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_x, \boldsymbol{\Phi}_z]$.

After the kernel transformation, $J(\boldsymbol{w}_k)$ changes to

$$J(\boldsymbol{\theta}_k) = \frac{\boldsymbol{\theta}_k^{\mathrm{T}} \Lambda_E \boldsymbol{\theta}_k}{\boldsymbol{\theta}_k^{\mathrm{T}} \Lambda_I \boldsymbol{\theta}_k} = \frac{\boldsymbol{\theta}_k^{\mathrm{T}} \Lambda_I^{-1} \Lambda_E \boldsymbol{\theta}_k}{\boldsymbol{\theta}_k^{\mathrm{T}} \boldsymbol{\theta}_k} \tag{16}$$

Kernel cost function $J(\boldsymbol{\theta}_k)$ is the form of the generalized Rayleigh quotient, so the optimal solution of $\boldsymbol{\theta}_k$ corresponds to the eigenvectors of the first b largest eigenvalues in $\Lambda_I^{-1} \Lambda_E$, $\Lambda_I \in \mathbb{R}^{(n+m) \times (n+m)}$.

After kernelization, the distance metric function $d(\boldsymbol{x}_i, \boldsymbol{z}_j)$ can be expressed as

$$d(\phi(\boldsymbol{x}_i), \phi(\boldsymbol{z}_j)) = (\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{z}_j))^{\mathrm{T}} W_\phi (\Sigma_I^{'-1} - \Sigma_E^{'-1}) W_\phi^{\mathrm{T}} (\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{z}_j)) \tag{17}$$

where $\Sigma_I^{'} = W_\phi^{\mathrm{T}} \Sigma_I W_\phi$, $\Sigma_E^{'} = W_\phi^{\mathrm{T}} \Sigma_E W$, $W_\phi = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_b)$.

3.2.3 Sample determination

All the intrinsic matrix dimensions of k-XQDA method depend on the size of samples, which greatly reduces the amount of calculation compared with the XQDA method depending on the feature dimension.

On the basis of subsection 3.2.1 and subsection 3.2.2, considering the different focus of the two metric learning methods, in order to integrate the advantages of the two and make the actual person re-identification task a better match, this paper proposes a sample determination method, that is, when the size of training set S satisfies the Eq. (18), using the corresponding metric learning method will make a better effect in the corresponding dataset.

$$d = \begin{cases} d_{\mathrm{XQDA}} & s \leqslant S \\ d_{\mathrm{k-XQDA}} & s > S \end{cases} \tag{18}$$

where S is the sample size to be determined, s is the current sample size.

4 Experiments

To evaluate the performance of the method fairly, all the comparison methods run in the same environment. The hardware environment is Intel Core i7-9700F CPU@3.00 GHz, 8 GB RAM. The operating system is Windows 10 64 bit, and the software environment is Matlab 2019b.

4.1 Datasets and evaluation protocol

The effectiveness of the proposed method is demonstrated by three publicly available datasets, they are VIPeR[19], PRID450S[20] and CUHK01[21]. The VIPeR dataset contains 632 persons with different identi-

ties. Each person involves two images captured from two disjoint camera views, including variations in background and illumination. The PRID450S dataset contains 450 persons with different identities. Each person covers two images captured by two non-overlapping cameras with a single background. The CUHK01 dataset consists of 971 persons with a total of 3884 shots captured by two non-overlapping cameras with an average of two images for each person, and the person poses vary greatly.

To evaluate the results of the features in different metric learning, cumulative match characteristics (CMC) curve is used as the evaluation protocol.

4.2 Comparison with state-of-the-art

All images are normalized to the same size of 128 ×48 pixels. The datasets of VIPeR, PRID450S and CUHK01 are randomly divided into two equal parts, one half for training and the other for testing. The size of images in the training set of the three data sets is 632, 450 and 972 respectively. To eliminate the performance difference caused by randomly dividing the training set and the testing set, the process is repeated 10 times, and the average cumulative matching accuracies at rank 1, 5, 10 and 20 are reported over 10 runs. In addition, the corresponding CMC curves are shown.

4.2.1 Evaluation of the mapping space

To analyze the effectiveness of the proposed mapping space, the output features of the mapping space are sent to the XQDA metric learning method to verify the performance of the method. Since the method is iterative, different weights are looped in different datasets to retain the one with the highest performance. Furthermore, showing the Rank-1 values corresponding to various weights may indicate that the weights are not constant and change between datasets. This paper selects three different datasets and compares the results with state-of-the-art approaches.

VIPeR dataset: to analyze the influence of weight a on the performance of the wLOMO, the Rank-1 under different weight on VIPeR dataset are shown in Fig. 3. It can be seen the introduction of mean information has a certain impact on the method performance. When a is in range of 0.1 – 0.2, the performance of the method is optimal, and increasing a continually the performance of the method declines.

The compared methods and their matching rates on VIPeR are shown in Table 1 and Fig. 4. The results are reported in Table 1, the Rank-1 of LOMO, LSSCDL, DNS and GOG are better, all exceeding 40%. The proposed approach achieves 50.63% in Rank-1, which is 2.37% better than GOG.

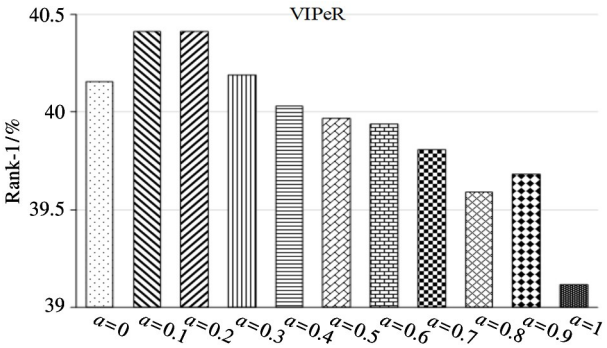

Fig. 3 Rank-1 matching rates

Table 1 Comparison of Rank results with other methods on VIPeR dataset

Methods	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
MidFilter [22]	29.11	52.34	65.95	79.87
gBiCov [23]	22.80	48.78	64.10	77.80
kCCA [24]	30.16	62.69	76.04	86.80
HSCD [25]	31.15	57.33	69.54	81.22
LOMO [16]	40.16	68.16	80.98	91.04
LSSCDL [26]	42.66	72.42	84.27	91.93
DNS [27]	42.28	71.46	82.94	92.05
GOG [17]	48.26	77.37	86.39	94.11
CAMEL [28]	26.70	50.46	63.91	79.93
VS-SSL [29]	44.80	72.30	79.30	86.10
CRAFT [30]	47.82	77.53	87.78	94.84
Proposed method	50.63	80.53	88.89	95.66

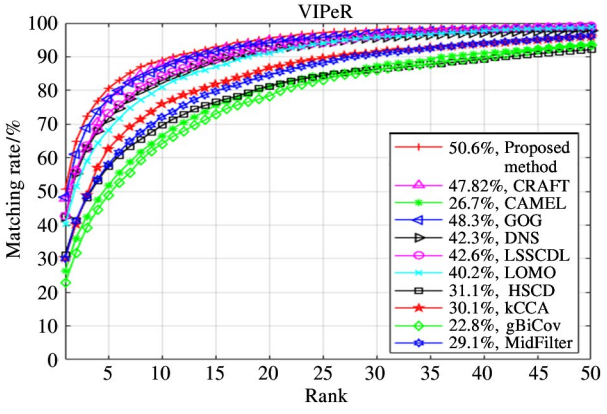

Fig. 4 CMC curves

PRID450S dataset: Fig. 5 shows the performance comparison of the wLOMO under different weight values. When the weight value is 0.3 – 0.4, the method performance is optimal.

The comparison methods and their matching rates results on PRID450S dataset are shown in Table 2 and Fig. 6. Different from the person images in VIPeR and

CUHK01 datasets, the background of person images in PRID450S dataset is relatively simple, and the background interference to all methods is small, the final matching results are generally better. For the proposed method with mean information, the matching rate of Rank-1 is 71.42%, outperforming the second best one GOG by 3.6%.

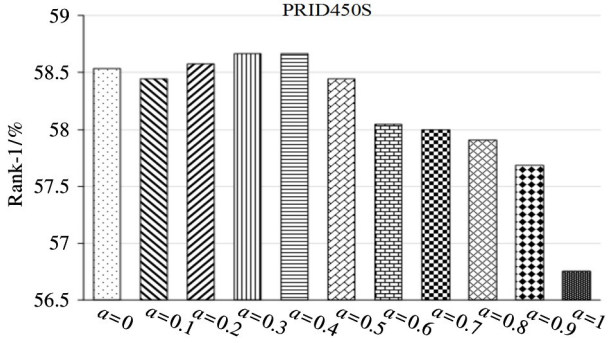

Fig. 5 Rank-1 matching rates

Table 2 Comparison of Rank results with other methods on PRID450S dataset

Methods	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
MidFilter[22]	35.26	55.62	67.45	80.33
gBiCov[23]	27.94	52.38	67.22	76.81
kCCA[24]	53.72	75.64	81.26	85.40
HSCD[25]	50.42	73.15	79.63	84.37
LOMO[16]	58.53	81.73	88.93	94.31
LSSCDL[26]	60.49	86.17	88.58	93.60
DNS[27]	61.20	85.51	91.16	95.60
GOG[17]	67.82	89.20	94.62	97.87
CAMEL[28]	31.56	54.85	66.43	82.30
MVLDML[31]	66.80	88.80	94.80	97.70
TDRP[32]	68.89	88.98	93.80	97.88
Proposed method	71.42	91.20	95.20	97.96

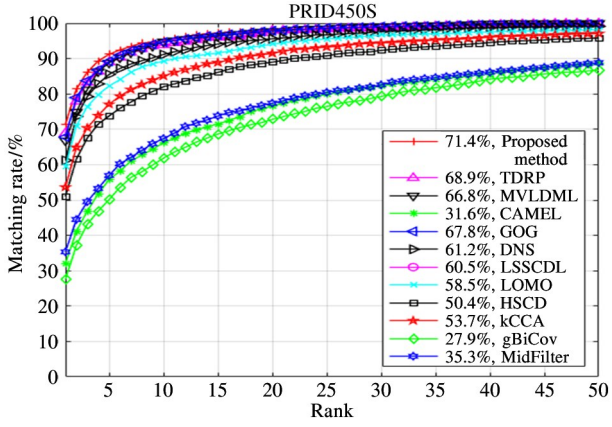

Fig. 6 CMC curves

CUHK01 dataset: the performance of the wLOMO has been declining with a increasing, because the person background information is more complex than the first two datasets in Fig. 7, and the introduction of mean information leads to performance degradation. Thus, the combination with GOG can strengthen the feature expression and weaken the error caused by mean information.

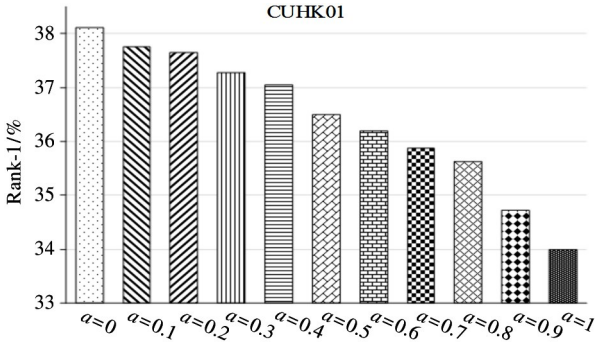

Fig. 7 Rank-1 matching rates

The compared methods and their matching rates on CUHK01 dataset are shown in Table 3 and Fig. 8. Considering that each person in the CUHK01 dataset contains four images, the first two images contain one front/back view, the last two images contain one side view, and the overall difference between them is little. Therefore, in the experiment, one is randomly selected from the foreground and background images of each person, and one is randomly selected from the side images of each person. The training sets contain 486 pairs of person images, and the test sets contain 485 pairs of person images. As listed in Table 3, the performance of proposed method is better than other methods, outperforming the second-best method with improvements of 5.65%.

Table 3 Comparison of Rank results with other methods on CUHK01 dataset

Methods	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
MidFilter[22]	28.26	51.54	60.86	70.89
gBiCov[23]	21.12	50.43	58.15	64.77
kCCA[24]	29.25	56.37	67.50	73.22
HSCD[25]	35.88	58.16	69.15	74.16
LOMO[16]	38.10	62.35	71.81	81.67
LSSCDL[26]	42.77	69.64	79.23	86.14
DNS[27]	43.40	68.35	77.75	85.38
GOG[17]	46.80	70.35	78.82	86.26
CAMEL[28]	30.02	52.10	64.55	81.27
PE[33]	39.20	-	-	-
Proposed method	52.45	75.61	83.57	89.67

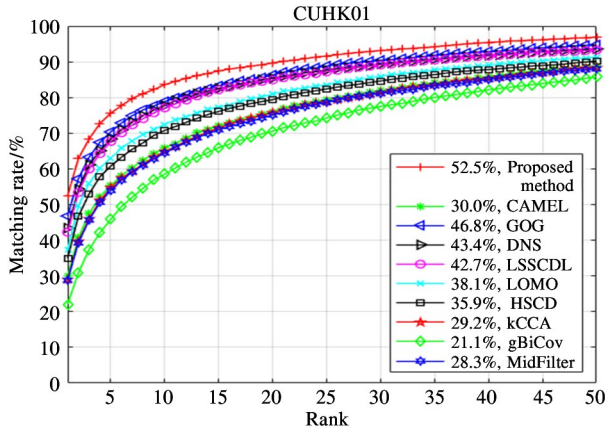

Fig. 8 CMC curves

4.2.2 Evaluation of the sample determination

The proposed method has achieved state-of-the-art performance, with inputting the output features of the mapping space into XQDA in the above experiment. Then, in order to verify the effectiveness of the proposed sample determination method, the output features of the mapping space are sent to XQDA and k-XQDA respectively to compare the performance of the methods. The experiment results are shown in Table 4, Table 5 and Table 6, in which the size of samples is the number of sample.

VIPeR dataset: in Table 4, when the size of training set samples is gradually increased, Rank-1 of the two metric learning methods is also increasing during the experiment on the VIPeR dataset. According to the Rank-1, the matching rate of XQDA is greater than that of k-XQDA even with the increase of training set samples. However, the increase of XQDA is 6.87% and 15.3%, the increase of k-XQDA is 7.97% and 16.93%. The increase extent of k-XQDA is greater than that of XQDA. Thus, when the size of training set samples increases to a certain size, k-XQDA can show better accuracy than XQDA.

Table 4 Ranks matching rates versus different size of samples on VIPeR dataset

Size of samples	Methods	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
316	Proposed method + XQDA	50.63	80.53	88.89	95.66
	Proposed method + k-XQDA	48.32	78.16	87.69	94.91
400	Proposed method + XQDA	57.50	86.29	93.66	98.23
	Proposed method + k-XQDA	56.29	86.47	93.41	97.97
532	Proposed method + XQDA	72.80	94.40	97.80	99.80
	Proposed method + k-XQDA	73.22	95.37	98.65	99.84

Table 5 Ranks matching rates versus different size of samples on PRID450S dataset

Size of samples	Methods	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
225	Proposed method + XQDA	71.42	91.20	95.20	97.96
	Proposed method + k-XQDA	66.27	89.11	94.58	97.87
300	Proposed method + XQDA	77.80	94.40	97.46	99.20
	Proposed method + k-XQDA	74.33	93.33	96.73	98.67
436	Proposed method + XQDA	94.12	99.65	100	100
	Proposed method + k-XQDA	95.27	99.81	100	100

Table 6 Ranks matching rates versus different size of samples on CUHK01 dataset

Size of samples	Methods	Rank-1/%	Rank-5/%	Rank-10/%	Rank-20/%
300	Proposed method + XQDA	45.63	68.79	77.08	84.49
	Proposed method + k-XQDA	43.28	66.62	75.81	83.46
486	Proposed method + XQDA	52.45	75.61	83.57	89.67
	Proposed method + k-XQDA	54.25	78.31	86.25	92.10

PRID450S dataset: when the size of samples in the training set increases from 225 to 300 and 436, the Rank-1 of XQDA is better than that of k-XQDA, reported in Table 5. In terms of the extent of Rank-1 increases, XQDA increases by 6.38% and 16.32%, k-XQDA increases by 8.06% and 20.94%. According to the experiment results on PRID450S dataset, when the size of training sets increases to a certain size, the Rank-1 of k-XQDA can exceed that of XQDA.

CUHK01 dataset: the output features of the mapping space are calculated by XQDA and k-XQDA respectively on CUHK01 dataset. When the size of train-

ing set samples is 486, the Rank-1 of k-XQDA exceeds that of XQDA by 1.8%, reported in Table 6.

In summary, when the size of training set samples is about 532, the performance of k-XQDA is better than that of XQDA in Table 4. Here, the k-XQDA can obtain better results. When the size of training sets is less than 532, the performance of XQDA is better than that of k-XQDA. On PRID450S dataset, when the size of training set samples is bigger than 436, the performance of k-XQDA method is better than that of XQDA method, and better results can be obtained by using k-XQDA. When the size of training sets is less than 436, the performance of XQDA is better than that of k-XQDA in Table 5. According to the results in Table 6, when person Re-ID is conducted on CUHK01 dataset, the size of training set samples is about 486, k-XQDA can obtain good results.

5 Conclusion

Based on multi-feature extraction, an effective feature mapping space and a sample determination method is proposed to solve the problem of visual ambiguities in person re-identification. The feature mapping space simplifies the process of complex feature extraction, which takes the basic features in person images as input and outputs the mapped features through the feature mapping space. The mapped features are discriminated by the proposed metric learning method to complete the similarity ranking. Compared with the existing correlation methods, the proposed method improves matching rate effectively. In the future, it is proposed to further study the determination method of metric learning and optimize the performance of the algorithm.

References

[1] LI H, XU J, ZHU J, et al. Top distance regularized projection and dictionary learning for person re-identification [J]. *Information Sciences*, 2019, 502:472-491

[2] FARENZENA M, BAZZANI L, PERINA A, et al. Person re-identification by symmetry driven accumulation of local features[C]//Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010: 2360-2367

[3] WEINBERGER K Q, SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. *Journal of Machine Learning Research*, 2009, 10(2): 207-244

[4] GRAY D, TAO H. Viewpoint invariant pedestrian recognition with an ensemble of localized features[C]//Proceedings of the 2008 European Conference on Computer Vision, Marseille, France, 2008: 262-275

[5] MA B, SU Y, JURIE F. Bicov: a novel image representation for person re-identification and face verification[C]//Proceedings of the 2012 British Machine Vision Conference, Surrey, UK, 2012: 1-11

[6] PEDAGADI S, ORWELL J, VELASTIN S, et al. Local fisher discriminant analysis for pedestrian re-identification [C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013: 3318-3325

[7] LIAO S, HU Y, ZHU X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 2197-2206

[8] MATSUKAWA T, OKABE T, SUZUKI E, et al. Hierarchical Gaussian descriptor for person re-identification[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1363-1372

[9] MA B, SU Y, JURIE F. Local descriptors encoded by fisher vectors for person re-identification[C]//Proceedings of the 2012 European Conference on Computer Vision Workshops and Demonstrations, Florence, Italy, 2012: 413-422

[10] MA B, SU Y, JURIE F. Discriminative Image Descriptors for Person Re-identification[M]. London: Springer-Verlag, 2014: 23-42

[11] CHEN Y, ZHENG W, LAI J. Mirror representation for modeling view-specific transform in person re-identification[C]//Proceedings of the 2015 International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015: 3402-3408

[12] WU S, CHEN Y, LI X, et al. An enhanced deep feature representation for person re-identification[C]//Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision, Lake Placid, USA, 2016: 1-8

[13] SU C, LI J, ZHANG S, et al. Pose-driven deep convolutional model for person re-identification[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 3980-3989

[14] KÖESTINGER M, HIRZER M, WOHLHART P, et al. Large scale metric learning from equivalence constraints [C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 2288-2295

[15] ZHAO C, WANG X, WONG W K, et al. Multiple metric learning based on bar-shape descriptor for person re-identification[J]. *Pattern Recognition*, 2017, 71: 218-234

[16] XIONG F, GOU M, CAMPS O, et al. Person re-identification using kernel-based metric learning methods[C]//Proceedings of the 2014 European Conference on Computer Vision, Zurich, Switzerland, 2014: 1-16

[17] NGUYEN B, BAETS B D. Kernel distance metric learning using pairwise constraints for person re-identification [J]. *IEEE Transactions on Image Processing*, 2019, 28: 589-600

[18] ALI T M F, CHAUDHURI S. Cross-view kernel similarity metric learning using pairwise constraints for person re-identification[EB/OL]. https://arxiv.org/abs/1909.11316v1:arXiv, (2019-09-25),[2021-09-22]

[19] GRAY D, BRENNAN S, TAO H. Evaluating appearance models for recognition, reacquisition, and tracking[C]// Proceedings of the 10th International Workshop on Performance Evaluation for Tracking and Surveillance, Riode Janeiro, Brazil, 2007: 41-47

[20] ROTH P M, HIRZER M, KÖESTINGER M. et al. Mahalanobis distance learning for person re-identification [J]. *Person Re-Identification*, 2014, 2014: 247-267

[21] LI W, ZHAO R, WANG X. Human reidentification with transferred metric learning[C]// Proceedings of the 2012 Asian Conference on Computer Vision, Daejeon, Korea, 2012: 31-44

[22] ZHAO R, OUYANG W, WANG X. Learning mid-level filters for person re-identification[C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 144-151

[23] MA B, SU Y, JURIE F. Covariance descriptor based on bio-inspired features for person re-identification and face verification[J]. *Image and Vision Computing*, 2014, 32 (6-7):379-390

[24] LISANTI G, MASI I, BIMBO A D. Matching people across camera views using kernel canonical correlation analysis[C]// Proceedings of the 2014 International Conference on Distributed Smart Cameras, Venezia Mestre, Italy, 2014: 10:1-10:6

[25] ZENG M, WU Z, TIAN C, et al. Efficient person re-identification by hybrid spatiogram and covariance descriptor[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, USA, 2015: 48-56

[26] ZHANG Y, LI B, LU H, et al. Sample-specific SVM learning for person re-identification[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1278-1287

[27] ZHANG L, XIANG T, GONG S. Learning a discriminative null space for person re-identification[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1239-1248

[28] YU H, WU A, ZHENG W. Cross-view asymmetric metric learning for unsupervised person re-identification[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 994-1002

[29] JIA J, RUAN Q, JIN Y, et al. View-specific subspace learning and re-ranking for semi-supervised person re-identification [J]. *Pattern Recognition*, 2020, 108: 107568

[30] CHEN Y C, ZHU X, ZHENG W S, et al. Person re-identification by camera correlation aware feature augmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(2): 392-408

[31] YANG X, WANG M, TAO D. Person re-identification with metric learning using privileged information [J]. *IEEE Transactions on Image Processing*, 2018, 27(2): 791-805

[32] LI H, XU J, ZHU J, et al. Top distance regularized projection and dictionary learning for person re-identification [J]. *Information Sciences*, 2019, 502: 472-491

[33] ZHUANG W, WEN Y, ZHANG S, et al. Joint optimization in edge-cloud continuum for federated unsupervised person re-identification [C]// Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 2021: 433-441

HOU Wei, born in 1991. He is a Ph. D. candidate at the School of Artificial Intelligence, Henan University. He received his B. S. and M. S. degrees from Henan University in 2014 and 2019 respectively. His research interests include target tracking and complex system modeling and estimation.