Multiple-Shot Person Re-identification by Features Learned from Third-party Image Sets

Yanna Zhao[1, 2], Lei Wang[1, 2], Xu Zhao[2]* and Yuncai Liu[2]
[1] School of Information Science and Engineering, Shandong University
Jinan 250100, China
[2] School of Automation, Shanghai Jiao Tong University
Shanghai 200240, China
[e-mail: zhaoxu@sjtu.edu.cn]
*Corresponding author: Xu Zhao

Received November 8, 2014; accepted January 21, 2015; published February 28, 2015

Abstract

Person re-identification is an important and challenging task in computer vision with numerous real world applications. Despite significant progress has been made in the past few years, person re-identification remains an unsolved problem. This paper presents a novel appearance-based approach to person re-identification. The approach exploits region covariance matrix and color histograms to capture the statistical properties and chromatic information of each object. Robustness against low resolution, viewpoint changes and pose variations is achieved by a novel signature, that is, the combination of Log Covariance Matrix feature and HSV histogram (LCMH). In order to further improve re-identification performance, third-party image sets are utilized as a common reference to sufficiently represent any image set with the same type. Distinctive and reliable features for a given image set are extracted through decision boundary between the specific set and a third-party image set supervised by max-margin criteria. This method enables the usage of an existing dataset to represent new image data without time-consuming data collection and annotation. Comparisons with state-of-the-art methods carried out on benchmark datasets demonstrate promising performance of our method.

Keywords: Person re-identification, Appearance modeling, Max-margin feature learning, Covariance matrix

A preliminary version of this paper appeared in IAPR ACPR 2013, November 5-8, Okinawa, Japan. This version includes a concrete analysis and more experimental results. This research has been partially supported by the grants of China 973 project 2011CB302203, NSFC 61375019, NSFC 61273285 and NSFC 61105001.

http://dx.doi.org/10.3837/tiis.2015.02.017 **ISSN : 1976-7277**

1. Introduction

In many video-surveillance applications, it is desirable to determine if a presently visible person has already been observed somewhere else in the network of cameras. This kind of problematic is commonly known as person re-identification. It represents a valuable task in video surveillance scenarios such as airports, streets, parking lots and shopping mall, where cameras with non-overlapping views are mounted. A typical re-identification method aims to establish the correspondence between probe and gallery sets. A gallery is a set of ID templates (in the form of feature vectors), representing objects in a database. These gallery feature vectors are used to recognize unknown feature vectors, which come from a probe set. Each feature vector in the probe set has a corresponding ID template in the gallery set.

Due to coarse resolution or low frame rate of cameras in video surveillance, biometric cues such as face, iris or gait might not be available. Person re-identification based on appearance information is of particular interest. In this case, it is assumed that objects do not change their clothes between different sightings. Generally speaking, appearance-based re-identification can be categorized into single-shot methods and multiple-shot methods depending on the sample size [1]. In single-shot methods, each object is represented by only one image. However, in many surveillance scenarios, it is not only possible but also easy to get multiple images for both the probe and gallery objects in field of view. Re-identification approaches using multiple images or image set are referred to as multiple-shot methods. Intuitively, more images contain more information, which could result in better performance. This intuition has been proved by recent works from both person re-identification and face recognition [1-5].

Compared with single-shot methods, recognition approaches using image set can extract more reliable and comprehensive information. However, using image set for recognition brings new challenges because images in the same set are probably taken from different viewpoints and illumination conditions, where intra-personal variations may be more significant than inter-personal variations. How to make full use of the image set while discriminate the inter-personal and intra-personal variations simultaneously are the key issues. Previous recognition literatures using multiple images generally focused on elaborately designed features, such as color and edgel histograms, weighted color histograms, maximally stable color regions, epitomic analysis, region covariance [1-6], etc. Although these features are directly compatible with image set based matching algorithms, it is hard to make them both representative and discriminative. The CHISD method [7] was proposed for face recognition by finding the between-set distance. Compared with the feature designing methods, this method can make a better use of the image sets, but the discriminative power of the image representation has not been fully exploited.

In this paper, we design a novel unsupervised feature learning method to strengthen the discriminative power of image representation. We start to construct feature representations from densely sampled image patches because patch based representation provides a well suited mechanism to capture the visual information of the object appearance. Each patch is represented by the combination of chromatic information and region covariance feature. Different from [5], we transform covariance matrices from the convex cone to the vector space by matrix logarithm and half-vectorization [8], so that combing the covariance feature with other features is possible. Besides, standard machine learning algorithms in the vector space can be utilized.

Our feature learning method takes inspirations from the support vector framework [9] in representing image sets [10, 11]. The basic idea behind these methods is that, the decision boundary between a pair of image sets contains rich information for classification/recognition. Each component of the normal vector of the decision boundary measures the contribution of the corresponding feature dimension for classification/recognition. Motivated by these considerations, we propose our third-party feature learning method for person re-identification. **Fig. 1** illustrates the framework of our method. In this framework, multiple image sets taken from different objects are firstly built to serve as the third-party reference. Both probe and gallery image sets are compared with this common point of reference, and the differences are utilized later to formulate feature vectors for both of the image sets. In our method, the differences are measured by the decision boundary between a given probe/gallery set and a third-party set. The boundary can be obtained by maximizing the margin between these two kinds of sets. Finally, the average of all decision boundaries obtained with respect to different third-party image sets gives rise to the informative representation for the probe/gallery set.

The contributions of this work can be summarized as follows. First, we propose a multiple-shot person re-identification method by exploring the discriminative features contained in the decision boundary between a given image set and the third-party image sets. The semantic and temporal information contained in the image set has already been automatically incorporated by the features. Second, we propose to use the covariance features in the Log-Euclidean space for person re-identification. By performing matrix logarithm and half-vectorization, covariance matrices are transformed to the vector space, which is then concatenated with other features. Finally, we use our third-party feature learning method to learn the most discriminative features from the transformed covariance feature space in a unified framework supervised by max-margin criteria, which is more efficient compared with learning in the Riemannian space.

The rest of the paper is organized as follows. Section 2 describes the state-of-the-art re-identification methods. Section 3 details the proposed method, and Section 4 reports the related experiments. Finally, conclusions and future perspectives are given in Section 5.

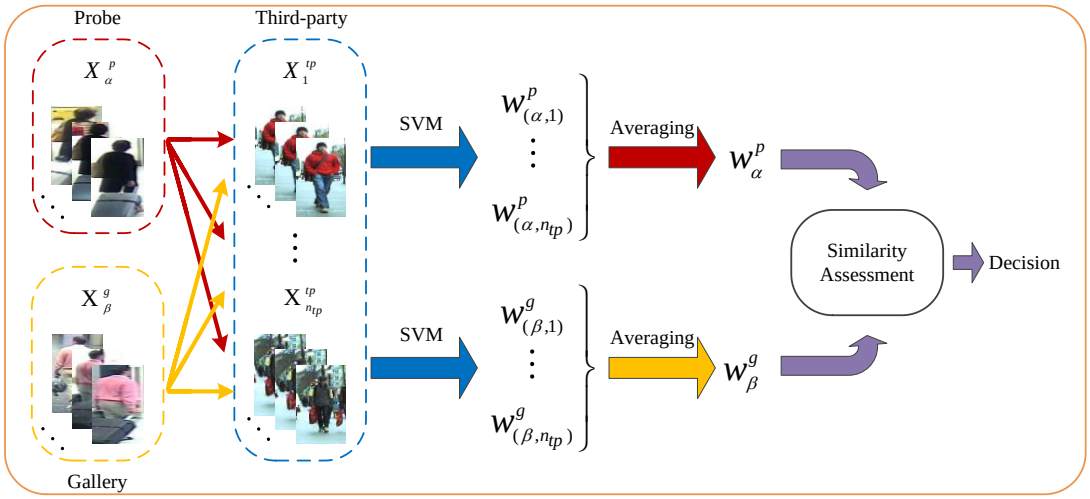


Fig. 1. The framework of our SVM-LCMH method for person re-identification. See Section 3 for details.

2. Related Works

Recent studies on person re-identification concentrate on appearance-based methods. Appearance-based re-identification methods focus either on distance learning regardless of the image representation [12-15, 21-25], or on appearance modeling [1-4, 16-19], producing a discriminative and invariant representation for human images. Learning methods utilize training data to exploit strategies that maximizing inter-class variation at the same time minimizing intra-class variation. Instead, feature-oriented methods focus on invariant image representations, which should be able to deal with pose and illumination changes. Further categorization of appearance-based methods distinguishes the single-shot and the multiple-shot methods. The former extracts appearance information employing a single image [6, 16-19], while the latter uses multiple images of the same object to acquire a robust representation [1-5].

Concerning single-shot methods, shape and appearance context was proposed in [17]. Spatial distributions of appearance relative to body parts were modeled so that discriminative features robust to misalignment can be extracted. This method is not flexible enough and only applicable if the system considers a frontal viewpoint. In [16], human body parts were detected to establish the correspondence between appearances. Covariance descriptor [20] was tailored to represent the appearance of each part. In order to improve discriminative matching of two sets of body parts, a spatial pyramid scheme was adopted. The color distribution structure of the human body was shown to be invariant and discriminative for recognition [6]. Fisher vector was employed in [18] to encode higher order statistics of local features using generative information. In [19], a descriptor combining Gabor filters and covariance descriptor was developed to handle illumination changes and background variations. These methods focused on feature design, but did not exploit the rich information of the features by discriminative learning.

Discriminative models like boosting and SVM are extensively used for feature learning [21-24]. An ensemble of localized features (ELF) was proposed in [21]. Instead of characterizing human appearance by a specific feature, a machine learning algorithm constructed a model by combining spatial and color information. The re-identification was reformulated as a ranking problem in [22]. An informative subspace was learned in which the potential true match gets highest ranking. Similarly, in [23], a discriminative model was obtained by projecting a high dimensional signature composed of multiple features into a low-dimensional space using a statistical tool called Partial Least Squares (PLS). Another branch of methods use metric learning algorithms to learn a task-specific distance functions. The Large Margin Nearest Neighbor (with Rejection) - LMNN (-R) [12] was used for viewpoint invariant object recognition. In [13], a transferred metric learning framework was proposed to learn a specific metric for every probe-gallery settings. A relaxed distance metric learning based on statistical inference was used to address person re-identification in [14]. In [15], a metric specially designed for re-identification was learned under pairwise constraints in high dimensional space.

Single-shot methods recognize a single instance at each time. However, the rich information contained in the image set cannot be fully exploited. Usually, users prefer to use as many images as possible, because they think that more images means more information and better recognition performance.

Multiple-shot methods are designed to make better use of the image sets. In [3], ten consecutive frames were used to generate the spatiotemporal graph. Region grouping over a certain time window was applied to achieve a stable foreground-background separation. The

SDALF method [1] adopted unsupervised Gaussian clustering to select key frames from a sequence of consecutive images. These key frames were used to build the signatures for each object. A similar proposal [2] combined color histograms with epitome - highly informative patches from a set of images to enhance appearance representation. Pictorial Structure was customized in [4] to finely localize the human body parts when multiple images are available for each person. Finally, a signature was built by concatenating features from different parts. However, this approach is only applicable when the pose estimator performs accurately. In [24], human blobs were extracted from accumulated images of tracking results. Then, sets of human blobs were used by boosting strategy to create a reliable visual signature. A set-based discriminative ranking model was proposed in [25], which simultaneously optimizes the feature space projection and set-to-set distance finding, obtaining a discriminative set-distance-based model.

Learning methods concentrate on distance metrics regardless of the image representation choice. Usually, those methods use very simple features such as texture and color histograms to perform re-identification. On the other hand, feature oriented methods focus on feature design without taking into consideration discriminative analysis. In fact, learning using sophisticated features is very hard or even inapproachable due to the curse of dimensionality.

We propose to combine the advantages of invariant feature design and discriminative feature learning. Besides a novel distinctive feature to capture the appearance information of an object, we also design an efficient learning method to distill the most discriminative information from the feature. The extracted features for image sets naturally benefit the multiple-shot based re-identification task because of the implicit discriminative measurement inside.

3. The Proposed Method

In this section, we detail our method for person representation and re-identification. This method directly handles the input images corresponding to object detection and tracking results. We assume that the bounding boxes of the objects have already been extracted. First, we extract histogram and covariance matrix features to describe the object. Then, we use third-party image sets and support vector machine (SVM) to learn the most discriminative feature representations. Finally, matching result is given by taking the minimum Euclidean distance of the average value of the features. See **Fig. 1** for illustration of our method.

3.1 Feature Extraction

Many possible cues could be used for a fine visual characterization. We focus on two kinds of information based on the previous researches in human appearance modeling [1, 6], that is, chromatic (histograms) information and statistical properties of image regions. In real word scenarios, human images tracked by far-field surveillance cameras are of moderate resolution. To ensure robustness in matching, each image is densely segmented into a grid of local patches.

Dense Color Histogram. Previous studies [1, 4, 6, 21] demonstrated that, color is the most powerful cue for person re-identification. Especially, studies carried out by Gray and Tao [21] showed that over 75 percent of the classifier weight was devoted to chromatic information, with the highest weight given to hue and saturation. In our method, the chromatic information of each patch is encoded by color histograms. We evaluate different color spaces, namely, HSV, RGB and YCbCr. Among them, HSV has shown to be superior. Besides, HSV color

space allows an intuitive quantization against different illumination conditions and camera settings. Therefore, we use HSV color space to extract histograms of each channel separately for each patch. For the purpose of combination with other features, all the histograms are L_2 normalized.

Log-Covariance Matrices. In order to handle viewpoint and illumination changes, region covariance [20] is used as a complementary feature to color histogram. Similar to the extraction of dense color histograms, covariance matrices are also extracted on a dense grid of local patches.

Covariance matrix $\mathbf{C}$ is a $d \times d$ symmetric matrix computed as:

$$\mathbf{C} = \frac{1}{N_p - 1} \sum_{i=1}^{N_p} (\boldsymbol{f}_i - \boldsymbol{\mu})(\boldsymbol{f}_i - \boldsymbol{\mu})^T, \tag{1}$$

where N_p is the number of pixels in the region. $\{\boldsymbol{f}_i\}_{i=1,2,\cdots,N_p}$ is the d dimensional feature vector for each pixel. $\boldsymbol{\mu} = \frac{1}{N_p} \sum_{i=1}^{N_p} \boldsymbol{f}_i$ is the mean feature vector. Covariance matrix presents a natural way of combining multiple heterogeneous features together with a relatively low dimensionality. The dimension of the covariance matrix is only related to the dimension of the feature vectors. Due to its symmetry, $\mathbf{C}$ has only $(d^2 + d)/2$ independent numbers. We follow the formulation of [5], which defines the feature vector of each pixel as:

$$\boldsymbol{f}_i = [x, y, R, G, B, \nabla^R, \theta^R, \nabla^G, \theta^G, \nabla^B, \theta^B] \tag{2}$$

where x and y are pixel locations, R, G, B are RGB channel values, ∇ and θ correspond to gradient magnitude and orientation in each channel respectively.

Covariance matrices are symmetric and non-negative definite. In our case, it is usually a symmetric positive definite (SPD) matrix. The set of all covariance matrices of a given size does not form a vector space because it is not closed under multiplication with negative scalars. In order to combine with other features and utilize the standard vector space learning algorithms, a key idea is to map the covariance matrices to the vector space by using the matrix logarithm proposed in the Log-Euclidean framework [26]. The matrix logarithm is computed as follows. Given a $d \times d$ covariance matrix $\mathbf{C}$, its eigen-decomposition is $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$, where the columns of $\mathbf{V}$ are orthonormal eigenvectors and $\mathbf{D} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ is the diagonal matrix of (non-negative) eigenvalues. The matrix logarithm is defined as:

$$\log(\mathbf{C}) = \mathbf{V} \cdot \mathrm{diag}\left(\log(\lambda_1), \log(\lambda_2), \cdots, \log(\lambda_d)\right) \cdot \mathbf{V}^T. \tag{3}$$

The eigenvalues of $\mathbf{C}$ are real and positive. The log mapping makes the eigenvalues of $\log(\mathbf{C})$ be real, which can be positive, negative or zero. Thus, $\log(\mathbf{C})$ is a symmetric matrix of the vector space, which can be handled with operations in Euclidean space. Compared with the Affine-Riemannian framework used in [5], the Log-Euclidean framework does not involve intensive computations of matrix square root or matrix inverse.

Because $\log(\mathbf{C})$ is symmetric, half-vectorization is carried out, denoted by $v\log(\mathbf{C})$. The upper triangular part of $\log(\mathbf{C})$ are packed into a vector by their column order. The obtained Log Covariance Matrix (LCM) feature is represented as:

$$v\log(\mathbf{C}) = [v\log\mathbf{C}_{1?}v\log\mathbf{C}_{?}\cdots v\log\mathbf{C}_{d\ d+1)/2?}]. \tag{4}$$

The LCM feature is L_2 normalized and combined with HSV histogram to get the feature vector for each patch. The final descriptor for each image is the concatenation of the feature vectors for all the patches, denoted as LCMH.

3.2 Max-margin Feature Learning Using Third-party Image Sets

After feature extraction, every image is represented by a feature vector $\mathbf{x} \in R^D$. An image set with n images belonging to the same object can be denoted as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, ?\mathbf{x}_n\} \in R^{D\times n}$. A probe is denoted as $\boldsymbol{X}^p$ and a gallery is denoted as $\boldsymbol{X}^g$. Instead of directly assessing similarity using LCMH descriptor, we propose to learn the most discriminative information of the probe/gallery using a third-party image set $\boldsymbol{X}^{tp}$. To obtain discriminative representation, each image set in both the probe sets $\mathbb{X}^P = \left\{\boldsymbol{X}_1^p, ?\boldsymbol{X}_2^p \cdots, ?\boldsymbol{X}_{n_p}^p\right\}$ and the gallery sets $\mathbb{X}^G = \left\{\boldsymbol{X}_1^g, ?\boldsymbol{X}_2^g \cdots, ?\boldsymbol{X}_{n_g}^g\right\}$ are described by measuring the difference with respect to the images contained in the third-party sets $?\mathbb{X}^{TP} = \left\{\boldsymbol{X}_1^{tp}\ ?\boldsymbol{X}_2^{tp} \cdots, ?\boldsymbol{X}_{n_{tp}}^{tp}\right\}$, where n_p, n_g and n_{tp} are the size of the probe, gallery and third-party image sets respectively, and the superscripts indicate the category of these sets. To learn the difference of the objects' feature more efficiently, the identities of the objects from the third-party image sets are totally different from those in both the gallery and probe image sets. The size of each image set can be different.

Our goal is to learn a discriminative feature representation for a given image set. Intuitively, a unique decision boundary that best separates an input set and a third-party set is appropriate for discriminative features of the input. The process is shown in **Fig. 1**, where we try to find the decision boundary between each image set in $\mathbb{X}^P$ and $\mathbb{X}^G$, and all the third-party image sets in $\mathbb{X}^{TP}$. The linear decision boundary is:

$$\boldsymbol{w}^p{}_{(\alpha,\gamma)}\bar{\mathbf{x}} + b = 0 \tag{5}$$

where $\bar{\mathbf{x}}$ is a vector in the LCMH feature space spaned by $\boldsymbol{X}_\alpha^p$ and $\boldsymbol{X}_\gamma^{tp}$, and $\boldsymbol{w}^p{}_{(\alpha,\gamma)}$ is the normal vector of the classifier boundary, which can be visualized as salient feature maps for a discriminative representation (**Fig. 2**).

In principle, any classifier could be used to form the classifier boundary. Due to its effectiveness and efficiency, we consider linear SVM that maximizes the margin between each pair of image set. The decision boundary between a probe $\boldsymbol{X}_\alpha^p$ and a third-party $\boldsymbol{X}_\gamma^{tp}$ is learned by using $\boldsymbol{X}_\alpha^p$ as positives and $\boldsymbol{X}_\gamma^{tp}$ as negatives. Learning the normal vector

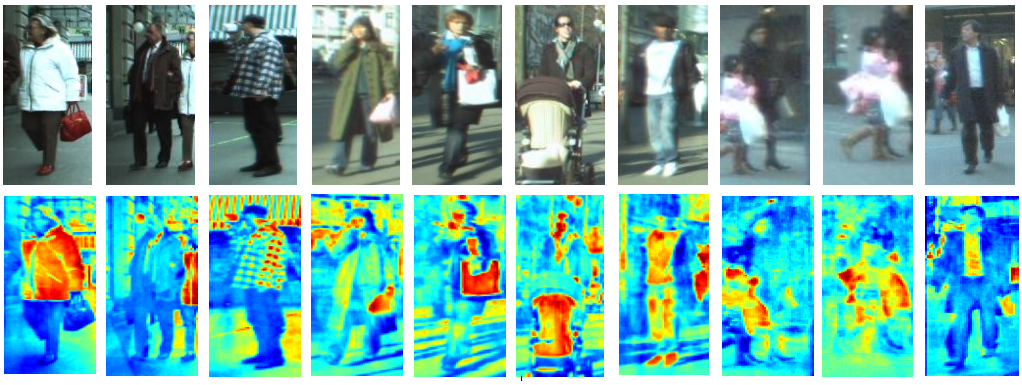


Fig. 2. Original images (top) and corresponding images generated from normal vectors of the decision boundaries (bottom). In the bottom row, red color indicates that large weights are given to the corresponding region. So these regions are more salient and discriminative for person re-identification.

$\boldsymbol{w}_{(\alpha,\gamma)}^{p} \in R^{D\times 1}$ of the decision boundary amounts to minimizing the following convex objective function:

$$L\left(\boldsymbol{w}_{(\alpha,\gamma)}^{p}\right)=\sum_{\mathbf{x}_i\in X_{\alpha}^{p}} h\left(\left(\boldsymbol{w}_{(\alpha,\gamma)}^{p}\right)^{T}\mathbf{x}_i\right)+\sum_{\mathbf{x}_j\in X_{\gamma}^{tp}} h\left(\left(-\boldsymbol{w}_{(\alpha,\gamma)}^{p}\right)^{T}\mathbf{x}_j\right)+\lambda\left\|\boldsymbol{w}_{(\alpha,\gamma)}^{p}\right\|^{2}, \tag{6}$$

where $h(x)=\max(0,1-x$ is the standard hinge loss function which is widely adopted by researchers, λ is the regularization parameter. The libSVM software [27] with the default settings of linear kernel is employed to solve the optimization problem. Once the decision boundary, which is represented by its normal vector $\boldsymbol{w}_{(\alpha,\gamma)}^{p}$, is determined, we can use it as an informative representation of the probe set $\boldsymbol{X}_{\alpha}^{p}$ against the third-party $\boldsymbol{X}_{\gamma}^{tp}$. As the size of the third-party sets is n_{tp}, we could get n_{tp} normal vectors $\left\{\boldsymbol{w}_{(\alpha,1)}^{p},\ \boldsymbol{w}_{(\alpha,2)}^{p},\cdots,\ \boldsymbol{w}_{(\alpha,n_{tp})}^{p}\right\}$ for a given probe $\boldsymbol{X}_{\alpha}^{p}$. Changing the positives to $\boldsymbol{X}_{\beta}^{g}$ and performs the same operation, we can obtain n_{tp} normal vectors $\left\{\boldsymbol{w}_{(\beta,1)}^{g},\ \boldsymbol{w}_{(\beta,2)}^{g},\cdots,\ \boldsymbol{w}_{(\beta,n_{tp})}^{g}\right\}$ for a gallery $\boldsymbol{X}_{\beta}^{g}$.

Essentially, the process of SVM learning is to find a decision boundary in D-dimensional space that best separates two image sets. SVM supervises this process to select a weight for each of the D-dimensions. Since our input images are the detecting or tracking results of the objects, regions in the same position roughly correspond to certain areas of the object. Besides, LCMH features are extracted on dense grids in the same manner. Therefore, different weights of feature dimensions from the normal vector of the decision boundary can be viewed as the saliency of the corresponding region (**Fig. 2**). Intuitively, images of the same object would be more likely to have similar salient feature maps than those of different objects. For example, an object only with salient upper body and an object only with salient lower body must have different identities. Human eyes can recognize objects' identities based on salient regions. Therefore, the normal vector of the decision boundary that best separates an input set in X^{P} and X^{G}, and a third-party set could be reasonably used as the discriminative feature for the input.

Fig. 2 shows the salient feature maps estimated by our third-party SVM feature learning method. The first row shows the original images. For each figure in the second row, pixels take the corresponding values of the normal vector with the same dimension as the original image. Raw pixel values are treated as image features for simplicity. Red indicates large weights of the region. We found that, our third-party feature learning gives higher weights to regions that are more salient and discriminative for each object.

3.3 Decision Making

As already mentioned in the previous section, given different third-party image sets, we could achieve a set of normal vectors for a probe/gallery. These normal vectors contain different aspects of the most discriminative features for the specific image set. When we compare the probe/gallery with more different third-party sets, we could draw more conclusions for the object. With human vision experience, we usually combine these conclusions to form a unique description for this object. In our work, we summarize the normal vectors and take the average value of them in order to blend the appearance information from different third-party sets. For a probe $\boldsymbol{X}_{\alpha}^{p}$ and gallery $\boldsymbol{X}_{\beta}^{g}$, the final feature vectors $\boldsymbol{w}_{\alpha}^{p}$ and $\boldsymbol{w}_{\beta}^{g}$ for them are obtained using the following equations:

$$\boldsymbol{w}_{\alpha}^{p} = \frac{1}{n_{tp}} \sum_{i=1}^{n_{tq}} \boldsymbol{w}_{(\alpha,i)}^{p}, \tag{7}$$

and

$$\boldsymbol{w}_{\beta}^{g} = \frac{1}{n_{tp}} \sum_{i=1}^{n_{tq}} \boldsymbol{w}_{(\beta,i)}^{g}. \tag{8}$$

Any similarity metric can be used to measure the distance between $\boldsymbol{X}_{\alpha}^{p}$ and $\boldsymbol{X}_{\beta}^{g}$ using the feature vectors $\boldsymbol{w}_{\alpha}^{p}$ and $\boldsymbol{w}_{\beta}^{g}$. For simplicity, Euclidean distance is adopted. Final recognition objective function can be written as:

$$I_{\alpha} = \arg\min_{i} \left\| \boldsymbol{w}_{\alpha}^{p} - \boldsymbol{w}_{i}^{g} \right\|_{2}, \tag{9}$$

where I_{α} is the identity of the probe $\boldsymbol{X}_{\alpha}^{p}$.

4. Experimental Results and Analysis

The evaluation of our method is carried out on ETHZ [30], CAVIAR4REID [4], and a modified version of iLIDS datasets [28]. These datasets reflect different challenges in real-world person re-identification, such as pose, viewpoint and illumination variations, low resolution, and occlusions (see **Fig. 3**). Identification performances are evaluated in terms of recognition rate by the Cumulative Match Curve (CMC) [17], and the normalized Area Under Curve (nAUC) for the CMC. CMC represents the expectation of finding the correct match in the top r matches. It expresses the performance of identification systems that return ranked list of candidates. The higher the CMC curve, the better the performance. On the other hand, nAUC gives an overall score of an identification system. Larger value of nAUC means better

Fig. 3. Sample images taken from ETHZ, iLIDS, and CAVIAR4REID datasets respectively with five pairs for each. Two images in the same column belong to the same person captured from different camera views.

performance. The proposed approach is denoted as SVM-LCMH. Re-identification by taking the minimum Euclidean distance of LCMH is denoted as LCMH. Comparisons with the state-of-the-art feature based methods and other feature selection methods are provided.

Human images are scaled into a fixed size of 48×128 pixels for ETHZ and iLIDS datasets and 32×64 pixels for CAVIAR4REID dataset respectively. Local patches of 16×16 pixels are extracted on a fixed step of 8 pixels. 8 bin histograms are estimated in each channel of the HSV color space. Before computing covariance feature, color equalization is applied to the V channel of the images.

4.1 ETHZ Dataset

This dataset has been originally used for pedestrian detection [30]. It has been used in [23] to test the PLS method for pedestrian recognition. Since the video sequences are captured from moving cameras, images have certain variations in resolution and illumination, some may suffer from severe occlusions (see **Fig. 3**). The dataset is structured as follows: SEQ. #1 contains 4,857 images for 83 persons; SEQ. #2 contains 1,936 images for 35 persons; SEQ. #3 contains 1,762 images for 28 persons.

Following the same multiple shot experimental protocol in [1], we randomly select a subset of N=5 images for each object to build the gallery set and probe set. When we match objects from any one of the three sequences, image sets in the other two sequences are treated as the third-party. For example, when the probe/gallery comes from SEQ. #1, images contained in SEQ. #2 and SEQ. #3 are taken as the third-party sets, and vice versa. All experiments are repeated 10 times to obtain reliable statistics. Comparisons are made with PLS, HPE and SDALF. **Fig. 4** and **Table 1** show CMC curves and nAUC of different methods on the three sequences. SVM-LCMH obtains the best results on SEQ. #1 and SEQ. #2. In particular, on SEQ. #1, the mean recognition rate of SVM-LCMH is 3% higher than SDALF for ranks between 2 and 7. The rank 1 recognition rate is 90% for SVM-LCMH, versus 85% for HPE. On SEQ. #2, SVM-LCMH significantly outperforms the other three methods. The rank 1 matching rate is around 97% for SVM-LCMH, versus 80% for HPE and 91% for SDALF respectively. Especially, for ranks greater than 3, recognition rate is 100%. On SEQ. #3, SVM-LCMH performs comparatively with SDALF as their nAUC are 99.23 and 99.07 respectively.

The performance of LCMH was also evaluated. Results are shown in **Fig. 4**. We can see that, on all the three sequences, SVM-LCMH gets superior recognition rate over LCMH. This

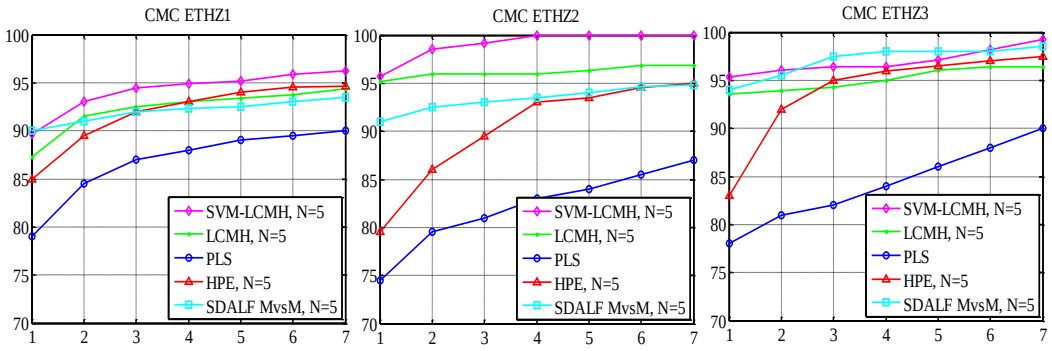


Fig. 4. CMC curves obtained on ETHZ dataset in multiple-shot case. Our methods are denoted as SVM-LCMH and LCMH. Comparisons are made with PLS [23], HPE [2], and SDALF [1].

Table 1. The nAUC of different methods on the three sequences of ETHZ dataset.

Sequence	PLS	HPE	SDALF	LCMH	SVM-LCMH
SEQ. #1	86.71	91.8	92.04	97.45	**98.22**
SEQ. #2	82.07	90.14	93.34	98.22	**99.81**
SEQ. #3	84.14	93.85	99.07	97.1	**99.23**

observation verifies the effectiveness of our adoption of SVM and third-party image sets for discriminative feature learning. It is also interesting to notice that, LCMH significantly outperforms PLS. At the same time, it performs on par with HPE and SDALF in multiple-shotcases. These results support our claim that, LCMH descriptor is robust to illumination changes, pose variation and occlusion. **Table 1** details the identification rate of different methods. Our SVM-LCMH gets the highest values on all the three sequences.

4.2 CAVIAR4REID Dataset

CAVIAR4REID [4] has been extracted from the CAVIAR dataset, which consists of several sequences filmed in the entrance lobby of the INRIA Labs and in a shopping center in Lisbon. Image sizes vary from 17×39 to 72×144 pixels. 50 different objects are captured by two camera views with 10 images each view. The other 20 objects are captured by only one camera, each with 10 images. These images vary a lot in terms of pose, illumination and viewpoint. These variations make the person re-identification issue more difficult. Sample images are shown in **Fig. 3**.

For this dataset, we randomly choose N=5 images for each person from the 50 objects with two camera views to form our probe and gallery. Image sets from the remaining 20 objects form our third-party sets. So the third-party number is $n_{tp} = 20$. **Fig. 5** (left) reports the CMC curves of SVM-LCMH and three benchmark methods: SDALF, HPE and CHISD. For CHISD, we use the same LCMH feature as our SVM-LCMH method. The overall identification rate on this dataset is not very high, indicating that, person re-identification in real world scenario remains to be a very challenging and open problem. Overall, SVM-LCMH outperforms CHISD. The matching rate of SVM-LCMH, AHPE and SDALF are comparative, as the

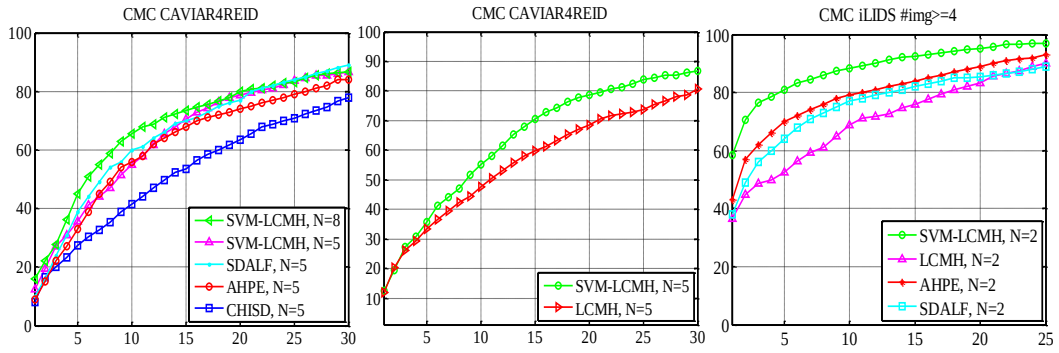


Fig. 5. CMC curves on CAVIAR4REID and $\text{iLIDS}_{\geq 4}$ datasets. Comparisons are made with SDALF [1], AHPE [29], and CHISD [7].

nAUC are 75.94, 76.24, and 72.92 respectively. SVM-LCMH achieves the highest matching rate at rank 1. The matching rate is around 14% for SVM-LCMH, versus 9% for SDALF and 10% for AHPE. For SVM-LCMH, when we set the number of images to N=8, we get consistent improvement in recognition rate, the nAUC increases to 77.81. The intuition is that, more images means more information can be used, which is critical to reduce ambiguity.

Fig. 5 (middle) gives the comparison between SVM-LCMH and LCMH on CAVIAR4REID. As expected, the performance of SVM-LCMH on this dataset is greatly improved compared with the performance of LCMH. In particular, SVM-LCMH is 10% better than LCMH for ranks greater than 15. On the other hand, the nAUC increases from 70.42 for LCMH to 75.94 for SVM-LCMH. The results again bear out the advantage of our SVM and third-party feature learning.

4.3 iLIDS Dataset for Re-identification

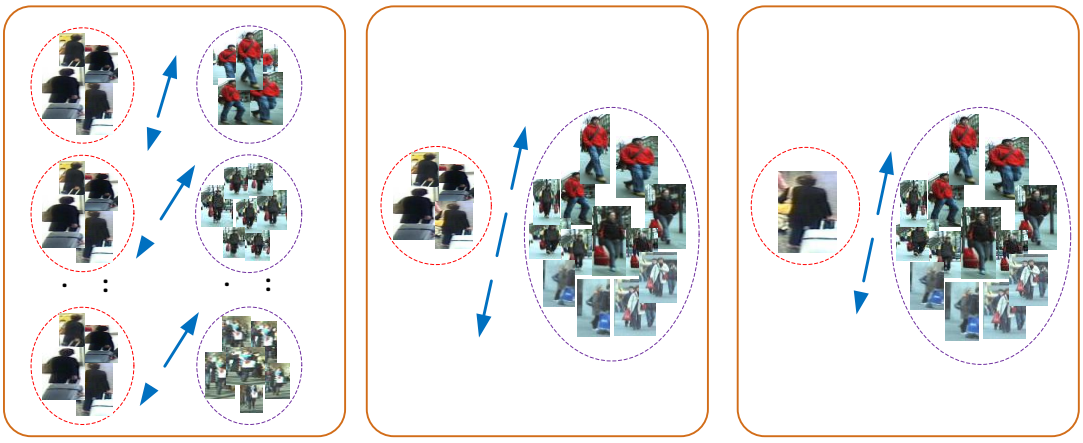
Images of iLIDS dataset for re-identification [28] are captured indoor at a busy airport arrival hall. There are 479 images of 119 objects in total. As images are taken from non-overlapping camera views, illumination changes and occlusions are quite large for this dataset (see **Fig. 3**). The number of images for each person is very low (4 in average). It does not fit very well with our multiple-shot method. So we use the modified version of the dataset [29]. Objects with more than 4 images are chosen, named $\text{iLIDS}_{\geq 4}$.

As most of the objects contain only four images, we randomly select 2 images for each person to build the probe sets, while another two images form the gallery sets. Image sets contained in the three sequences of ETHZ are taken to be the third-party image sets. In total, the third-party number is $n_{tp} = 246$. The matching between a probe and a gallery is estimated. Since SDALF and AHPE are the only methods published their results on this dataset, the comparisons are made between our methods and them. Experimental results are shown in **Fig. 5** (right).

This figure shows clearly that SVM-LCMH yields the best rank 1 identification rate and overall much better performance to the compared methods. The rank 1 matching rate is 58% for SVM-LCMH versus 43% for AHPE and 38% for SDALF. We also report the performances of LCMH. It is obvious that SVM-LCMH is greatly superior to LCMH. This witnesses again the fact that person re-identification benefits from SVM and third-party feature learning. Examples of matching results using the proposed method on iLIDS and CAVIAR4REID datasets are shown in **Fig. 6** and **Fig. 7** respectively.

Fig. 6. Results of person re-identification on iLIDS dataset using SVM-LCMH. In each row, the left-most image is the probe, the other images are the top 20 matched gallery. The true match is highlighted with a red box.

Fig. 7. Results of person re-identification on CAVIAR4REID dataset using SVM-LCMH. In each row, the left-most image is the probe, the other images are the top 20 matched gallery. The true match is highlighted with a red box.



SVM-LCMH SVM2-LCMH SVM3-LCMH

Fig. 8. Different SVM and third-party feature learning schemes. SVM-LCMH: For a probe/gallery image set (red dotted), a set of boundaries are achieved against different third-party sets (purple dotted). SVM2-LCMH: Given a probe/gallery image set, a unique boundary is learned against a third-party set, which is the combination of all the images of the reference person. SVM3-LCMH: All the images of the reference person compose a third-party set. A boundary is learned between a given image of the probe/gallery and the third-party set.

4.4 Discussions

SVM and third-party feature learning schemes. In our SVM-LCMH method, discriminative features are learned by taking images from a probe $\boldsymbol{X}^p$ or a gallery $\boldsymbol{X}^g$ as positives, while images from a third-party set $\boldsymbol{X}^{tp}$ as negatives. Experiments have been done on iLIDS$_{\geq 4}$ dataset to demonstrate the advantage of this scheme. Here, we consider two other different schemes: a) all images in the third-party sets $\mathbb{X}^{TP}$ compose the negatives. The other formulations are the same as SVM-LCMH. This scheme is denoted as SVM2-LCMH. In this setting, we get only one normal vector for a probe $\boldsymbol{X}^p$ or gallery $\boldsymbol{X}^g$. This normal vector is used to assess similarity between two objects. b) We use SVM in a one-against-all manner. In particular, every image in the probe/gallery is treated to be positive, all images in the third-party sets $\mathbb{X}^{TP}$ compose the negatives. This scheme is denoted as SVM3-LCMH. The learned normal vector serves as the feature for each image. Minimum pairwise distance determines the similarity between the probe and gallery. **Fig. 8** illustrates the three different feature learning schemes.

Identification rates of different schemes are given in **Fig. 9**. The overall recognition performance was much better for scheme SVM-LCMH when compared with SVM2-LCMH and SVM3-LCMH. For example, rank 1 recognition results are 20% and 8% for SVM2-LCMH and SVM3-LCMH, versus 58% for SVM-LCMH. This can be attributed to the fact that multiple image sets cover more different aspects of within-class variations of the input. Discriminative feature learning using scheme SVM-LCMH is better at exploiting these variations.

Selection of third-party image sets. In the proposed SVM-LCMH feature learning method, third-party image sets are selected for discriminative feature learning. For ETHZ and CAVIAR4REID datasets, the third-party images are taken under similar imaging conditions as the probe/gallery images. For iLIDS$_{\geq 4}$ dataset, the third-party images are taken from the ETHZ dataset with different imaging conditions. Good third-party image sets are expected to

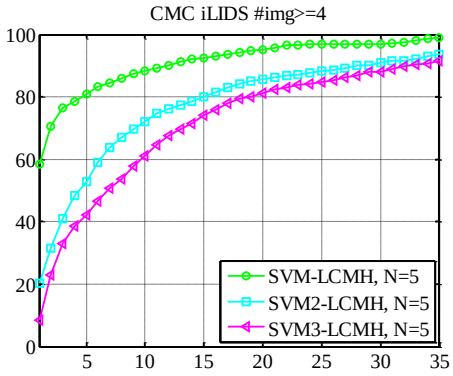


Fig. 9. CMC curves on iLIDS dataset using different SVM and third-party feature learning schemes illustrated in **Fig. 8**.

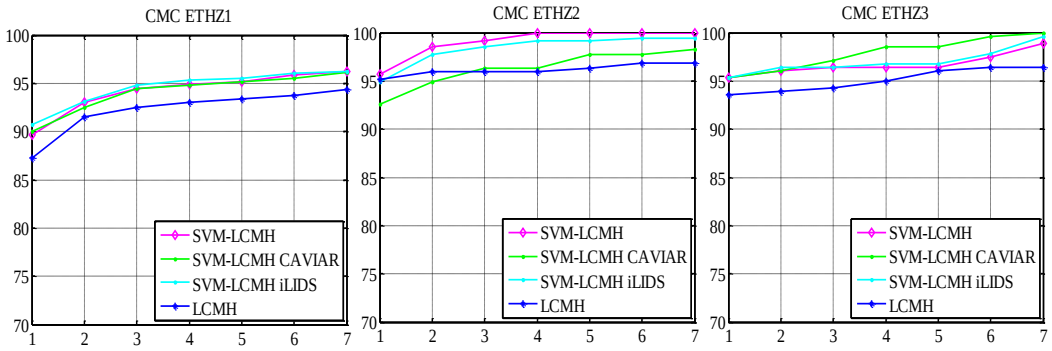


Fig. 10. CMC curves on ETHZ dataset using different third-party image sets.

have certain amount of persons with enough images for each of them covering different kinds of variations of the probe and gallery. This is to make it possible that for any given probe or gallery object, there is a sufficient representation of it by a few samples of the third-party image sets with similar visual aspects. Experiments have been conducted to test the performance of SVM-LCMH with different selection of third-party image sets. Dataset name is added to "SVM-LCMH" to denote different third-party sets, e.g., "SVM-LCMH CAVIAR" denotes third-party image sets taken from CAVIAR4REID dataset. "SVM-LCMH" and "LCMH" have the same meaning as stated before. Experimental results on ETHZ dataset are given in **Fig. 10**. From the figure, we can draw two conclusions: first, LCMH with SVM and third party feature learning outperforms LCMH; second, our SVM-LCMH has some robustness to the quality of different third-party image sets. The second conclusion implies that existing re-identification datasets can be employed to serve as the common third-party reference.

Run time analysis. For our SVM-LCMH method, the evaluation time are dominated by LCM feature computation and max-margin feature learning from the third-party image sets. For each image, computing the LCM feature takes <0.5 seconds using integral images. For a probe with 5 images from SEQ. #1, the third-party number is 70. Learning the feature using libSVM software takes <0.1 seconds using MATLAB implementation. All timings are run on a 2.67GHz Intel CPU and 4 GB RAM. So, our method is suitable for real-time surveillance applications.

5. Conclusion

In this paper, we proposed a descriptive feature for characterizing the appearance of an object and a discriminative feature learning method with the adoption of third-party image sets for person re-identification. The feature utilizes the power of region covariance matrix while at the same time transforming it from the Riemannian manifold to vector space. The feature learning method is conceptually simple and computationally efficient for multiple-shot person re-identification. Discriminative information is explored from the decision boundary obtained by maximizing the margin between a probe/gallery set and a third-party image set. Experimental results on three challenging datasets show that, with either similar or dissimilar image sets as third-party, the proposed approach gets promising performance on person re-identification. Future work includes finding a representative dataset to serve as the common third-party reference for person re-identification.

References

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino and M. Cristani, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130-144, 2013. Article (CrossRef Link).

[2] L. Bazzani, M. Cristani, A. Perina, M. Farenzena and V. Murino, "Multiple-shot person re-identification by hpe signature," in *Proc. of International Conference on Pattern Recognition (ICPR)*, pp. 1413-1416, 2010. Article (CrossRef Link).

[3] N. Gheissari, T. Sebastian and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1528-1535, 2006. Article (CrossRef Link).

[4] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani and V. Murino, "Custom pictorial structures for re-identification," in *Proc. of British Machine Vision Conference (BMVC)*, 2011. Article (CrossRef Link).

[5] S. Bak, E. Corvee, F. Bremond and M. Thonnat, "Multiple-shot human re-identification by Mean Riemannian Covariance Grid," in *Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pp. 179-184, 2011. Article (CrossRef Link).

[6] I. Kviatkovsky, A. Adam, Amit and E. Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 7, pp. 1622-1634, 2013. Article (CrossRef Link).

[7] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2567–2573, 2010. Article (CrossRef Link).

[8] K. Guo, P. Ishwar and J. Konrad, "Action Recognition from Video using Feature Covariance Matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479-2494, 2013. Article (CrossRef Link) .

[9] V. Vapnik, "The nature of statistical learning theory," *Springer*, 2000. Article (CrossRef Link).

[10] X. Li, X. Zhao, Y. Fu and Y. Liu, "Bimodal gender recognition from face and fingerprint," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2590–2597, 2010. Article (CrossRef Link).

[11] M. Ma, M. Shao, X. Zhao and Y. Fu, "Prototype based feature learning for face image set classification," in *Proc. of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1-6, 2013. Article (CrossRef Link).

[12] M. Dikmen, E. Akbas, T. Huang and N, Ahuja, "Pedestrian recognition with a learned metric," in *Proc. of Asian Conference on Computer Vision (ACCV)*, pp. 501-512, 2010. Article (CrossRef Link).

[13] W. Li, R. Zhao and X. Wang, "Human reidentification with transferred metric learning," in *Proc. of Asian Conference on Computer Vision (ACCV)*, pp. 31-44, 2012. Article (CrossRef Link).

[14] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2288-2295, 2012. Article (CrossRef Link).
[15] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2666-2672, 2012. Article (CrossRef Link).
[16] S. Bak, E. Corvee, F. Bremond and M. Thonnat, "Person Re-identification Using Spatial Covariance Regions of Human Body Parts," in *Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pp. 435-440, 2010. Article (CrossRef Link).
[17] X. Wang, G. Doretto, T. Sebastian, J. Rittscher and P. Tu, "Shape and appearance context modeling," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 1-8, 2007. Article (CrossRef Link).
[18] B. Ma, Y. Su and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proc. of European Conference on Computer Vision Workshops and Demonstrations*, pp. 413-422, 2012. Article (CrossRef Link).
[19] B. Ma, Y. Su and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *Proc. of British Machine Vision Conference (BMVC)*, 2012. Article (CrossRef Link).
[20] O. Tuzel, F. Porikli and P. Meer, "Region covariance: a fast descriptor for detection and classification," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 589-600, 2006. Article (CrossRef Link).
[21] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 262-275, 2008. Article (CrossRef Link).
[22] B. Prosser, W. Zheng, S. Gong and T. Xiang, "Person re-identification by support vector ranking," in *Proc. of British Machine Vision Conference (BMVC)*, pp. 1-11, 2010. Article (CrossRef Link).
[23] W. Schwartz and L. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. of Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pp. 322-329, 2009. Article (CrossRef Link).
[24] S. Bak, E. Corvee, F. Bremond and M. Thonnat, "Person re-identification using Haar-based and DCD-based signature," in *Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pp.1-8, 2010. Article (CrossRef Link).
[25] Y. Wu, M. Minoh, M. Mukunoki and S. Lao, "Set based discriminative ranking for recognition," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 497-510, 2012. Article (CrossRef Link).
[26] V. Arsigny, P. Pennec and X. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp.411–421, 2006. Article (CrossRef Link).
[27] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no.3, 2011. Article (CrossRef Link).
[28] W. Zheng, S. Gong and T. Xiang, "Person Re-identification by Probabilistic Relative Distance Comparison," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 649-656, 2011. Article (CrossRef Link).
[29] L. Bazzani, M. Cristani, A. Perina and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, vol.33, no.7, pp. 898-903, 2012. Article (CrossRef Link).
[30] A. Ess, B. Leibe and L. Van Gool, "Depth and appearance for mobile scene analysis," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 1-8, 2007. Article (CrossRef Link).

Yanna Zhao received the MS degree in signal and information process from Shandong Normal University, Jinan, China, in 2010. She is currently a PhD candidate at School of Information Science and Engineering, Shandong University, Jinan, China. She is now doing research in Institute of Image Processing and Pattern Recognition at Shanghai Jiao Tong University. Her research interests include computer vision, pattern recognition and image processing.

Lei Wang received the B.Sc. and M.Sc. degrees from Shandong University, Jinan, China, in 2002 and 2005, respectively. She is currently a PhD candidate at School of Information Science and Engineering, Shandong University, Jinan, China. Since 2005, she has been a Lecturer in the School of Electrical engineering and Automation, Qilu University of Technology, Jinan, China. She also takes part in the research of Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include computer vision, pattern recognition, and human activity analysis.

Xu Zhao received the Ph.D. degree in pattern recognition and intelligence system from Shanghai Jiao Tong University in 2011. He is currently an Associate Professor in the department of Automation at Shanghai Jiao Tong University. He was a visiting scholar at the Beckman Institute for Advanced Science and Technology at University of Illinois at Urbana-Champaign from 2007 to 2008. He had been the postdoctoral research fellow in the Northeastern University from 2012 to 2013. His research interests include visual analysis of human motion, machine learning and image/video processing.

Yuncai Liu received his PhD in the Department of Electrical and Computer Science Engineering from the University of Illinois at Urbana-Champaign in 1990 and worked as an associate researcher at the Beckman Institute of Science and Technology from 1990 to 1991. Since 1991, he was a system consultant and then chief consultant of research at Sumitomo Electric Industries, Ltd., Japan. In October 2000, he joined Shanghai Jiao Tong University as a distinguished professor. His research interests are in image processing and computer vision, especially in motion estimation, feature detection and matching, and image registration. He also has made great progress in the research of intelligent transportation systems.