PERSON RE-IDENTIFICATION BY FREE ENERGY SCORE SPACE ENCODING

Yanna Zhao

Shandong University
School of Information Science and Engineering
Jinan, China

Xu Zhao, Yuncai Liu

Shanghai Jiao Tong University
School of Electronic Information
and Electrical Engineering
Shanghai, China

ABSTRACT

Person re-identification is an important and challenging computer vision problem. Recent progress in this area is due to new visual features and models that deals with cross-view variations. Instead of working towards more complex models, we focus on low level features and their encoding. Low level features capturing the color and structural information are first extracted from human images. Gaussian Mixture Model (GMM) is then employed to approximate the distribution of the features, providing a relatively comprehensive statistical representation. Finally, low level features are mapped to a space by computing free energy score of the GMM. The mapped features are encoded into a fixed-length feature vector for person re-identification. Extensive experiments are conducted on several public datasets. Comparisons with benchmark person re-identification methods show the promising performance of our approach.

Index Terms— person re-identification, appearance modeling, Gaussian Mixture Model, free energy score space

1. INTRODUCTION

Recent studies on person re-identification have concentrated on the appearance based methods. In this situation, it is assumed that individuals do not change their clothes among different sightings. Two lines of research have been pursued here. On one hand, there are learning based methods, irrespective of the image representation [1, 2, 3, 4, 5, 6, 7, 8]. An ensemble of localized features has been proposed in [1]. Instead of characterizing human appearance by a specific feature, a machine learning algorithm was used to construct a model by combining spatial and color information. The re-identification has been reformulated as a ranking problem in [4]. An informative subspace was learned in which the potential true match gets highest ranking. PRDC [2] formulated person re-identification as a probabilistic relative distance comparison problem, which aims to maximize the probability that a pair of true match has a smaller distance than that of a wrong match pair. In [8], a transferred metric learning framework was proposed to learn a specific metric for every query-gallery pair. In [3], image spaces of two camera views were partitioned into different configurations and a specific metric was learned in the locally aligned common feature space. Learning based methods may suffer from generalization problems. When new targets enter, the model has to be re-trained and updated.

The other big branch of appearance based methods is appearance modeling, producing a discriminative and invariant representation for human images [9, 10, 11, 12, 13, 14, 15, 16]. The SDALF method [9] combined weighted color histograms with maximal stable color region descriptors and recurrent local color patches, setting the state-of-the-art performance on several benchmark datasets. A similar proposal [10] was to combine color histograms with epitome - highly informative patches from a set of images to enhance appearance representation. Pictorial structure was adopted in [11] to finely localize the human body parts, obtaining distinctive features for matching signatures of individuals. Fisher vector was employed in [13] to encode higher order statistics of local features using generative information. In [14], a descriptor combining Gabor filters and the covariance descriptor was developed to handle illumination changes and background variations. In [15], part-based models were adopted to handle pose variation.

In this paper, we propose a new appearance modeling approach for person re-identification. We extract low level features to capture color and structural cues of an image. These low level features are mapped to a score space based on the free energy score [17] measure of Gaussian Mixture Model (GMM). GMM approximates the generation process of the low lever features, providing a good representation of the data distribution. Free energy score space (FESS) [17] feature mapping measures how well a sample fits a random variable and how uncertain the fit is and produces discriminative information. Finally, this information is encoded into a fixed-length feature vector for person re-identification.

This research has been partially supported by the grants of China 973 project 2011CB302203, STCSM project 11JC1405200 and NSFC 61375019 and NSFC 61273285.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

2. FREE ENERGY SCORE SPACE ENCODING

In this section, we detail our person re-identification approach by computing FESS of GMM. First, low level features are extracted from the input images. Second, GMM is employed to model the distribution of the low level features. Finally, discriminative information is first obtained by FESS mapping and then encoded into fixed-length feature vector for person re-identification.

2.1. Low Level Image Feature

The input images corresponding to human detection or tracking results are divided into 3×4 blocks. Low level features are extracted from the HSV color space using the 7-d feature vector proposed in [13]:

$$\begin{aligned} & f(x, y, I) \\ & = (x, y, I(x, y), I_x(x, y), I_y(x, y), I_{xx}(x, y), I_{yy}(x, y)). \end{aligned} \tag{1}$$

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

To extract discriminative information from the low level features, we propose a mapping method by calculating FESS [17] of the GMM. FESS is a generative model that derives feature maps based on the log likelihood function of a model. The lower bound of log likelihood (see Eq. (7)) is expanded according to the random variable, and each resulting term becomes a feature map. FESS has been successfully used in various vision tasks [17, 18, 19]. We transform the low level feature vectors based on this mapping to obtain discriminative information for person re-identification. The generative model in our method is GMM as it has been widely used in modeling the distribution of image features [20].

2.2. Gaussian Mixture Model

Let $\mathbf{x} \in R^D$ be the observed random variable. In our context of person re-identification, $\mathbf{x}$ denotes the low level image features. Let $\mathbf{z} = \{z_1, \cdots, z_K\}$ be a set of K hidden variables. Where $z_k = 1$ indicates that the k-th mixture center is selected to generate the sample $\mathbf{x}$ and $z_k = 0$ otherwise. The prior distribution of $\mathbf{z}$ is typically chosen to be:

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

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

Let $\boldsymbol{\mu}_k$ and Σ_k be the mean and variance matrix of the k-th mixture center respectively. Here we assume that the covariance matrices are diagonal. We do this for two reasons: (1) a weighted sum of Gaussians with diagonal covariances can approximate any distribution with an arbitrary precision; (2) the computational cost of diagonal covariance is much lower [20]. We use the notation $\Sigma_k = diag(\sigma_{k1}^2, \cdots, \sigma_{kD}^2)$. Then the distribution of $\mathbf{x}$ conditioned on the hidden variable $\mathbf{z}$ can be written as:

$$\begin{aligned} P(\mathbf{x}|\mathbf{z}) &= \prod_{k=1}^{K} N(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k)^{z_k} \\ &= \prod_{k=1}^{K} [\frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\}]^{z_k} \end{aligned} \tag{3}$$

The joint distribution of GMM can be formulated as:

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

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

$$P(\mathbf{x}|\theta) = \prod_{k=1}^{K} \frac{a_k}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\} \tag{5}$$

GMM is trained for each block in each one of the three color channels. With the obtained GMM, we get our feature mapping by computing FESS of GMM.

2.3. Free Energy Score Space Encoding

FESS intends to derive feature maps based on the variational EM algorithm [21]. It derives a lower bound for the likelihood function so that learning and inference can be performed on the lower bound instead of the log likelihood.

For any observed sample $\mathbf{x}^i$, let $Q^i(\mathbf{z})$ denotes the approximate distribution of the posterior distribution $P(\mathbf{z}|\mathbf{x}^i)$. The approximate posterior distribution $Q^i(\mathbf{z})$ is assumed to take the same form with $P(\mathbf{z})$ but with different parameter $\mathbf{g}^i = (g_1^i, \cdots, g_k^i)^T$ [21], that is:

$$Q^i(\mathbf{z}) = \prod_{k=1}^{K} {g_k^i}^{z_k}. \tag{6}$$

The variational lower bound of the log likelihood function is:

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

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

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

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

FESS uses the resulting terms as feature maps [17]. The elements of the obtained score function contain three groups:

$$\Phi_x^{fit} = \sum_{k,d=1}^{K,D} g_k^i(-\frac{(x_d^i-\mu_d)^2}{2\sigma_d^2} - \log\sigma_d(2\pi)^{\frac{D}{2}}) = \sum_{k=1}^{K}\Phi_{x_k}^{fit}. \tag{9}$$

$$\Phi_z^{fit} = \sum_{k=1}^{K} g_k^i \log a_k = \sum_{k=1}^{K}\Phi_{z_k}^{fit}. \tag{10}$$

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

The fitness group Φ^{fit} measures how well the sample fits the model, the entropy group Φ^{ent} measures how uncertain the fit is. Therefore, for input sample $\mathbf{x}^i$, we obtain a set of feature mapping under the GMM:

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

The dimension of the mapping is $3 \times K$.

Discriminative features are extracted by using Eq. (12) rather than directly stacking the low level feature vector. By using this feature mapping, we get two benefits. First, the feature mapping by (9) includes a data normalization procedure $(x_d^i - \mu_d)^2/(2\delta_d^2)$, which reduces the metric difference among different features. Second, the feature mapping by Eq. (10) and Eq. (11) exploit the additional information contained in the hidden variable $\mathbf{z}$. This information usually represents higher level concepts hid in the observed random variables, like the cluster or mixture center in image representation using the bag-of-words model [22].

Similar to [20], FESS maps obtained for each pixel in a certain block are summed up and averaged to give the FESS encoding for this block:

$$\Phi = \frac{1}{N_p}\sum_{i=1}^{N_p}\Phi(\mathbf{x}^i), \tag{13}$$

where N_p is the number of pixels in the block. The FESS encodings obtained on all the blocks in the three channels are concatenated to get the final image representation. The size of the representation is $3 \times K \times 3 \times 12$.

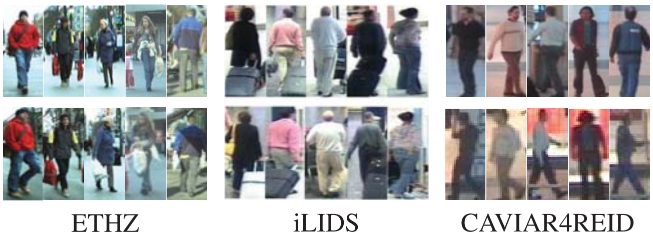

Fig. 1. Sample images from ETHZ, iLIDS, and CAVIAR4REID datasets. Images in the same column belong to the same person. For each dataset, five pairs of images are given.

3. EXPERIMENTAL ANALYSIS

The evaluation of the proposed method is carried out on three publicly available datastes: ETHZ [23], a variant of iLIDS for re-identification [2] and CAVIAR4REID [11]. Example images are shown in Fig.1. Every image is scaled into a fixed size of 128 × 64 pixels. The number of GMMs is set to 16 in all the experiments. 1-Nearest Neighbor classifier is adopted for decision making. The results are shown in terms of recognition rate, by the Cumulative Matching Characteristic (CMC) curve, which represents the expectation of finding the correct match in the top r matches.

ETHZ dataset [23]. This dataset contains three subdatasets: ETHZ1 contains 83 persons, with 4857 images; ETHZ2 contains 35 persons, with 1961 images; ETHZ3 contains 28 persons, with 1762 images. The moving cameras setup provides a rang of variations in peoples' appearance. The challenges are illumination changes, occlusion and low resolution.

We randomly selected 5 images for each person to form the probe and gallery. Comparisons are made with two benchmark methods: SDALF [9] and AHPE [10]. Experimental results are shown in Fig.2. As can be seen, our method gives much better results than SDALF and AHPE on all of the three sequences. Especially, on ETHZ2 and ETHZ3, our method gets 100% recognition rate for ranks greater than 2. We also compare our methods with two other descriptors using the same local features: eLDFV [13] and eBiCov [14]. From Fig.2, we can see that, our method outperforms eBiCov in multiple shot case. The rank 1 recognition rate on ETHZ1 and ETHZ2 for our method are 95% and 98% versus 92% and 91% for eBiCov. Our method is inferior to eLDFV on ETHZ1, but on ETHZ2 and ETHZ3, our method alone performs comparatively with eLDFV, which is the combination of LDFV, wHSV and MSCR [13].

iLIDS dataset for re-identification [2]. Images of this dataset are captured indoor at a busy airport arrival hall. As images are taken from non-overlapping cameras, illumination changes and occlusions are quite large for this dataset (see Fig. 1). We use the modified version of the dataset. Individu-

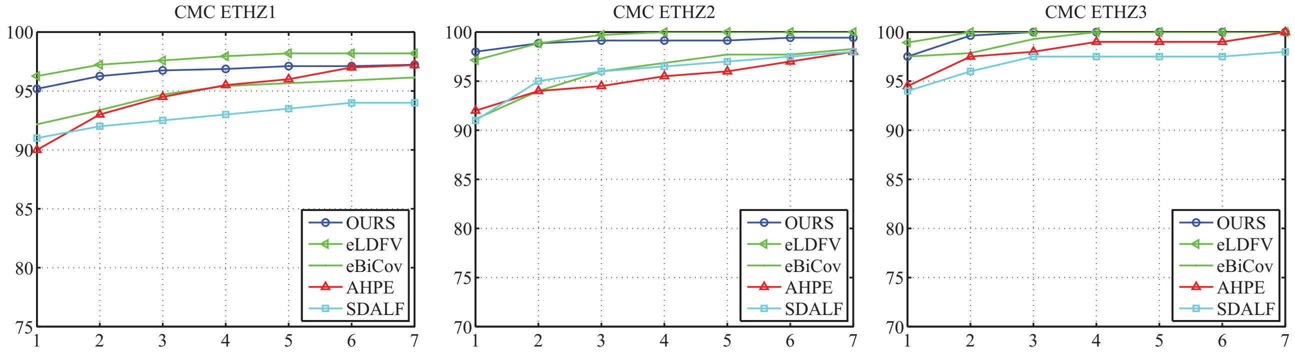

Fig. 2. CMC curves obtained on ETHZ dataset. All the results are obtained using multiple images per person, the number of images is set to 5. Comparisons are made with SDALF [9], AHPE [10], eLDFV [13] and eBiCov [14].

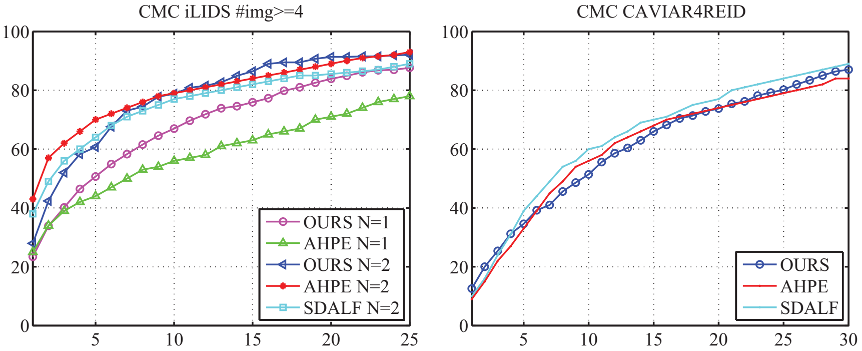

Fig. 3. CMC curves obtained on iLIDS$_{\geq 4}$ (left) and CAVIAR4REID datasets. For CAVIAR4REID dataset, the number of images is set to 5. Comparisons are made with SDALF [9] and AHPE [10].

als with at least 4 images are chosen, named iLIDS$_{\geq 4}$ [10].

Fig.3 (left) illustrates the performance of the proposed method and two state-of-the-art methods on this dataset. In the single shot case, our method outperforms AHPE [10], especially for ranks greater than 5. Increasing the number of images, the performances of the three methods are all improved. AHPE and SDALF outperform our method for ranks between 1 and 6. The recognition rate of our method is comparative with AHPE for ranks greater than 6, showing consistent improvements over the results obtained by SDALF.

CAVIAR4REID dataset [11]. This dataset 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 pixels to 72 × 144 pixels. 50 different individuals are captured from two camera views, with 10 images for each view. Challenges of dataset are low resolution, occlusion, and pose variation.

Fig.3 (right) gives the performance of our method, SADLF, and AHPE on CAVIAR4REID dataset. The overall recognition rate on this dataset is not high. Among the three methods, SDALF performs the best. In multiple shot cases, our method performs comparatively with AHPE. Those results indicate that, in real world scenario, person re-identification is more challenging.

4. CONCLUSION

In this paper, we propose a new image representation for person re-identification. The new representation is obtained by encoding local image features through free energy score space computing of the log likelihood of Gaussian Mixture Model. We use free energy score space to exploit more information rather than using low level image features, generating a discriminative fixed-length feature vector. Experiments are conducted on three benchmark datasets against several state-of-the art methods. Convincing experimental results demonstrate the effectiveness of the proposed method for person re-identification. In the future, we will design more effective methods for person re-identification.

5. ACKNOWLEDGMENT

The iLIDS dataset used in our experiments is by courtesy of the i-LIDS team.

6. REFERENCES

[1] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, pp. 262–275. 2008.

[2] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.

[3] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013, pp. 3594–3601.

[4] B. Prosser, W. Zheng, S. Gong, T. Xiang, and Q.Mary, "Person re-identification by support vector ranking," in *BMVC*, 2010, pp. 1–11.

[5] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *Image and Vision Computing*, vol. 30, no. 6, pp. 443–452, 2012.

[6] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2666–2672.

[7] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, pp. 780–793. 2012.

[8] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, pp. 31–44. 2012.

[9] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer vision and image understanding*, vol. 117, no. 2, pp. 130–144, 2012.

[10] 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.

[11] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *BMVC*, 2011.

[12] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[13] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *ECCV Workshop*, 2012, pp. 413–422.

[14] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *BMVC*, 2012.

[15] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *AVSS*, 2010, pp. 435–440.

[16] C. Liu, S. Gong, and C. Loy, "On-the-fly feature importance mining for person re-identification," *Pattern Recognition*, vol. 47, no. 4, pp. 1602–1615, 2014.

[17] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score spaces: using generative information in discriminative classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1249–1262, 2012.

[18] X. Li, T. Lee, and Y. Liu, "Hybrid generative-discriminative classification using posterior divergence," in *CVPR*, 2011, pp. 2713–2720.

[19] Xiong Li, Bin Wang, Yuncai Liu, and Tai Sing Lee, "Learning discriminative sufficient statistics score space for classification," in *ECML*, pp. 49–64. 2013.

[20] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007, pp. 1–8.

[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[22] C. Zhang, X. Li, X. Ruan, Y. Zhao, and M. Yang, "Discriminative generative contour detection," in *BMVC*, 2013.

[23] W. Schwartz and L. Davis, "Learning discriminative appearance-based models using partial least squares," in *SIBGRAPI*, 2009, pp. 322–329.