Fusing Multiple Statistical Features via Explicit Feature Mapping for Person Re-Identification

Hongli Zhang, Honggang Zhang, Jianlou Si

Beijing University of Posts and Telecommunications, Beijing 100876, China
zhanghl_bupt@163.com

Abstract: Person re-identification (Re-ID) across non-overlapping camera views is one of the challenging problems in surveillance video analysis. In this paper, we propose to combine multiple statistical features via explicit kernel feature mapping, and learn a linear metric model by local fisher discriminant analysis (LFDA) for person Re-ID. To strengthen the robustness of our representation, three complementary statistical characteristics, including histogram-like features, covariance matrix and expectation vector, were extracted from multiple spatial scales for each person image. Experimental results show that the proposed method works effectively on the popular benchmark data sets VIPeR and CUHK01 and yield impressive performance measured with Cumulative Match Characteristic curves (CMC).

Keywords: Feature fusion; Feature mapping; Person re-identification

1 Introduction

Person re-identification (Re-ID) aims at automatically matching people across disjoint camera views, which plays a critical role in many visual surveillance tasks, such as human retrieval, long-term tracking, and behavior analysis. Over the last decade, many efforts have been devoted to addressing this problem via either constructing an invariant feature representation or learning a discriminative distance metric.

However, due to the visual ambiguity caused by clothing similarity among different people and appearance variations from different view angle, pose, illumination, and occlusion, it is still difficult to make accurate matching. Fortunately, this visual ambiguity can be alleviated by comprehensively utilizing multiple features. To this end, one way is to concatenate multiple complementary features to a comprehensive representation directly; another way is to utilize multiple kernel learning, where the individual's similarity measured via each feature is computed in the corresponding non-linear kernel space and different features are fused in the kernel level. While, it should be noted that, previous works prefer to only use different histogram-like features (e.g. color or gradient histogram) to represent the appearance, which are not enough to describe people's characteristics; and the original kernel methods always have high storage complexity, even worse for multiple features.

In this paper, we propose to address the mentioned problems by fusing multiple statistical features via explicit feature mapping for person Re-ID. Specifically, various multi-scale statistical features are extracted firstly (including histogram-like features, covariance matrix and expectation vector); and then they are mapped to the higher dimensional measurement space with the corresponding kernel function via explicit mapping [1], in which the inner product approximates the kernel distance well; after that a common used linear metric learning model LFDA [8] can be applied to explore a proper metric space effectively. In addition, extensive experiments on two benchmark data sets VIPeR and CUHK01 demonstrate the effectiveness of our method. The overview of our proposal is showed in Fig. 1.

The rest of this paper is organized as follows. In Sec. 2, we review the related works. In Sect. 3, we present our proposal. Experimental evaluations are provided in Sec. 4. We conclude with discussion in Section 5.

2 Related work

Person Re-ID has attracted many attentions in the computer vision community and many methods have been proposed. Roughly, the existing works can be divided into two categories: feature representation based methods (e.g. [2,3,4,21]) and metric learning based methods (e.g. [5,6,8,9]). For example, [2] proposed an effective feature representation called Local Maximal Occurrence, which maximized horizontal occurrence of local color histograms and was stable for viewpoint changing; [21] extracted three complementary kinds of features based on the localization of perceptual relevant human parts, driven by asymmetry/symmetry principles; Zheng et. al. [6] formulated person Re-ID as a relative distance comparison learning problem, where an optimal similar measurement was learned; Pedagadi et. al. [8] presented a manifold learning approach, in which the dimension reduction and distance metric learning were simultaneously conducted.

To resist the diverse interferences causing visual ambiguity, multiple feature descriptors are always jointly utilized in person Re-ID. For instance, [2, 6, 9] combined color histograms from different color space as the color visual cues. Obviously, the representation

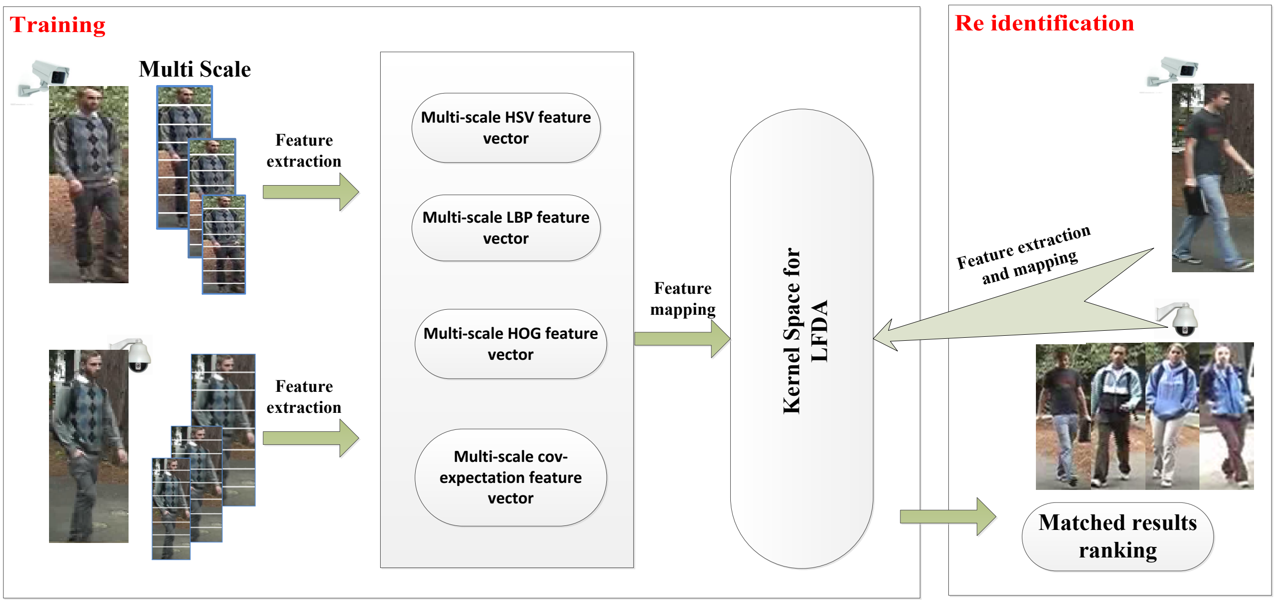
Training
Multi Scale
Feature extraction
Feature extraction
Multi-scale HSV feature vector
Multi-scale LBP feature vector
Multi-scale HOG feature vector
Multi-scale cov-expectation feature vector
Feature mapping
Kernel Space for LFDA
Re identification
Feature extraction and mapping
Matched results ranking

Figure 1 System Overview

could be further enriched by introducing more statistical features such as covariance matrix [13] or expectation vector. Besides, although [9] presented many kernel versions for linear metric learning models, it did not implement multi-feature fusion in the kernel-level and yet had high storage complexity.

In [1], explicit feature mapping was proposed to approximate large scale non-linear kernel models using suitable feature maps, which benefited from both the simplicity and speed of linear model and the non-linearity of kernel space. Motivated by this idea, we adopt the explicit feature mapping to approximate the multiple kernel learning based multi-feature fusion method for person Re-ID.

3 The proposed method

In this section, we will give the details of our method from three successive steps: multiple statistical feature extraction, explicit feature mapping and feature fusion, and linear metric learning.

3.1 Feature extraction

As in Fig. 1, we implemented three histogram-like features (i.e. color histogram, HOG, and LBP) and one covariance-expectation feature (i.e. combination of covariance matrix and mean vector) for each person image. Each feature was calculated in a multi-scale convention. Specifically, the person image is first resize to 128*48; and then divided into dense half overlapping patches with 3 different sizes, i.e. 8*8, 16*16, and 32*32; finally, local features are extracted from each patch and aggregated to a whole feature vector. In addition, after the feature extraction, we apply the principle component analysis (PCA) to reduce the dimensionality of each feature. Some additional remarks for feature extraction are given as below.

3.1.1 Histogram-like features

Color histogram feature: Color histogram depicts the color distribution in the three-channel color image. Considering HSV color space is much more fit for human's visual characteristics and it is more robust to illumination, we extract the HSV color histogram in this paper.

HOG feature: It describes the gradient orientation distribution of an intensity image. To further combine color and texture features, the HOG features are extracted from YUV color images and computed by concatenating the orientation histogram of each intensity channel.

LBP feature: The distribution of local binary pattern is represented by the LBP histogram. In this paper, we also compute the LBP histogram in YUV color space by concatenating each intensity one.

3.1.2 Covariance-expectation feature

In addition to utilizing the histograms to approximate the feature distribution within a certain region, we also characterize the distribution by some other statistics, such as expectation and variance.

The covariance matrix was first introduced to represent the correlation between multiple features and was demonstrated good performance in pedestrian detection task in [13,14]. Different from the former works, we utilize the covariance matrix in person Re-ID incorporated with the mean vector of each feature, which improves the performance considerably than the single covariance feature. In practice, we ignore the geometric configuration of covariance matrix and vectorize it by simply stacking the upper triangular part into a feature vector.

In this paper, the multiple low-level features used to compute the covariance and mean are as:

$$[x, y, \boldsymbol{D_x}, \boldsymbol{D_y}, \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{V_1}, \boldsymbol{L}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{V_2}]$$

where x and y represent the pixel location, $\boldsymbol{D_x}$ and $\boldsymbol{D_y}$ are intensity gradient in two orientations, the rest are color values extracted from different color spaces.

3.2 Feature mapping and fusion

Because of most visual features have complex nonlinear structures, it is difficult to discriminate similar individuals via a common linear model. To solve this problem, kernel based methods, such as KLFDA [9] and multiple kernel learning (MKL) [15], are widely applied in computer vision tasks, where the non-linearity of features is explored in the measurement distance space. What is more, MKL can naturally achieve multi-feature fusion. While these methods demand for high computation and memory capacity.

To incorporate the non-linearity and the convenience in multi-feature fusion of kernel methods, without being affected by their storage complexity, we utilize the explicit feature mapping as in [1] to approximate the kernel function $k(\boldsymbol{x}, \boldsymbol{y})$. It is found experimentally and empirically that radial basis function (RBF) kernels are suitable for exploring the non-linear structure of histogram-like and covariance-expectation features, so we focus on the explicit feature maps for RBF kernels.

A generalized RBF kernel (GRBF) has the formulation as followed

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{1}{2\sigma^2} D^2(\boldsymbol{x}, \boldsymbol{y})\right), \qquad (1)$$

where $D^2(\boldsymbol{x}, \boldsymbol{y})$ is a particular distance metric. As demonstrated in [1], GRBF kernels based on additive distances such as χ^2 can be computed by combining the random Fourier features and the homogeneous features for the additive kernels. Specifically, let $\widehat{\Psi}(\boldsymbol{x}) \in \mathbb{R}^n$ be an approximated feature map for the metric $D^2(\boldsymbol{x}, \boldsymbol{y})$ and let $\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_n \in \mathbb{R}^n$ be sampled from the Gaussian density $\kappa(\boldsymbol{\omega})$. Then the approximated feature map for GRBF is

$$\widehat{\Psi}_{GRBF}(\boldsymbol{x}) = \frac{1}{\sqrt{n}}\left[e^{-i<\omega_1, \widehat{\Psi}(x)>}, \dots, e^{-i<\omega_n, \widehat{\Psi}(x)>}\right]^T. \quad (2)$$

In our work, we utilize RBF-χ^2 and Gaussian kernel for histogram-like features and covariance feature separately, which correspond χ^2 and Euclidean distance metric respectively. As in [1], the approximated feature map for χ^2 is computed as

$$\widehat{\Psi}(\boldsymbol{x}) = \bigoplus_{l=1}^{d} \Psi(x_l), \qquad (3)$$

$$\Psi(x_l) = e^{i\omega \log x_l}\sqrt{x_l \operatorname{sech}(\pi\omega)}, \qquad (4)$$

where d is the dimension of the feature and $\oplus$ is the direct sum; and the approximated feature map for Euclidean distance is

$$\widehat{\Psi}(\boldsymbol{x}) = \boldsymbol{x}. \qquad (5)$$

According to the Eq. (2)(3)(4), each feature can be approximately mapped into the proper kernel space. Then multiple features are directly concatenated into the final feature representation, and linear metric learning model can be used.

3.3 Metric learning

Different from the linear discriminant analysis, LFDA introduces the local affinity matrix assisting in separating the different classes apart and compacting the same classes together simultaneously. Let $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{1, \dots, C\}$ denote the final fused feature and the corresponding label respectively. The within-class and between-classes scatter matrices are defined as

$$\boldsymbol{S}^W = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^W (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T, \qquad (6)$$

$$\boldsymbol{S}^B = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{A}_{i,j}^B (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T, \qquad (7)$$

where $\boldsymbol{A}_{i,j}^W$ and $\boldsymbol{A}_{i,j}^B$ are the corresponding local affinity matrix.

An optimal transformation matrix can be found as

$$\boldsymbol{Q}_* = \operatorname{argmax}_{\mathbf{Q}} \left((\boldsymbol{Q}\boldsymbol{S}^W\boldsymbol{Q})^{-1}(\boldsymbol{Q}\boldsymbol{S}^B\boldsymbol{Q})\right), \qquad (8)$$

where $\boldsymbol{Q} \in \mathbb{R}^{d0 \times d}$. The estimation of $\boldsymbol{Q}_*$ can be solved as a generalized eigenvalue problem.

Considering the ranking defect of within class scatter matrix $\boldsymbol{S}^W$, a regularized $\tilde{\boldsymbol{S}}^W$ defined below is used instead

$$\tilde{\boldsymbol{S}}^W = (1-\beta)\boldsymbol{S}^W + \beta\frac{Tr(\boldsymbol{S}^W)}{n}\boldsymbol{I}, \qquad (9)$$

where β ($0 \le \beta \le 1$) is the regularization parameter, and we set it to 0.5 by default.

4 Experiments

4.1 Data sets and evaluation protocol

The proposed method is evaluated on two popular benchmark data sets: VIPeR and CUHK01. They both provide the pedestrian images across two non-overlapping cameras with large variations in illumination, viewpoints and poses. Specifically, VIPeR contains 632 pairs of person images; CUHK01 contains 971 individuals and has two images for each person from each camera. We randomly and evenly divide each data set into training and testing set, and randomly select two images for each person as probe and gallery image separately. The experimental results are reported by Cumulative Match Characteristic curves (CMC). Considering the randomness, this process is repeated 10 trials and the average result is obtained.

4.2 Results

We compare the final fused feature, and other 4 single descriptors including multiple scale LBP feature, HOG feature, HSV feature and covariance-expectation feature on data sets VIPeR and CUHK01. Experimental results are showed with CMC curves in Fig. 2 and Fig. 3. We also compare our method with three popular metric learning methods: KLFDA [9], LFDA [16], LADF [17] and other three state-of-the-art non-metric learning methods SCNCD [18], SDALF [19], SDC [20]. The comparison is listed in Table I.

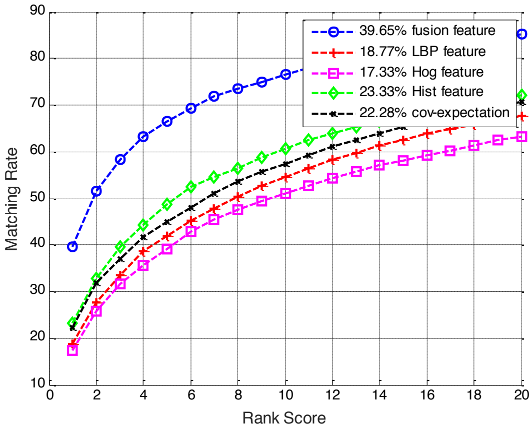

Figure 2 Performance of the proposed method on CUHK01

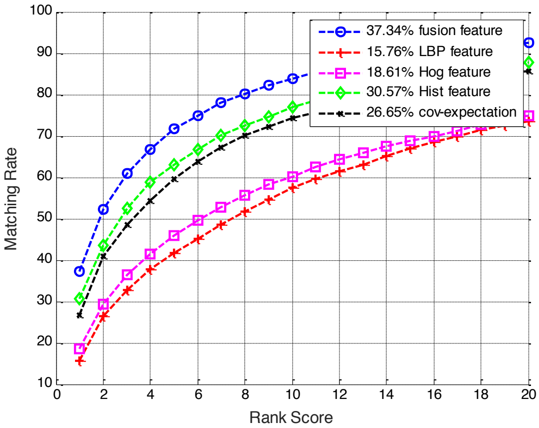

Figure 3 Performance of the proposed method on VIPeR

Table I Matching Rate on VIPeR

Method	Rank1	Rank5	Rank10	Rank20
LFDA	21.5%	49.6%	64.6%	79.1%
KLFDA(RBF-χ2)	32.3%	65.8%	79.7%	90.9%
LADF	30.1%	63.2%	77.4%	88.1%
SCNCD	33.7%	62.7%	74.8%	85%
SDALF	19.9%	38.9%	49.4%	65.7%
SDC	26.7%	50.7%	62.4%	76.4%
Our method (color feature)	**30.1%**	**63.1%**	**76.9%**	**83.1%**
Our method	**37.3%**	**71.7%**	**83.7%**	**92.5%**

In our experiment, fusion feature greatly improves the matching accuracy than each single feature, and rank1 matching rate is improved by about 19.23% on VIPeR and 14.44% on CUHK01. Particularly, HSV color descriptor contributes the most to the improvement, and that is reasonable since the color information is the most salient characteristic. While HOG and LBP features, which represent the texture information, are difficult to capture these details of individuals due to the poor resolution and illumination, so they contribute less than others. As we can see, the covariance-expectation feature we extracted is also effective, and it almost performs as well as the color features.

In Table I, we compare our method with other popular metric learning and non-metric learning methods. It is obvious that our method with feature fusion via explicit feature mapping catches up with or even surpasses others. Considering that RBF-$\chi 2$ kernel LFDA gives the best performance and only color histogram features are used in [9], we compare it with our method using single color feature. As Table I shows that, our single color feature with explicit feature maps performs almost as well as the exact kernel method but in exchange of less training and testing time in experiment, which demonstrated the effectiveness and efficiency of the explicit kernel approximation method in person Re-ID.

5 Conclusions

In this paper, considering the visual ambiguity of appearance representation and storage complexity of non-linear metric learning in person Re-ID, we propose to combine multiple statistical features via explicit kernel feature mapping, and learn a linear metric model by local fisher discriminant analysis (LFDA). Extensive experiments conducted on two popular benchmark data sets VIPeR and CUHK01 demonstrate simplicity and effectiveness of our method.

Acknowledgements

References

[1] Vedaldi A, Zisserman A. Efficient Additive Kernels via Explicit Feature Maps. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 34(3):480-492.

[2] S. Liao, Y. Hu, X. Zhu and S. Z. Li , Person re-identification by local maximal occurrence representation and metric learning, CVPR, pp.2197 -2206

[3] Bak S, Corvee E, Brémond F, et al. Person re-identification using spatial covariance regions of human body parts. IEEE Advanced Video and Signal Based Surveillance, 2010: 435-440.

[4] Liu C, Gong S, Chen C L, et al. Person Re-identification: What Features Are Important? ECCV, pp. 391-401.

[5] Dikmen M, Akbas E, Huang T S, et al.. Pedestrian recognition with a learned metric. ACCV, 2010: 501-51.

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

[7] M. Sugiyama, Local fisher discriminant analysis for supervised dimensionality reduction, in: Proceedings of

the 23rd International Conference on Machine Learning, ACM, 2006, pp. 905–912

[8] S. Pedagadi, J. Orwell, S. Velastin, et al., Local fisher discriminant analysis for pedestrian re-identification, CVPR, 2013, pp. 3318–3325.

[9] F. Xiong, M. Gou, O. Camps, et al., Person re-identification using kernel-based metric learning methods, Computer Vision - ECCV 2014. pp.1–16

[10] Chan C H, Kittler J, Tahir M A. Kernel Fusion of Multiple Histogram Descriptors for Robust Face Recognition. 2010. pp.718-727.

[11] Chan, C., Kittler, J., Messer, K.: multiple scale local binary pattern histograms for face recognition. In: ICB. Volume 4642. (2007) 809–818

[12] Chan, C., Kittler, J., Poh, N., Ahonen, T., Pietik¨ainen, M.: (multiscale) local phase quantization histogram discriminant analysis with score normalisation for robust face recognition. In: VOEC. (2009) 633–640

[13] Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on Riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. **30**(10) (2008) 1713–1727

[14] Paisitkriangkrai S, Shen C, Hengel A V D. Strengthening the Effectiveness of Pedestrian Detection with Spatially Pooled Features. Lecture Notes in Computer Science, 2014, 8692:546-561.

[15] Sonnenburg S, Rätsch G, Schäfer C, et al. Large Scale Multiple Kernel Learning. Journal of Machine Learning Research, 2010, 7(2006):1531-1565.

[16] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, “Local fisher discriminant analysis for pedestrian re-identification,” in *CVPR*, 2013.

[17] Z. Li, S. Y. Chang, F. Liang, T. S. Huang, L. L. Cao, and J. R. Smith, “Learning locally-adaptive decision functions for person verification,” in *CVPR*, 2013.

[18] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li, “Salient color names for person re-identification,” in *ECCV*, 2014.

[19] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, “Person reidentification by symmetry-driven accumulation of local features,” in *CVPR*, 2010.

[20] R. Zhao, W. L. Ouyang, and X. G. Wang, “Unsupervised salience learning for person re-identification,” in *CVPR*, 2013.

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