Multimed Tools Appl (2016) 75:4795–4813
DOI 10.1007/s11042-015-2503-y

Person Re-identification by encoding free energy feature maps

Yanna Zhao · Xu Zhao · Ruotian Luo · Yuncai Liu

Received: 7 August 2014 / Revised: 22 December 2014 / Accepted: 8 February 2015 /
Published online: 20 March 2015

Abstract Recognizing objects from disjoint camera views, known as person re-identification, is an important and challenging problem in the field of computer vision. Recent progress in person re-identification is due to new visual features and models that deal with cross-view differences. Existing appearance models focus on visual features in the normal sense, e.g., color histogram, Scale-invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG). In this paper, we propose a new appearance based method using the generative information of local image features and their encoding. In this paradigm, local image features which capture the color and structural cues of the human images are first extracted. A Gaussian Mixture Model (GMM) is then learned to approximate the generation process of these features. It provides a relatively comprehensive statistical representation. Finally, discriminative feature maps are obtained by calculating Free Energy Score Space (FESS) for GMM. The obtained feature maps are concatenated and encoded into a fixed-length feature vector for person re-identification. Our approach demonstrates promising performance on challenging datasets. It is also very practical: it has low computational cost both at training and testing. A GMM trained on images with different imaging conditions can be applied to other images without any significant loss in performance.

Keywords Person re-identification · Appearance modeling · Gaussian mixture model · Free energy score space

Y. Zhao
School of Information Science and Engineering, Shandong University, Jinan, China

Y. Zhao (✉) · X. Zhao (✉) · R. Luo · Y. Liu
School of Automation, Shanghai Jiao Tong University, Shanghai, China
e-mail: yannazhao@outlook.com
e-mail: zhaoxu@sjtu.edu.cn

R. Luo
e-mail: skylikelrt@sjtu.edu.cn

Y. Liu
e-mail: whomliu@sjtu.edu.cn

1 Introduction

Person re-identification is the problem of matching objects observed from different camera views. It is increasingly important due to its potential wide application in video surveillance, object tracking over non-overlapping camera views, multi-camera behavior analysis and target search in a collection of video sequences. However, this problem is also very challenging. First, surveillance cameras are of moderate resolution and low frame rate, biometric cues such as face or gait may not be available or difficult to catch. The main cues that can be relied on are the appearance information. Second, the huge amount of surveillance cameras may observe thousands of different objects in a public area within a single day, and some of them may have similar appearance. Finally, the same object observed under different cameras often undergoes large variations in illumination, poses, viewpoints, image resolutions and backgrounds. These make the inter-personal variations more significant than intra-personal variations.

Existing appearance based approaches mainly count on local and global image features to capture the visual cues of the human appearance. While local features provide raw and basic cues within body parts, global features provide the overall configuration of the body parts. The classical features for person re-identification include: color [4, 8, 10, 16, 20, 21, 38] (widely adopted since the color of clothing captures simple but efficient visual cues), textures [4, 8, 9, 16, 21, 38], covariance features [1, 16, 21], edges [9], co-occurrence matrix [32], Histogram of Oriented Gradients (HOG) signatures [27] and interest points, e.g., Speeded Up Robust Features (SURF) and Scale-invariant Feature Transform (SIFT) [9, 35]. Once local features are extracted, various strategies can be employed to combine them in order to capture different visual information. For instance, authors in [10, 38] combined 8 color histograms with 21 texture features. Farenzena et al. [8] combined weighted color histograms with maximally stable color regions, achieving state-of-the-art performance under unsupervised settings.

In typical object recognition literatures, some researchers focus on the encoding of local features into a global image signature. Perhaps the most common approach is the bag-of-visual-words (BoV) model introduced in [29]. In particular, local features are quantized, and their overall distribution in an image is represented by means of BoV histograms. The BoV model was used for person re-identification in [37], where the BoV description of an object is enriched by the contextual information coming from the surrounding people. Recent advances replace the hard quantization of features involved in this method with alternative encodings that retain more information about the original image features. Among them, Fisher vector (FV) [25], which encodes higher order statistics of local features, gives excellent performance on several challenging object recognition and image retrieval tasks [23, 28, 30]. In [20], FV was applied to person re-identification and showed promising performance on benchmark datasets.

Fisher kernel [13] was proposed to integrate generative models and discriminative models in a hybrid scheme. The basic idea is to represent a set of data by gradient of its log likelihood with model parameters. The fixed-length representation is called FV. Generative models are designed to model data distribution. They seek to explain data in terms of hierarchical model with hidden variables. These hidden variables encode higher order information related to the observed data that could be informative in recognizing data samples [17, 18, 24, 25]. In these methods, explicit feature mappings or score spaces are extracted from the generative models of the data distribution, producing a fixed-length feature vector in a highly informative space. The resulting features are not visual cues in the normal sense (e.g., SIFT, HOG), but are abstract ones with components determined by the generative model structure.

Free energy score space (FESS) [24] also seeks to derive feature maps based on the log likelihood function of a model. But they focus on the random variables, rather than on the

parameters in their derivation. FESS provides a principled way to derive feature maps from generative models, the mapping of which arises as a consequence of the factorization of the generative model being used. The feature maps measure how well a sample fits the model and how uncertain the fit is, thus giving rise to meaningful new features for vision tasks, e.g., scene classification and gene recognition. In this paper, we explore the potential of FESS feature mapping for person re-identification. Specifically, we use the Gaussian Mixture Model (GMM) to approximate the distribution of human image data. Then, we calculate FESS for GMM. With GMM, the dimension of the FESS feature maps is linear in the number of mixture centers and is independent of local feature dimension. The obtained feature maps are encoded into a fixed-length feature vector for person re-identification. We show that FESS encoding improves performance of BoV in person re-identification. Further experiments on several widely used datasets demonstrate that FESS encoding also outperforms benchmark methods.

In summary, the major contributions of this work are as follows:

1. While existing person re-identification methods focus on classical image features, we propose to use generative information for person re-identification. To the best of our knowledge, there is little work in person re-identification using this kind of methods.
2. We derive the FESS feature mapping for GMM using variational inference. The resulting feature maps, which measure how well a sample fits a random variable and how uncertain the fit is, are encoded into a fixed-length feature vector for person re-identification. The size of the encoded feature vector is much smaller than FV with the same number of mixture centers.
3. Our approach can be divided into offline and online parts, where the majority of computation cost is accomplished by the offline part, which is especially suited to real-time surveillance scenarios. Besides, the training data used to get the GMM may have different imaging conditions (such as camera parameters, illuminations, background, etc.) as those in the query and gallery sets without any significant loss in performance.

This paper is an extended version of the work of [36]. Besides the obvious increase of paper length, the major extensions of this journal paper compared with the original conference paper include: (1) More experimental results are conducted and analyzed by introducing a new dataset and two encoding alternatives: BOV encoding of the same local feature as our method and FESS encoding of dense SIFT local descriptors; (2) The proposed method is divided into the offline and online parts, which significantly reduce the computational cost of re-identification. The time complexity, performance and applicability of our method are also evaluated and discussed; (3) More analysis and discussions are added to further clarify the motivation of the proposed method. The rest of the paper is organized as follows. In Section 2, related works on person re-identification are described. Section 3 describes the proposed method in detail. Experimental validations are given in Section 4. Finally, Section 5 concludes the paper.

2 Related works

A categorization of recent appearance based methods in person re-identification is given in Table 1: the supervised methods and the unsupervised methods. In the former category [2, 3, 6, 10, 11, 15, 19, 22, 26, 27, 33, 38], a dataset is split into training (with identity labels) and testing sets. The training set is used to analyze the features and/or the policies for combining them that ensures high re-identification accuracies. The testing set is used as validation. Unsupervised methods either extract features directly [1, 5, 8, 9, 21, 32] or learn discriminative

Table 1 Categorization of existing re-identification methods

	Supervised	Unsupervised
Single-shot	[6, 10, 11, 15, 19, 22, 26, 27, 38]	[1, 5, 8, 20, 21, 35] Our approach
Multiple-shot	[2, 3, 11, 33]	[5, 8, 9, 20, 21, 32] Our approach

and reliable descriptors through unsupervised learning [20, 35]. A complementary classification separates the single-shot methods from the multiple-shot methods. The former focuses on associating pairs of images for each object, while the latter employs multiple images of the same object as probe and gallery.

For supervised methods, discriminative feature selection and metric learning have been widely used to pick up the most discriminative and reliable subset of features and to reduce cross-view variations. Considering feature selection methods, boosting and SVM are widely employed. In [10], an intelligently designed feature space was defined to represent an object. Instead of picking up a specific feature by hand, boosting strategy was used to find the best representation for re-identification. The same strategy has also been applied in [2, 3, 11]. The appearance models in [2] were haar-like features and dominant color descriptors. In [3], a novel appearance model based on Mean Riemannian Covariance (MRC) patches was extracted. The discriminative power of MRC patches was obtained by a boosting scheme. A rich set of feature descriptors based on color, textures and edges was proposed in [27]. The weighting of different features for a specific individual was got by using Partial Least Squares (PLS) analysis in a one-against-all scheme. The above mentioned methods learn discriminative features and deal with drastic viewpoint changes in a supervised way. A recently proposed approach [35] was to learn the salience information in an unsupervised manner. Color and SIFT features were extracted from densely sampled mid-level local patches. Dense correspondence was built between image pairs using adjacency constrained patch matching before saliency learning. The acquired saliency feature was combined with other features for re-identification. Person re-identification was formulated as a relative ranking problem in [26]. A subspace was learnt by giving the potential true match the highest ranking.

Metric learning plays an important role in machine learning. It is also particularly important for computer vision tasks, e.g., person re-identification, action recognition. The large margin nearest neighbor metric learning framework has been adopted and extended in [6] to learn the most effective metric to match data from two arbitrary camera views. In [38], a Probabilistic Relative Distance Comparison (PRDC) model has been introduced, which aims at maximizing the probability that a pair of truth match has a smaller distance than a wrong match. Instead of using a single generic metric for matching all the subjects, a transfer learning framework was put forward in [19] which learned a specific metric for different pairs of query-gallery. Pairwise constrained component analysis was used to learn distance metrics in high dimensional input space [22]. To deal with the scalability and required degree of supervision problem, a distance metric was proposed to learn from equivalence constraints [15]. The above mentioned methods focus on single-image based identification. Recently, both face recognition and person re-identification have been extended to multiple image problems due to the widespread of surveillance cameras. In [33], a set-based discriminative model was proposed. It simultaneously optimizes the set-to-set distance finding and feature space projection.

In general, supervised methods produce higher performance than unsupervised methods. However, since supervised methods require labeling new training data when camera settings change, they are impractical for applications especially for large-scale camera networks.

Unsupervised methods have been developed to achieve fast match. In [32], appearance models were developed based on the concept of shape and appearance context. Discriminative features were extracted by modeling the spatial distributions of appearance relative to body parts. In [9], every human image was fitted with an articulated model in order to establish correspondence. However, these two methods are not flexible enough and only suitable for frontal view re-identification. SDALF method [8] partitioned the human body into symmetry and asymmetry parts to handle pose variation problems. Weighted color histograms, maximal stable color region descriptors, and recurrent highly structured patches are combined to describe each part of the human body. Several other approaches were proposed to deal with pose variations [1, 5]. In [5], pictorial structures were customized to finely localize each human body part. In [1], HOG was used to automatically detect the human body parts. Spatial Covariance Region (SCR) was extracted from each part to model the appearance of the human body. However, these models will fail when the pose estimators work inaccurately. The BiCov descriptor [21] combined biologically inspired features (e.g., Gabor filters) and covariance descriptors to handle both background and illumination changes.

Modeling visual cues through generative score spaces or feature mappings have been proven to be valuable approaches in the literature [12, 17, 18, 23–25, 28, 30, 34]. The underlying methodology is to derive score functions or measures from generative models. The derived score functions provide features of fixed length for classification/recognition. In [20], FV [25] was employed to encode higher order statistics of local features. FV was also utilized to other vision tasks [23, 25, 28, 30]. FESS [36] is another generative score space technique which encloses the uncertainty that exists in the generative learning phase usually disregarded by FV. The variational free energy terms are treated as feature vectors, so that the degree of fitness of the data and the uncertainty of the fitness are included explicitly in the feature maps. In this paper, we derive feature maps by computing FESS for GMM and study their utility for person re-identification. Our method lies in the class of unsupervised methods, and working both in the single-shot and multi-shot modalities.

3 The proposed method

In this section, we detail our method for person re-identification. We divide our method into offline and online parts. The offline part includes two steps: (1) extract local features which encompass visual and structural cues from the input images; (2) learn GMM to model the distribution of the local features. The online part is getting the FESS feature mapping for each image and encode the mapping to a fixed-length feature vector, and use the feature vector for person re-identification. Figure 1 illustrates the flowchart of our method.

3.1 Local feature extraction

Our FESS encoding starts by extracting local image features from an image, which provides raw and basic cues for describing an object. In order to capture effective visual information, we extract the local image features by representing each pixel in the image I by a feature vector in 7-dimensional feature space (7-d feature) [20]:

$$\mathrm{f}(x,y,I) = \big(x, y, I(x,y), I_x(x,y), I_y(x,y), I_{xx}(x,y), I_{yy}(x,y)\big). \tag{1}$$

This simple feature vector contains pixel coordinates, raw pixel intensity value in the corresponding color channel, and the first-order and second-order derivatives with respect to

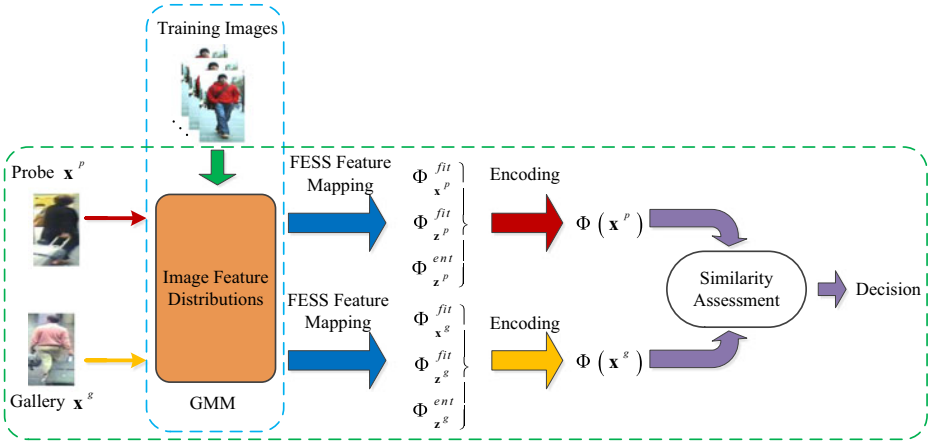

Fig. 1 An overview of the proposed FESS encoding method for person re-identification. The offline part is enclosed in *blue dotted box* and the online part is enclosed in *green dotted box*

pixel coordinates. Other image cues, e.g., gradient orientation or texture features can also be incorporated in this local feature vector. In our approach, this 7-d feature vector is adopted for its simplicity and effectiveness.

Features of different human body parts may have different distributions. In order to capture the spatial information, human images are usually divided into three parts roughly corresponding to head, shirt and pants [10]. In our approach, considering moderate resolution of human images captured by far-field surveillance cameras, we divide the input images generated from object detection or tracking results into 3×4 blocks. Our feature detection and description technique borrows the ideas developed for object recognition and further incorporates the spatial information.

Once the local features are extracted, we use them to construct a signature to characterize an object. For this step we calculate the FESS for a generative model. GMM, which has been widely used in modeling the distribution of image features, is adopted as the generative model to approximate the generating process of the above mentioned local features. The lower bound of the log likelihood (see Eq. (7)) function is expanded according to the random variable, and each resulting term becomes a feature map.

3.2 Gaussian mixture model

In this section, we give an exhaustive interpretation of GMM for ease of reading and for derivation of FESS feature mapping in the next section. Let $\mathbf{x} \in \mathscr{R}^D$ be the observed random variable. In our context of person re-identification, $\mathbf{x}$ denotes the local image features and $D=7$. Suppose $\mathbf{z}=\{z_1,\cdots,z_K\}$ is a set of hidden variables following the Multinomial distribution over K possible states. The random variable z_k indicates which one of the K Gaussians each $\mathbf{x}$ had come from, e.g., $z_k=1$ means that the k-th Gaussian distribution is selected to generate the sample $\mathbf{x}$. The prior distribution of $\mathbf{z}$ is typically selected to be:

$$\mathrm{P}(\mathbf{z}) = \prod_{k=1}^{K} a_k^{z_k}, \tag{2}$$

where $\boldsymbol{a}=(a_1,\cdots,a_K)^{\mathrm{T}}$ is the mixture prior satisfying $a_k = E_{P(\mathbf{z})}[z_k]$.

The conditional distribution of $\mathbf{x}$ given the hidden variable $\mathbf{z}$ is:

$$\mathrm{P}\left(\mathbf{x}\middle|\mathbf{z}\right) = \prod_{k=1}^{K} N(\mathbf{x};\boldsymbol{\mu}_k,\Sigma_k)^{z_k} = \prod_{k=1}^{K}\left[\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}\right]^{z_k}. \quad (3)$$

where $\boldsymbol{\mu}_k$ and Σ_k are the mean and covariance matrix of the k -th Gaussian distribution respectively. The diagonal covariance matrices are used when learning GMM. This is done mainly for two reasons. First, a weighted sum of Gaussians with diagonal covariance can approximate any distribution with an arbitrary precision. Second, the computational cost of estimating diagonal covariance is much lower than computing full covariance [25]. For later use, we introduce the notation of $\Sigma_k = \mathrm{diag}(\sigma_{k1}^2,\cdots,\sigma_{kD}^2)$, where σ_{kd} is the standard deviation of the k -th Gaussian in dimension d.

With the prior distribution P($\mathbf{z}$) and the conditional distribution P($\mathbf{x}$|$\mathbf{z}$), the joint distribution of GMM can be formulated as:

$$\begin{aligned}\mathrm{P}\left(\mathbf{x},\mathbf{z}\middle|\theta\right) &= P\left(\mathbf{x}\middle|\mathbf{z}\right)P(\mathbf{z}) = \prod_{k=1}^{K} N(\mathbf{x};\boldsymbol{\mu}_k,\Sigma_k)^{z_k}\prod_{k=1}^{K}a_k^{z_k}\\ &= \prod_{k=1}^{K}\left[\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}\right]^{z_k}\prod_{k=1}^{K}a_k^{z_k}, \end{aligned} \quad (4)$$

where $\theta = \{a_k,\boldsymbol{\mu}_k,\Sigma_k\}_{k=1}^{K}$ is the parameter of the joint distribution. The likelihood function P($\mathbf{x}$|θ) is the integration of P($\mathbf{x}$,$\mathbf{z}$|θ) over $\mathbf{z}$:

$$\mathrm{P}\left(\mathbf{x}\middle|\theta\right) = \sum_{k=1}^{K}\frac{a_k}{(2\pi)^{\frac{D}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}. \quad (5)$$

A GMM is learned for each block in each one of the three color channels. With the obtained parameters, we can derive our FESS feature mapping online in the next section.

3.3 Free energy score space feature mapping and encoding

FESS seeks to derive feature maps based on the log likelihood function of a generative model, which in our case is the GMM. The feature maps are derived using variational EM algorithm [14] that deals with generative models whose log likelihood functions are intractable to be integrated. It derives a tractable lower bound for the intractable likelihood function so that learning and inference can be performed on the lower bound instead of the log likelihood.

For an observed sample $\mathbf{x}^i$, let $Q^i(\mathbf{z})$ denotes the approximate distribution of the posterior P($\mathbf{z}$|$\mathbf{x}^i$). In variational inference, $Q^i(\mathbf{z})$ is usually assumed to take the same form as P($\mathbf{z}$) but with different parameter $\mathbf{g}^i = (g_1^i\cdots,g_K^i)^{\mathrm{T}}$, so that:

$$Q^i(\mathbf{z}) = \prod_{k=1}^{K} g_{\mathrm{k}}^{i\,z_{\mathrm{k}}}. \quad (6)$$

The variational algorithm [14] derives a lower bound from Jensen's inequality to approximate the log likelihood function:

$$\log \mathrm{P}\left(\mathbf{x}^i \middle| \theta\right) \geq -\mathrm{KL}\left(Q^i(\mathbf{z}) \middle\| \mathrm{P}\left(\mathbf{x}^i, \mathbf{z}; \theta\right)\right) = -\mathcal{F}\left(Q^i, \theta\right), \tag{7}$$

where KL denotes the Kullback–Leibler divergence and $\mathcal{F}$ the variational free energy.

Based on the above joint distribution and approximate posterior distribution, the free energy function $\mathscr{F}$ for a given sample $\mathbf{x}^i$ can be formulated as:

$$\begin{aligned} \mathcal{F}\left(Q^i, \theta\right) &= \mathrm{E}_{Q^i(\mathbf{z})}\left[\log Q^i(\mathbf{z}) - \log \mathrm{P}\left(\mathbf{x}^i, \mathbf{z}; \theta\right)\right] \\ &= \mathrm{E}_{Q^i(\mathbf{z})}\left[\sum_{k=1}^{K} z_k \left(\sum_{d=1}^{D} -\frac{\left(x_d^i - \mu_d\right)^2}{2\sigma_d^2} - \log(2\pi)^{\frac{D}{2}} \prod_{d=1}^{D} \sigma_d\right) + \sum_{k=1}^{K} z_k \log \frac{g_k^i}{a_k}\right]. \end{aligned} \tag{8}$$

In Eq. (8), the lower bound of the log likelihood function is directly expanded according to the random variables. The resulting terms compose the feature map for FESS. The elements of the obtained FESS feature maps contain three groups according to the random variables:

$$\Phi_{\mathbf{x}}^{fit} = \sum_{k,d=1}^{K,D} g_k^i \left(-\frac{\left(x_d^i - \mu_d\right)^2}{2\sigma_d^2} - \log \sigma_{\mathrm{d}} (2\pi)^{\frac{D}{2}}\right) = \sum_{k=1}^{K} \Phi_{x_k}^{fit}, \tag{9}$$

$$\Phi_{\mathbf{z}}^{fit} = \sum_{k=1}^{K} g_k^i \log a_{\mathrm{k}} = \sum_{k=1}^{K} \Phi_{z_k}^{fit}, \tag{10}$$

$$\Phi_{\mathbf{z}}^{ent} = \sum_{k=1}^{K} g_k^i \log g_k^i = \sum_{k=1}^{K} \Phi_{z_k}^{ent}. \tag{11}$$

The fitness groups $\Phi_{\mathbf{x}}^{fit}$ and $\Phi_{\mathbf{z}}^{fit}$ measure how well the sample fits the model, the entropy group $\Phi_{\mathbf{z}}^{ent}$ measures how uncertain the fit is. The complete FESS feature mapping for sample $\mathbf{x}^i$ is the concatenation of the three groups:

$$\Phi\left(\mathbf{x}^i\right) = \mathbf{vec}\left(\left\{\Phi_{x_k}^{fit}, \Phi_{z_k}^{fit}, \Phi_{z_k}^{ent}\right\}_k\right). \tag{12}$$

The dimension of the mapping is $3 \times K$, where K denotes the mixture number of GMM.

Similar to FV [25], FESS feature mappings obtained for each pixel i in a certain block bj are summed up and averaged to give the FESS encoding for this block:

$$\Phi_{bj} = \frac{1}{n_b} \sum_{i=1}^{n_b} \Phi\left(\mathbf{x}^i\right), \tag{13}$$

where n_b is number of 7-d feature vectors in the block. The FESS encoding for each channel $\Phi_{ck}\{k=1,2,3\}$ is the concatenation of $\Phi_{bj}\{j=1,\cdots,12\}$ on all the blocks. Finally, $\Phi_{ck}\{k=1,2,3\}$ is L_2 normalized and stacked to get the feature vector Φ for re-identification, the size of which is $3 \times \mathrm{K} \times 3 \times 12$.

So far, discriminative features for re-identification are extracted by using Eq. (13) rather than directly stacking the local features in a long vector. By doing this, we get the following benefits. First, FESS encoding retains more information about the original image features, e.g., mean, covariance and second order statistics contained in Eq. (9). Second, feature mapping by Eq. (9) includes a data normalization procedure $(x_d^i - \mu_d)^2/(2\sigma_d^2)$, which reduces the metric difference among different feature dimensions. Finally, the feature mapping by Eqs. (10) and

(11) exploit the additional information contained in the hidden variable **z**. This information usually represents higher level concepts hid in the observed random variables, like the cluster or mixture center in image representation used in BoV model [29].

3.4 Similarity assessment and decision making

The derived feature vectors in Section 3.3 are used for person re-identification. In general, we have two sets of images: a probe set A and a gallery set B. Re-identification consists in associating each object of A to the corresponding object of B. The matching of two images I^A and I^B is carried out by estimating the feature matching distance d:

$$d\left(I^A, I^B\right) = d\left(\Phi\left(I^A\right), \Phi\left(I^B\right)\right). \tag{14}$$

In practice, any similarity metric can be used to measure the distance. For simplicity, Euclidean distance is adopted in our method. Final recognition objective function can be written as:

$$ID\left(I^A\right) = \arg\min_i \left\|\Phi\left(I^A\right) - \Phi\left(I_i^B\right)\right\|_2 \tag{15}$$

where $ID(I^A)$ is the identity of the probe I^A.

3.5 Time complexity analysis

In the offline part, the main source of computational cost is GMM learning for local feature vectors in each block. The time complexity for GMM learning is $O(K \cdot N_b \cdot I \cdot D)$, where K is the mixture number of GMM, N_b is the total number of local feature vectors used for training GMM of the corresponding block b, I is the number of iterations required for convergence, and D is the dimensionality of the local feature. Since we should learn GMM in each block, the time complexity of GMM learning can be estimated by $O\left(\sum_b K \cdot N_b \cdot I \cdot D\right)$.

In the online portion, the computational cost of FESS feature mapping for each image is $O(K \cdot N_{total} \cdot D)$, where N_{total} is the total number of local feature vectors in each image. The computational cost of person re-identification lies in the similarity computations between the target object and each object in the gallery. The time complexity is $O(N_P \cdot D_{im})$, where N_P is the number of objects in the gallery and D_{im} is the extracted feature length of each image. With the above two-stage implementation manner, the whole complexity of our method can be greatly reduced and its online part is very efficient, which is especially suited for real-time applications.

4 Experimental results and analysis

In this section, we conduct quantitative evaluation of the proposed method on several public datasets:

ETHZ [7], i-LIDS [38], i-LIDS-MA [1] and CAVIAR4REID [5]. Sample images of the datasets evaluated in our experiments are shown in Fig. 2. These datasets reflect different aspects of the above-described issues in person re-identification applications: viewpoint changes, illumination variations, occlusions, low resolution, etc. Furthermore, these datasets have been used in the recent literatures, allowing comparisons of different methods.

(a) ETHZ (b) i-LIDS (c) CAVIAR4REID

Fig. 2 Sample images taken from: a ETHZ, **b** i-LIDS, and **c** CAVIAR4REID datasets, with five pairs for each. Two images in the same column belong to the same person from different camera views

Our FESS encoding of the 7-d feature vector is denoted as "FESS+7d". We compare our method with another two encoding alternatives: FESS encoding of dense SIFT descriptors and BoV encoding of the 7-d feature vector, which are denoted as "FESS+dSIFT" and "BOV+7d" respectively. We also compare with other benchmark methods on these datasets. We implement our method and the two encoding alternatives by ourselves. All experiments are carried out in the HSV color space. To account for illumination changes that frequently occur in re-identification, a color correction technique is applied. Histogram equalization is performed in each RGB channel independently to obtain an invariant image representation. This step is helpful to diminish the appearance differences and make images captured from different camera views much more similar.

To simplify reproducibility, the VLFeat toolbox [31] is used to generate the dictionary for BoV and the GMM for FESS encoding. Dense SIFT descriptors are computed by using the vl_phow command included in the toolbox.

The mixture number of GMM is set to 16. This setting is a good tradeoff between efficiency and performance. We also test the identification performance using other settings (32, 48 and 64), and obtain similar results. For fair comparison, the dictionary size of BOV is also set to 16. Dense SIFT descriptors are extracted in the same manner as the 7-d feature vector. In detail, they are extracted with a spatial stride of one pixel, and at one scale, defined by setting the width of the spatial bin to 8 pixels. For computational efficiency, dense SIFT descriptors are reduced to 64-dimensions by performing PCA. PCA is performed separately on each channel and block.

For other methods, the results are given in the related references. We report re-identification results using cumulative matching characteristic (CMC) curve [32]. The CMC curve represents the expectation of finding the correct match in the top r matches. This evaluation technique is particularly used in identification systems where the input is a probe. The system has to return the matching results in descending order according to their similarity to the probe.

4.1 ETHZ dataset

ETHZ dataset was originally built for human detection purpose [7]. Video sequences contained in this dataset were captured from moving cameras. Schwartz et al. extracted a set of sample images for each different person in the video to test their PLS method for person re-identification [27]. The camera setup of this dataset provides a range of variations in people's appearance. Other challenges include illumination and background changes, low resolution and occlusions. The dataset is structured as follows: SEQ. #1 contains 83 persons, with 4857 images in total; SEQ. #2 contains

35 persons, with 1936 images in total; SEQ. #3 contains 28 persons, with 1762 images in total.

The same experimental settings in [4, 8] are reproduced to make fair comparisons. A subset of 5 images is randomly selected to create a signature for the probe and gallery. The results are shown in Fig. 3. On all the three sequences, our FESS encoding of the 7-d local feature vector significantly outperforms BoV. This is because FESS encoding captures more information than BOV. Also, FESS encoding of the 7-d feature vectors significantly outperforms FESS encoding of dense SIFT descriptors. Moreover, as the computational cost of GMM learning is $O(K \cdot N_b \cdot I \cdot D)$, training a GMM for the 7-d feature vectors is much faster than that for the dense SIFT descriptors. Consequently, the 7-d feature vector, in addition of being very compact and simple to compute, gives much better results than dense SIFT descriptors. These bear out our adoption of the 7-d feature vector for FESS encoding.

Further comparisons are made with two benchmark methods on these datasets: SDALF [8] and AHPE [4]. It is obvious that, our method outperforms SDALF and AHPE on all the three sequences. On SEQ. #1, the first rank matching rate is 95 % for our method versus 91 and 90 % for AHPE and SDALF respectively. On SEQ. #2 and SEQ. #3, our method gets 100 % recognition rate for ranks greater than 2. In general, our FESS encoding, which derive feature maps using the generative information of local image features, performs better than the method based on classical image features. It is also worthwhile to point out that, SDALF and AHPE need to partition the silhouette according to symmetry and asymmetry principles. Our method does not use foreground or body part segmentation. However, incorporating the body parts would be possible and could make the results even better.

We also compare our methods with two other approaches using similar local features as our method: eLDFV [20] and eBiCov [21]. eLDFV and our method use the same kind of local feature. To make fair comparisons, the mixture number of GMM is also the same. The difference lies in the encoding way of both methods: FV encoding of eLDFV and FESS encoding of our method. From Fig. 3 we can see that, our method outperforms eBiCov in multiple-shot cases on all the three sequences. Compared with eLDFV, FESS encoding gets inferior performance on SEQ. #1. On SEQ. #2 and SEQ. #3, it performs comparatively with eLDFV. Both of them get almost 100 % matching rate on these two sequences. It is difficult to account precisely for the reason of why FESS encoding gets inferior performance compared with eLDFV. One hypothesis is that, the approximated Fisher information matrix [25] introduced better normalization to the range of different dimensions of FV. Even though, the computational cost of eLDFV is much higher than our method. Besides computing the FV encoding of each image, eLDFV also needs to calculate the weighted color histograms and

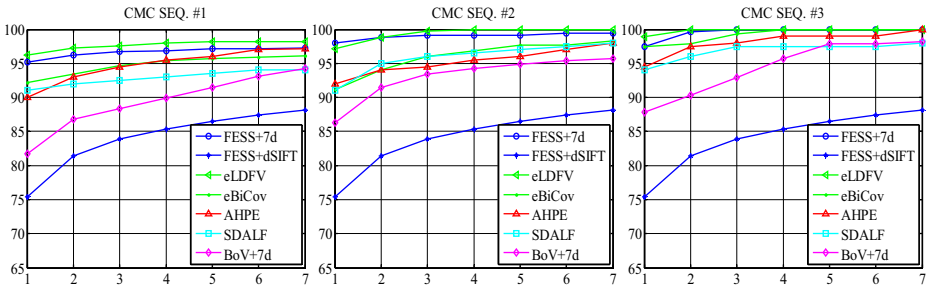

Fig. 3 CMC curves on SEQ. #1, SEQ. #2, and SEQ. #3 of the ETHZ dataset. To make fair comparisons with other methods, only the first 7 ranks are shown. All the compared methods are reported under multiple-shot settings

maximal stable color regions, both of which are very time consuming. In addition, the extracted feature vector of eLDFV is of higher dimensionality than ours. As the computation cost of re-identification is $O(N_P \cdot D_{im})$, higher dimensional feature vector will lead to much slower re-identification process, which will restricts the applicability of eLDFV.

It is obvious to notice that, the overall identification rate on SEQ. #1 is lower than that on SEQ. #2 and SEQ. #3. Especially on SEQ. #3 (with 28 persons), the identification rate for our method, eLDFV and eBiCov reach to 100 % for ranks greater than 2. The reason is that, SEQ. #1 has the largest number of persons, which makes the re-identification problem more difficult. These results confirm our statement that, for public surveillance cameras which record huge amount of objects every day, re-identification is challenging.

4.2 i-LIDS dataset

Images of this dataset are automatically extracted in [38] from the 2008 i-LIDS Multiple Camera Tracking Scenario (MCTS) dataset. As images are taken from multiple camera views, this dataset is very challenging. There are illumination variations and occlusions in this dataset. Unfortunately, it does not fit very well for person re-identification as the number of images per individual is very low. The reasonable assumption is that, it is possible and easy to get multiple images for a given individual with surveillance cameras. We do experiments on the modified version of this dataset. Individuals with at least 4 images are chosen, named i-LIDS$_{\geq 4}$ [4]. It total i-LIDS$_{\geq 4}$ dataset contains 59 individuals.

As SDALF and AHPE have reported their results on this dataset, comparisons are made with them, BOV encoding and FESS encoding of dense SIFT descriptors. We use the same experimental settings as SDALF and AHPE. In the single-shot case, 1 image is chosen to build the signature. In the multiple-shot case, as most objects have only 4 images, we choose N=2 images per person to build the signatures. The evaluation results are presented in Fig. 4a. Again, our method outperforms the two encoding alternatives. It is also worth noticing that, performance is not very high since images were captured from non-overlapping camera views subject to large variations on both view points and illumination conditions. In detail, when we choose N=1 image to build the signature, our method and BoV encoding outperform AHPE, especially for ranks greater than 5. Increasing the number of images for each object, the identification rate for our method, SDALF and AHPE are all improved. For ranks lower than 10, our method performs comparatively with AHPE, which is inferior to SDALF. The matching rate for our method gets consistent improvement over SDALF and AHPE for ranks

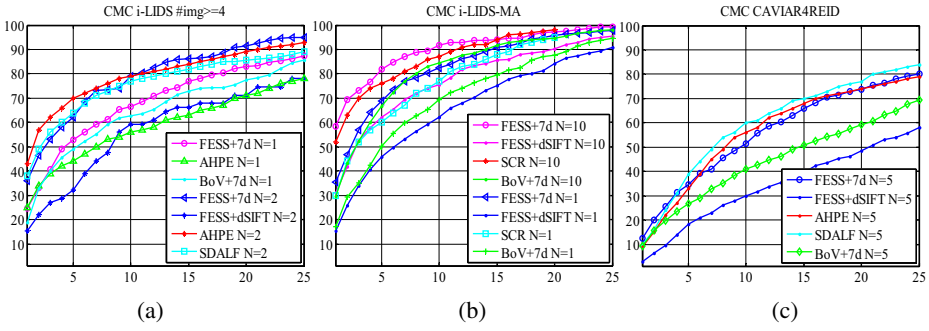

Fig. 4 CMC curves on different datasets: **a** results on i-LIDS$_{\geq 4}$, **b** results on i-LIDS-MA, and **c** results on CAVI AR4REID. We compare our method with SDALF [8], AHPE [4], and SCR [1]. The value of N is the number of images used to build to the signature for both probe and gallery

greater than 10. The scare amount of images per person in this dataset results in the low identification rate for the three approaches. To test the performance of our method with more images per person in real surveillance scenario, we choose another public dataset.

4.3 i-LIDS-MA dataset

This dataset is extracted by Bak et al. [3] from the MCTS dataset. It contains 40 objects extracted from two cameras. For each individual, there are 46 annotated images from both cameras views. Therefore, i-LIDS-MA dataset contains $40\times2\times46$ images. In [3], authors created human signatures using $N=1$ or $N=10$ randomly selected images, and evaluated the performance of their method on this dataset. To make fair comparisons with it, we follow the same experimental settings when evaluating our method and the two encoding alternatives. The average CMC curves of our method, BOV encoding of the 7-d feature, FESS encoding of dense SIFT descriptors and SCR are given in Fig. 4b.

In both the single-shot and the multiple-shot cases, our method outperforms all the compared methods. For our approach and SCR, similar local image features are used. The difference lies in the encoding way and different means to capture the spatial information. SCR employed region covariance to encode the local image features into a compact descriptor, and used a body part detector to detect different human body parts. The problems with SCR are that, computing region covariance and their distance is time consuming. Besides, the approach is not flexible enough and only applicable when the part detector works accurately.

4.4 CAVIAR4REID dataset

The last experiment is carried out on CAVIAR4REID dataset [5]. This dataset is a recently published dataset for person re-identification evaluation, which contains images of 72 objects and 50 of them were captured by two surveillance cameras mounted in a shopping center in Lisbon. The main complexity of this dataset is the presence of illumination changes, large pose variations and low resolution (e.g., varying from 17×39 pixels to 72×144 pixels). See Fig. 2 for visualization. The dataset is very challenging compared with other existing dataset in person re-identification literatures. We did experiments using images of these 50 objects. Images were resized to a size of 48×128 pixels in all experiments.

Following the same experimental settings of SDALF [8] and AHPE [4], we create human signatures using $N=5$ randomly selected images. Then, signatures from one camera compose our query set, while signatures from the other camera form the gallery set. The average CMC curve is given in Fig. 4c. Comparisons are made with the two encoding alternatives, SDALF and AHPE. Compared with the re-identification results on ETHZ and i-LIDS datasets, the recognition rate on this dataset is not high. Among all the three methods, SDALF performs the best. This is can be attribute to (1) images of CAVIAR4REID is of low resolution, which makes color the most effective cues for re-identification and (2) SDALF combined weighted color histograms and maximal stable color regions in an effective way by using symmetry and asymmetry information of the human body. Having good body segmentation is essential for re-identification. Our method performs comparatively with AHPE while at the same time outperforms BoV encoding. FESS encoding of dense SIFT descriptors performs the worst. It is worth noting that recognition performance can be greatly impacted by both the choice of local features and encoding methods. The top rank identification rate for SDALF is 10 %. The results also indicate that, in realistic scenarios, person re-identification is still an open and challenging problem. Examples of matching people using the proposed FESS encoding on i-LIDS$_{\geq4}$ and CAVIAR4REID datasets are shown in Figs. 5 and 6 respectively.

Fig. 5 Examples of person re-identification on i-LIDS$_{\geq 4}$ dataset using our FESS encoding method. In each row, the left-most image is the probe, the other images are the top 20 matched gallery images. The true match is highlighted with a *red box*

4.5 Other issues

Discriminative power of each block. In order to roughly capture spatial information, we divide the images into 3×4 blocks. For each block, we compute the FESS encoding of the local features in the corresponding color channel. We do experiments to test the performance of each block in our re- identification task. The feature vector for each block is the concatenation of the FESS encoding of the three color channels, the dimension of which is $3\times K\times 3$. The performance of each block is measured by CMC and the normalized Area Under Curve (nAUC) for CMC. nAUC gives the overall score of an identification method. Larger value of nAUC means better performance.

Table 2 reports the values of nAUC for different blocks on i-LIDS$_{\geq 4}$ dataset. In this table, $B_{mn}\{m=1,2,3,n=1,\cdots,4\}$ represents the feature vector for block in the i -th row and j -th column. From the table, we can draw the conclusion that, different part of the human body has different importance in re-identification. B_{12} and B_{13}, corresponding to the features of the upper body part, get the highest score in terms of nAUC. This coincides with how our humans perform re-identification. We usually

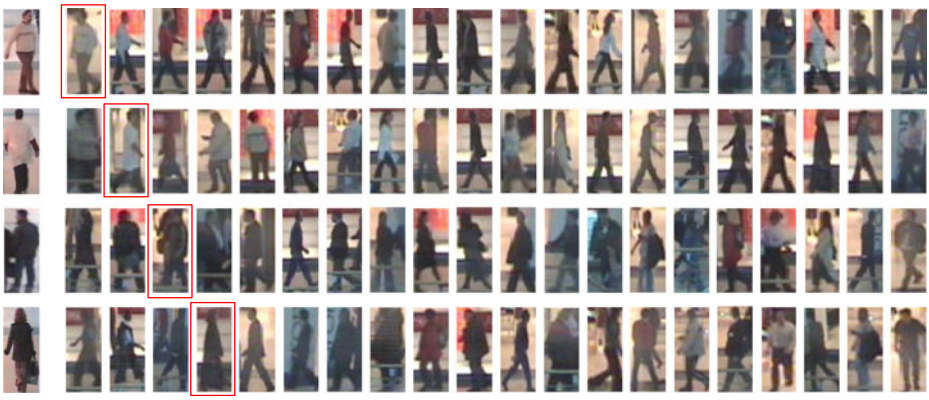

Fig. 6 Examples of person re-identification on CAVIAR4REID dataset using FESS encoding

Table 2 Valuses of nAUC on i-LIDS$_{\geq 4}$ dataset using different block features for re-identification

Block features	B_{11}	B_{12}	B_{13}	B_{14}	B_{21}	B_{22}	B_{23}	B_{24}	B_{31}	B_{32}	B_{33}	B_{34}
nAUC	59.52	72.14	67.29	55.10	61.12	63.78	62.31	61.43	59.37	63.52	62.53	56.49

focus our attention on the upper part of the body. Besides, for an input image, features from the four corners (i.e., B_{11},B_{14},B_{31},B_{34}) get lower performances. The reason is that, the corner of the image is usually related to the background, which changes under different camera views. In practice, when we concatenate the feature vectors from the other 8 blocks (i.e., B_{12},B_{13},B_{21},B_{22},B_{23},B_{24},B_{32},B_{33}), we get better identification results compared with concatenating all feature vectors of the 12 blocks.

Learning of GMM model. Another issue concerning our FESS encoding for re-identification is the learning of GMM model for calculating FESS feature mapping. In the previous experiments, every GMM is trained using images coming from the same dataset being evaluated. We do experiments to see the performance of using GMM models learned with different datasets. Figure 7 shows the CMC curves on the three sequences of ETHZ dataset using GMMs learned with different datasets. The name of the dataset used to train the GMM is added to "GMM" for comparison. Overall, our FESS encoding framework with either similar or dissimilar datasets to learn the GMMs gets higher performance than benchmark methods, e.g., SDALF. In practice, a representative dataset should be collected beforehand. GMMs for each block and each channel can be trained with this dataset. When performing re-identification, we only need to map the local features of the probe and gallery to the corresponding GMM and gets the FESS encoding feature vector.

Speed of encoding. For BOV the encoding time is dominated by nearest neighbor search, which increases not only with the size of the dictionary but also the number of nearest neighbor sought. Using 2000 images taken from the ETHZ dataset, the training of a dictionary with 16 visual words takes 23 s for BOV, and 73 s for training a GMM with 16 centers for FESS. After getting the dictionary or GMM model, both BOV and our FESS encoding take <0.1 s per image using MATLAB implementation. All timings are run on a 2.67GHz Intel CPU and 4 GB RAM. The encoding part requires less time compared with the GMM learning part, and is suitable for real-time video surveillance applications.

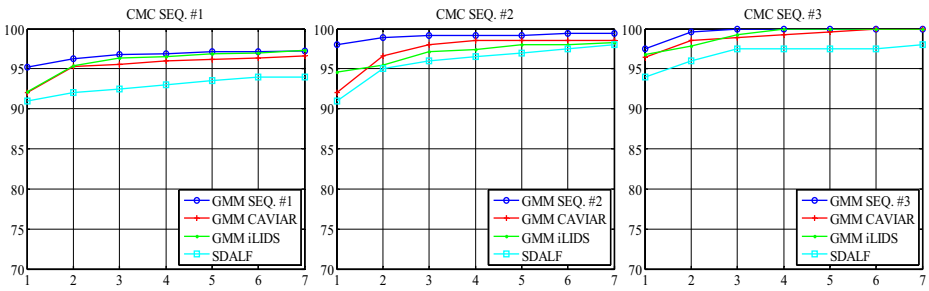

Fig. 7 CMC curves on SEQ. #1, SEQ. #2, and SEQ. #3 using different training datasets to get the GMM for calculating FESS feature mapping. The name of the dataset used for learning GMM is added to "GMM"

5 Conclusions

In this paper, we proposed a new appearance based method for person re-identification. This method derives high level meaningful new features from local visual cues by computing the free energy score space of the log likelihood function of a Gaussian Mixture Model. Free energy score space feature mapping is a generative score space that measures how well a sample fits a generative model and how uncertain the fit is. The derived feature maps are encoded into a fixed length feature vector for recognizing objects from disjoint camera views. Quantitative experiments are conducted on several public datasets against benchmark methods in the literature. Experimental results demonstrate that our method outperforms bag-of-visual-words in person re-identification. At the same time, it performs favorably against leading methods.

Our method can be divided into two parts: the offline part and the online part. The offline part consists of local feature extraction and GMM learning. The online part comprises FESS encoding and re-identification. Compared with the offline part, the online part has much lower computational cost, rendering our method appropriate for real time applications. Furthermore, the images used to learn the GMM model may have different imaging conditions from those in the probe/gallery, without any significant loss in performance. In practice, GMMs could be learned beforehand using representative datasets.

As future works, our FESS encoding may be embedded into supervised metric learning strategies, in order to further improve the performance. Foreground extraction and human body part division can also be incorporated. Currently, we apply the FESS encoding to person re-identification problem. As an encoding alternative, is can also be applied to other computer vision tasks, i.e., object recognition and image retrieval.

Acknowledgments This research has been partially supported by the funding from China 2011CB302203, NSFC 61375019 and NSFC 61273285.

References

1. Bak S, Corvee E, Bremond F, Thonnat M (2010) "Person re-identification using spatial covariance regions of human body parts." in Proc IEEE Int Conf Adv Video Signal-Based Surveill (AVSS): 435–440
2. Bak S, Corvee E, Bremond F, Thonnat M (2010) ". Person re-identification using Haar-based and DCD-based signature." in Proc IEEE Int Conf Adv Video Signal-Based Surveill (AVSS):1–8
3. Bak S, Corvee E, Bremond F, Thonnat M (2012) Boosted human re-identification using riemannian manifolds. Image Vis Comput 30(6):443–452
4. Bazzani L, Cristani M, Perina A, Murino V (2012) Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recogn Lett 33(7):898–903
5. Cheng D, Cristani M, Stoppa M, Bazzani L, Murino V (2011) "Custom pictorial structures for re-identification." in Proc Br Mach Vision Conf (BMVC)
6. Dikmen M, Akbas E, Huang T, Ahuja N (2010) "Pedestrian recognition with a learned metric," in Proc Asian Conf Comput Vision (ACCV): 501–512
7. Ess A, Leibe B, Van Gool L (2007) "Depth and appearance for mobile scene analysis." in Proc IEEE Int Conf Comput Vision (ICCV): 1–8
8. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2013) Symmetry-driven accumulation of local features for human characterization and re-identification. Comput Vis Image Underst 117(2):130–144
9. Gheissari N, Sebastian T, Hartley R (2006) "Person reidentification using spatiotemporal appearance." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR): 1528–1535

10. Gray D, Tao T (2008) "Viewpoint invariant pedestrian recognition with an ensemble of localized features." in Proc Eur Conf Comput Vision (ECCV): 262–275
11. Hirzer M, Beleznai C, Roth P, Bischof H (2011) "Person re-identification by descriptive and discriminative classification," in Image Anal: 91–102
12. Holub A, Welling M, Perona P (2008) Hybrid generative-discriminative visual categorization. Int J Comput Vis 77(1):239–258
13. Jaakkola T, Haussler D (1999) "Exploiting generative models in discriminative classifiers." Adv Neural Inf Process Syst: 487–493
14. Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) An introduction to variational methods for graphical models. Mach Learn 37:183–233
15. Kostinger M, Hirzer M, Wohlhart P, Roth P, Bischof H (2012) "Large scale metric learning from equivalence constraints," in Proc IEEE Conf Comput Vision Pattern Recog (CVPR): 2288–2295
16. Kviatkovsky I, Adam A, Rivlin E (2013) "Color invariants for person reidentification,". IEEE Trans Pattern Anal Mach Intell (PAMI) 35(7):1622–1634
17. Li X, Lee T, Liu Y (2011) "Hybrid generative-discriminative classification using posterior divergence." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR): 2713–2720
18. Li X, Wang B, Liu Y, Lee T (2013) "Learning Discriminative Sufficient Statistics Score Space for Classification." Mach Learn Knowl Discov Databases: 49–64
19. Li W, Zhao R, Wang X (2012) "Human reidentification with transferred metric learning." in Proc Asian Conf Comput Vision (ACCV): 31–44
20. Ma B, Su Y, Jurie F (2012) "Local descriptors encoded by fisher vectors for person re-identification." in Proc Eur Conf Comput Vision Workshops Demonstrations: 413–422
21. Ma B, Su Y, Jurie F (2012) "Bicov: a novel image representation for person re-identification and face verification." in Proc Br Mach Vision Conf (BMVC)
22. Mignon A, Jurie F (2012) "PCCA: A new approach for distance learning from sparse pairwise constraints." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR): 2666–2672
23. Oneata D, Verbeek J, Schmid C (2013) "Action and event recognition with Fisher vectors on a compact feature set." in Proc IEEE Int Conf Comput Vision (ICCV):1817–1824
24. Perina A, Cristani M, Castellani U, Murino V, Jojic N (2012) Free energy score spaces: using generative information in discriminative classifiers. IEEE Trans Pattern Anal Mach Intell 34(7): 1249–1262
25. Perronnin F, Dance C (2007) "Fisher kernels on visual vocabularies for image categorization." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR):1–8
26. Prosser B, Zheng W, Gong S, Xiang T (2010) "Person re-identification by support vector ranking." in Proc Br Mach Vision Conf (BMVC): 1–11
27. Schwartz W, Davis L (2009) "Learning discriminative appearance-based models using partial least squares." in Proc Braz Symp Comput Graph Image Process (SIBGRAPI):322–329
28. Simonyan K, Parkhi O, Vedaldi A, Zisserman A (2013) "Fisher vector faces in the wild." in Proc Br Mach Vision Conf (BMVC)
29. Sivic J, Zisserman A (2003) "Video google: a text retrieval approach to object matching in videos." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR): 1470–1477
30. Sun C, Nevatia R (2013) "Large-scale web video event classification by use of fisher vectors." in Proc IEEE Workshop Appl Comput Vision (WACV): 15–22
31. Vedaldi A, Fullerson B (2010) "VLFeat – An open and portable library of computer vision algorithms." in Proc ACM Int Conf Multimed
32. Wang X, Doretto G, Sebastian T, Rittscher J, Tu P (2007) "Shape and appearance context modeling." in Proc IEEE Int Conf Comput Vision (ICCV):1–8
33. Wu Y, Minoh M, Mukunoki M, Lao S (2012) "Set based discriminative ranking for recognition." in Proc Eur Conf Comput Vision (ECCV): 497–510
34. Zhang C, Li X, Ruan X, Zhao Y, Yang M (2013) "Discriminative generative contour detection." in Proc Br Mach Vision Conf (BMVC)
35. Zhao R, Ouyang W, Wang X (2013) "Unsupervised Salience Learning for Person Re-identification." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR)
36. Zhao Y, Zhao X, Liu Y (2014) "Person re-identification by free energy score space encoding." in Proc Int Conf Image Process (ICIP)
37. Zheng W, Gong S, Xiang T (2009) "Associating groups of people." in Proc Br Mach Vision Conf (BMVC): 6–16
38. Zheng W, Gong S, Xiang T (2011) "Person re-identification by probabilistic relative distance comparison." in Proc IEEE Conf Comput Vision Pattern Recog (CVPR): 649–656

Yanna Zhao received the M.S. 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.

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.

Ruotian Luo has been an undergraduate student in Shanghai Jiao Tong University since 2011, and will earn his bachelor degree in 2015. He is a student of IEEE Honor Class, and is IEEE student member. His primary area of study is computer science. He is currently a research assistant in VisionLab, Department of Automation at Shanghai Jiao Tong University. He was a research assistant at VIVA Lab, University of Ottawa in 2014 summer. His research interests include computer vision and machine learning.

Yuncai Liu received his Ph.D. 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.