An Ensemble of Invariant Features for Person Reidentification

Young-Gun Lee, *Student Member, IEEE*, Shen-Chi Chen,
Jenq-Neng Hwang, *Fellow, IEEE*, and Yi-Ping Hung, *Member, IEEE*

***Abstract*— This paper proposes an ensemble of invariant features (EIFs), which can properly handle the variations of color difference and human poses/viewpoints for matching pedestrian images observed in different cameras with nonoverlapping field of views. Our proposed method is a direct reidentification (re-id) method, which requires no prior domain learning based on prelabeled corresponding training data. The novel features consist of the holistic and region-based features. The holistic features are extracted by using a publicly available pretrained deep convolutional neural network used in generic object classification. In contrast, the region-based features are extracted based on our proposed two-way Gaussian mixture model fitting, which overcomes the self-occlusion and pose variations. To make a better generalization during recognizing identities without additional learning, the ensemble scheme aggregates all the feature distances using the similarity normalization. The proposed framework achieves robustness against partial occlusion, pose, and viewpoint changes. Moreover, the evaluation results show that our method outperforms the state-of-the-art direct re-id methods on the challenging benchmark viewpoint invariant pedestrian recognition and 3D people surveillance data sets.**

***Index Terms*— 3D people surveillance (3DPeS), deep convolutional neural network (DCNN), ensembles of invariant features (EIFs), person reidentification (re-id), two-way Gaussian mixture model fitting (2WGMMF), viewpoint invariant pedestrian recognition (VIPeR).**

I. Introduction

THERE is increasing interest in video surveillance due to the growing availability of cheap sensors and computing power, as well as a growing need for applications both in public and private environments, such as homeland security, business monitoring, smart city planning, crime prevention, monitoring patients, elderly and children at home, and so on. Person reidentification (re-id) is a problem of recognizing and matching a person at different locations and/or at different times in multiple cameras or in a video archive [1]. In video surveillance scenarios, the sizes of captured persons are usually too small to apply face recognition techniques; therefore, most of the existing person re-id methods [2]–[6] exploit descriptors of appearance to solve the problem. However, it is still challenging, because observed human objects undergo significant variations of pose, viewpoint, illumination, and color response among cameras, even though they do not change their clothes during the time being captured in multiple cameras.

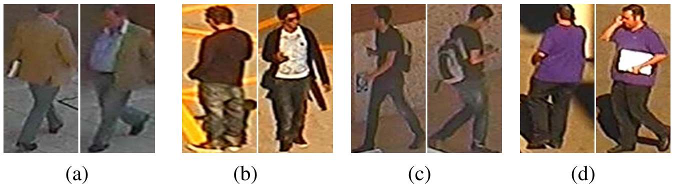


Fig. 1. Examples on the 3DPeS. (a) ID 22. (b) ID 195. (c) ID 59. (d) ID 197.

As shown in Fig. 1, four pairs of images from 3D people surveillance (3DPeS) (IDs, 22, 195, 59, and 197) [7] are captured in different poses with different cameras, so shirt appearances of Fig. 1(a) and (b) are only visible in the right-hand side images. Moreover, the bag and the book carried by the persons are available in the right-hand side images of Fig. 1(c) and (d). Since unseen region only exists in one of image pair, intra-personal variations can be even larger than inter-personal variations in most existing color feature representations (see Section IV-B1). To solve this problem, [8]–[11] exploit 3D body models by using multiple shots, motion detection, and camera calibration. However, the body pose or 3D information is not usually available in practical cases.

many cases, a person can easily be reidentified based on the possessed distinct features. For example, in Fig. 1, ID 22 dresses a brown jacket, ID 195 wears a dark colored cardigan, ID 59 carries a black and white colored bag in the back, and ID 197 puts on a purple shirt. In human perception, these features can be effectively exploited to find the correct pair. On the contrary, other salient regions, i.e., the blue shirt in Fig. 1(a), the white shirt in Fig. 1(b), and the book in Fig. 1(d), are less useful in the re-id process. Furthermore, even though white region of the bag, carried by ID 59, is shown slightly in the left-hand side image, human can utilize it for recognizing the same person. This motivates us to reidentify a person by finding distinct region and utilizing the small dominant color regions. We propose a novel direct re-id method based on specially designed invariant features.

Manuscript received December 10, 2015; revised April 29, 2016, July 15, 2016, and September 16, 2016; accepted November 3, 2016. Date of publication December 9, 2016; date of current version March 3, 2017. This paper was recommended by Associate Editor A. Cavallaro.

Y.-G. Lee and J.-N. Hwang are with the Department of Electrical Engineering, University of Washington, Seattle, WA 98105 USA (e-mail: lygstj@uw.edu; hwan@uw.edu).

S.-C. Chen is with Huawei Research and Development, Santa Clara, CA, USA. (e-mail: d98922030@csie.ntu.edu.tw).

Y.-P. Hung is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: hung@csie.ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2637818

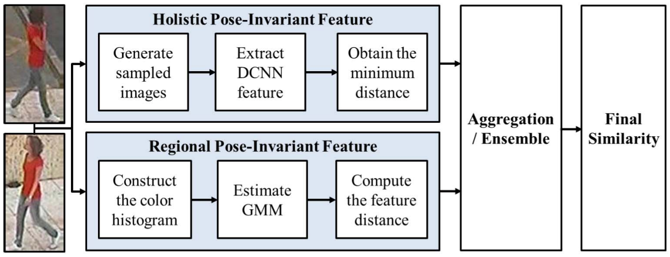

Holistic Pose-Invariant Feature
Generate sampled images
Extract DCNN feature
Obtain the minimum distance
Regional Pose-Invariant Feature
Construct the color histogram
Estimate GMM
Compute the feature distance
Aggregation / Ensemble
Final Similarity

Fig. 2. Overview of the proposed framework.

In this paper, we present a novel appearance-based re-id framework, which uses an ensemble of invariant features to achieve robustness against partial occlusion, camera color response variation, pose and viewpoint changes, and so on. The diagram of the proposed framework is shown in Fig. 2. The proposed method not only solves the problems resulted from the changing human pose and viewpoint, with some tolerance of illumination changes, but also can skip the laborious calibration effort and restriction.

An early version of this paper appeared in [12]. In addition to giving a more detailed description of our proposed ensemble of invariant features (EIFs), the main addition of contents in this paper includes: 1) investigation of the effectiveness of all possible deep convolutional neural network (DCNN) features with both general pretrained neural network and Siamese network; 2) extension of regional invariant feature with the joint histogram; 3) more in-depth discussions and analyses on the proposed features; and 4) more extensive experimental evaluations.

The main contributions of this paper can be summarized as follows.

1) We propose a way of describing distinct visual characteristics and utilizing the dominant color modes from pedestrian images based on human perception. This descriptor, called EIFs, is the core of our approaches. The proposed features consist of the holistic and regional pose-invariant features.
2) The invariant holistic features are extracted by using a publicly available pretrained DCNN, which is originally used in generic object classification. The DCNN gives rich and discriminative features, including color, texture, shape, and other visual cues to describe a human. We investigate the pretrained model, which is suitable to extract features for person re-id.
3) The regional features are extracted from partitioned body parts by exploiting Gaussian mixture models (GMMs) on the color histograms. Each Gaussian distribution of a GMM is exploited to represent a dominant color mode of a human subject. The proposed GMM-based feature matching method is applied on a two-way process, i.e., probe-to-gallery and gallery-to-probe, to compensate different poses of the same identity. A distance between two GMMs is computed by our proposed feature distance metric.
4) In the direct-method (without any specific data set training) category, our method has improved recognition rate over the state-of-the-art performance on the challenging public benchmark data sets, viewpoint invariant pedestrian recognition (VIPeR) [13], and 3DPeS [7] data sets.

This paper is organized as follows. Section II reviews some related research works. The holistic invariant features are proposed in Section III, while the regional invariant features are proposed in Section IV. The aggregating method is proposed in Section V and experimental results are shown in Section VI, followed by the conclusion in Section VII.

II. Related Works

In this section, we review the state-of-the-art person re-id approaches, which can be roughly divided into two main categories, the learning-based and the direct methods, whose main difference resides on with/without the requirement of the domain data training. The learning-based method uses the specific domain data to train the specific features, classifiers, or metrics. For each camera network environment, the learning-based method needs to collect many images associated with the same identity in different cameras to optimize the re-id performance. Instead of using any target-domain data, the direct method discovers the general and reliable visual descriptors to represent a person. The desired descriptor has the property that the intra-class (the same identity) pairs have more similar descriptors than the inter-class (different identities) pairs. The direct method aims to propose the invariant feature extraction and matching metric, which can handle the appearance variations caused by the different persons' pose, camera viewpoint, and environmental illumination change. The occlusion problem is another issue needed to be concerned; however, the existing method does not show the significant solution to this problem. The common solution is to segment one person image into many region parts and measure the similarity by accumulating each part's score to get the final decision.

The first category of approaches [14], [15] selects the domain-driven features or obtains specific models trained in advance by a data set separately to achieve high re-id accuracies if there are enough training data available and the testing environments are similar to the training environments. In [14], the generic descriptive statistical model and the discriminatively learned feature model are combined to attain better results. The generic descriptive statistical model can generate a rank list initially and the positive and negative training data can be chosen according to the rank list without any manual process for training the discriminating learned feature model. Gray and Tao [4] propose a viewpoint invariant model by integrating spatial and color information based on the boosting scheme. Prosser *et al.* [16] formulate the re-id problem as a ranking problem, where the Ensemble-Rank support vector machine (SVM) is used to learn a better subspace that the potential true match is given the highest ranking. Wang *et al.* [15] extract the 3D histogram of oriented gradient features from the discriminative video segments where the more reliable space-time features are derived to learn a video ranking function for re-id. Zhao *et al.* [6], [17] propose an unsupervised saliency learning framework for human matching. Some of the studies focus on metric learning to enhance the accuracies using both the similar pairs and dissimilar pairs of training data, e.g., large margin nearest neighbor [18], keep it simple and straightforward metric [19], and pairwise constrained component analysis [20]. In contrast to previously mentioned work that exploits hand-crafted features, some novel methods [21], [22] apply the deep convolutional network for learning the features and the similarity metric for person re-id simultaneously. In the existing person re-id data set, the image number is too less to train a deep convolutional network. To solve this problem, Li *et al.* [21] built the largest person re-id benchmark, CUHK03, which consists of 13 164 images of 1360 pedestrians. Based on these data sets, the deep convolution network structures with the paired images as input can thus be determined. In the supervised learning framework, each input pair has a binary label representing whether it is the same or different pair. The re-id problem is transformed to a two-class classification problem and the softmax loss minimization is performed in the top layer. When minimizing the loss, the stochastic gradient descent is adopted to update the weights and decrease the loss. Taking advantage of the end-to-end framework of the deep convolutional network, the discriminative and effective features and corresponding decision metric can be learned in the former layers and the top layer at the same time.

Although most learning-based methods achieve better performance than direct methods, it is worth noting that learning-based methods are strongly dependent on the training data sets and lacking generalization capabilities. Moreover, it is hard to obtain the samples with ground-truth label in the real scenarios where the conditions are dynamically changing, so they are not quite appropriate for practical use in the real-world surveillance applications. The second category of approaches, the direct methods [3], [5], [8], [23], directly runs on each person independently without the training process on a specific data set. These works mainly focus on proposing the novel and discriminative features. More specifically, an illumination-invariant color feature is proposed by Kviatkovsky *et al.* [5], where the signature is formed by the log-polar quantization in the log-chromatic color space. The symmetry-driven accumulation of local features (SDALF) by Farenzena *et al.* [3] divides the human body into head/torso/legs parts and extracts color features based on the horizontal and vertical asymmetries of the human silhouettes. Cheng *et al.* [23] use a pictorial structure (PS) to localize the parts, and then extract the part descriptor features to match objects. However, the performance of these approaches is quite unreliable, often subject to serious human pose and viewpoint issues. Recently, Baltieri *et al.* [8] propose a framework, which exploits a simplified nonarticulated 3D model to spatially map 2D appearance descriptors (color and gradient histograms) into the vertices of a regularly sampled 3D body model. This model effectively mitigates the problem of occlusion, partial views, and poses changes. However, the image-to-model mapping needs the perspective projection matrix (intrinsic parameters) and extrinsic calibration matrix to estimate the rotation and translation between the world reference coordinate and the model one, resulting in critical inconvenience to the practical use.

III. Holistic Pose-Invariant Features

The holistic features are desired to extract the meaningful information to describe the whole person. The information should include color, texture, shape, and other reliable visual cues. Recently, the DCNNs show its powerful capability to extract the rich and discriminative features from images [24], [25]. Based on the end-to-end structure of a DCNN, the local low-level features (e.g., specific orientation gradient on a person's shoulder) and the global high-level features (e.g., the heads and legs of pedestrians) can be trained and extracted hierarchically from the low-to-high layers [26] during the loss minimization.

Indeed, the DCNN is a supervised learning mechanism, which typically requires a lot of labeled training data to train model, i.e., learning the network weights parameters to minimize the objective loss. It is difficult and time-consuming to collect thousands of training data for the specific task or domain. Recently, domain transferring method has shown a significant boost performance for object classification [25], [27], [28]. When the labeled training data are scarce and insufficient to train complicated DCNNs, the domain-specific fine-tuning can transfer the discriminative features learned from a supervised pretraining DCNNs model to the target data set. Specifically, the transfer procedure can be achieved by the following two steps, pretraining on a large data set, e.g., ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [29], and fine-tuning on a smaller target data set. Although this approach shows a significant performance improvement, it still requires hundreds or thousands of target-domain data for performing the fine-tuning. Therefore, it is still very infeasible and unpractical to employ for the person re-id in real case.

Based on the most common scenario that no image can be obtained for training or fine-tuning the model, we want to discover a potential solution, using the existing pretrained DCNN models without domain fine-tuning. A large-scale DCNN is learned by great amount of image data, so the extracted features can inherently possess many different types of visual information. The rich features can thus enhance the invariance of different poses, lightings, and viewpoints. Therefore, many studies [25], [30], [31] for object detection, object segmentation, and object tracking have accommodated the DCNN features to achieve better results. Therefore, we evaluate the different features extracted from three popular large-scale DCNNs, Krizhevsky *et al.* AlexNet [24], Chatfield *et al.* visual geometry group (VGG) [32], and Szegedy *et al.* GoogLeNet [33]. All these networks are trained by the data of ILSVRC [29]. The ILSVRC 2014 classification challenge involves the task of classifying an image into one of 1000 leaf-node categories in the ImageNet hierarchy. There are about 1.2 million images for training, 50 000 for validation, and 100 000 images for testing. Each image has a label associated with its ground-truth category belonged to one of 1000 categories.

The classifier network is achieved by following the top feature extraction layer with an n-way, i.e., 1000, softmax layer. Given a training example x that produces n-dimensional feature map X_{soft} at the softmax layer, for each possible label

$i = 1, \ldots, n$, this layer computes probability distribution p_i over the n classes by

$$p_i = \frac{\exp(X_{\text{soft}}(i))}{\sum_{k=1}^{n} \exp(X_{\text{soft}}(k))}. \tag{1}$$

When training the model, it minimizes the cross-entropy loss, which we call the classification loss and is denoted as

$$loss_{\text{class}}(X_{\text{soft}}, y, \theta) = -\sum_{i=1}^{n} y_i \log p_i \tag{2}$$

where $y_i = 1$ if x has label i and 0 otherwise, and θ denotes the weight parameters of softmax layer. Note that this loss does not depend on just θ, because the computation of the feature map X_{soft} involves the weights of the early convolutional and fully connected layers.

To correctly classify all the classes simultaneously, the early layers will be updated to search the most discriminative and category-related features, i.e., features with large inter-category variations. In general, the higher layers can provide more global information with better discriminative power as holistic features [25], [26]. Therefore, we extract the activation of the top layer with/without employing the rectified linear unit (ReLU), as visual representations, which is 4096 dimensions features from FC7 (fully connected seventh layer) and FC7 + ReLU (ReLU7) of both AlexNet [24] and Chatfield *et al.* VGG [32], and 1024 dimensions of Pool5, which represents the Average-Pooling after inception 5 in GoogLeNet [33].

In order to evaluate the capability of representations extracted from different models or settings, we apply a specific distance measurement to compute the distances $d_{\text{DCNN}_k}(I_p(i), I_q(j))$, where k denotes the separate features extracted from the following DCNNs, {Alex FC7, Alex ReLU7, VGG FC7, VGG ReLU7, and Pool5 GoogLeNet}. Though the DCNN has pooling layers to eliminate the problem raised by person misalignment in the image, the performance can still easily be affected in the fine-grained recognition.

We thus propose to solve the misalignment problem by additionally generating ten sampled images with various geometric distortions before applying the DCNN. Given two images, the probe image I_p and the corresponding gallery image I_q, the additionally generated sampled images for both probe and gallery images are denoted as $I_p(i)$ and $I_q(j)$, where $1 \leq \{i, j\} \leq 10$, and the extracted DCNN feature vector for an image I is denoted as $f_{\text{DCNN}_k}(I)$. The holistic feature distance $d_{\text{DCNN}_k}(I_p, I_q)$ can be obtained by the minimum of the distance d_{metric} among all the sampled pairs. This data augmentation procedure is able to improve about 40% of accuracy in general

$$d_{\text{DCNN}_k}(I_p, I_q) = d_{metric}(f_{\text{DCNN}_k}(I_p(\hat{i})), f_{\text{DCNN}_k}(I_q(\hat{j}))) \tag{3}$$

where $\hat{i}, \hat{j} = \arg\min_{i,j} d_{metric}(f_{\text{DCNN}_k}(I_p(i)), f_{\text{DCNN}_k}(I_q(j)))$ and the $metric$ can be applied by L_1 and L_2, or Chi-squared distance metric. Fig. 3 shows an example of data augmentation procedure.

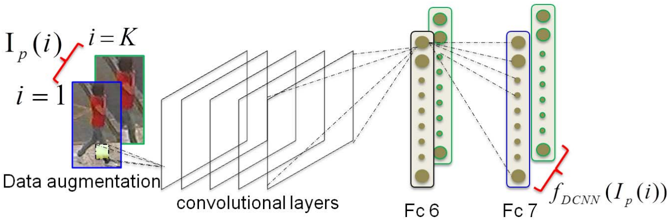


Fig. 3. Example to show feature from the AlexNet [24]. Data augmentation generates ten sampled images with various geometric distortions before applying the DCNN. The holistic features extracted the high-level representation from the top layers of the pretrained DCNNs.

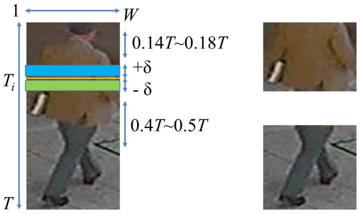


Fig. 4. Sample body partition using (4), where the left-hand side image is an input and the right-hand side images are partitioned parts.

IV. Regional Invariant Features

We propose regional invariant features based on color appearance for person re-id. Most persons wear two separate clothes for top and bottom bodies, e.g., a shirt and a pair of pants, which have several distinctive dominant colors. To extract discriminative information, we derive features after dividing a human body into three parts, head, torso, and legs. Even if the person wears a single piece dress or a long coat, this division can still be applicable.

A. Body Partition

We assume that bounding boxes and silhouettes of the person in our experiments are acquired and normalized to a fixed template size. Both 3DPeS and VIPeR data sets, which are used in evaluating our proposed algorithm, furnish the pedestrian images and the corresponding paired silhouette masks. Background subtraction is exploited to extract the foreground pixels. In general, a person re-id system, which utilizes color signature, subdivides the human body into salient parts in order to minimize the effects of mixing colors from different clothing articles [3], [5]. Since a pedestrian is commonly acquired at very low resolution, it is reasonable to notice that the most distinguishable body parts are three: head, torso, and legs [3]. Two boundary lines are taken on both head–torso and torso–legs, respectively. Fig. 4 shows an example of the body partition. From the top to the bottom of the rectangular foreground bounding box, we calculate the histogram distance line-by-line between two stripe regions, i.e., the blue stripe and green stripe regions. Each stripe is of the height δ and the width W. Intuitively, we expect that color similarity between two different body parts to be low. Therefore, a boundary line is located at height T_i, which is computed by solving the following problem, for both head–torso and torso–legs

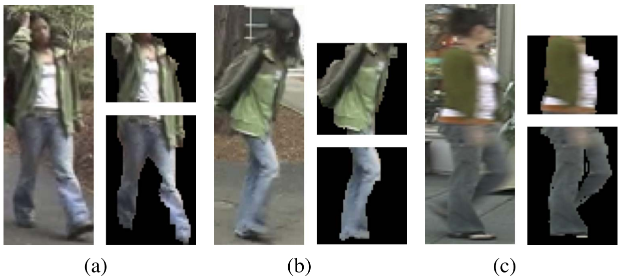


Fig. 5. Examples of the VIPeR data set. (a) Probe (ID 57). (b) Gallery (ID 57). (c) Gallery (ID 554).

TABLE I
DISTANCE IN FIG. 5 WITH HIST, EMD, KL, AND 2WGMMF

Method	$d\left(\mathbf{h}^{(a)}_{part}, \mathbf{h}^{(b)}_{part}\right)$ Total (torso, legs)	$d\left(\mathbf{h}^{(a)}_{part}, \mathbf{h}^{(c)}_{part}\right)$ Total (torso, legs)
Hist	1.29 (0.56, 0.73)	1.07 (0.55, 0.52)
EMD	18.74 (9.12, 9.62)	14.76 (5.99, 8.77)
KL	4.12 (2.14, 1.98)	4.00 (0.27, 3.73)
2WGMMF(1D)	1.03 (0.60, 0.43)	1.33 (0.67, 0.66)
2WGMMF(3D)	33432 (16285, 17147)	72103 (26063, 46040)

regions, respectively:

$$\max_{T_i \in \{A,B\}} d_{\text{Chisq}}(\mathbf{h}_{[T_i, T_i+\delta]}, \mathbf{h}_{[T_i-\delta, T_i]}) \tag{4}$$

where $d_{\text{Chisq}}(\cdot)$ denotes a function that computes Chi-squared distance and $\mathbf{h}_{[a,b]}$ denotes the color histogram derived from the stripes of a to b. Moreover, the boundary line is assumed to be located within $\{A, B\}$, e.g., $\{0.14T, 0.18T\}$ for head–torso and $\{0.4T, 0.5T\}$ for torso–legs. After partitioning the whole body into three parts, we extract features from the torso and the legs parts and discard the information of the head/face region, because standard biometric algorithms usually fail at low resolution [3]. Eight-bin hue, saturation, value (HSV) histogram is employed and the height δ value is empirically set to 5 pixels, when the entire segmented foreground is normalized with the height of $T = 128$ pixels.

B. 2WGMMF Features

The person re-id problem can be regarded as a task of discriminative feature design to decrease the distance between two snapshots of the same identity. The most difficult challenge is to handle the large variations of poses and viewpoints in feature extraction. Since a human body is a 3D object in real, some body parts can be occluded by other parts of the same body and the unseen parts in a probe image can be visible in a gallery image when their poses or camera viewpoints are changed. As shown in Fig. 5(a) and (b) is the same person, however, color appearances of top body are not similar with respect to the existing color representation due to viewpoint is changed. The existing color-based signatures, such as color histogram, earth movers distance (EMD), and GMM method [34], are not useful to deal pose variations across disjoint camera views. On the other hand, human eyes can recognize them as the same identity based on small distinctive regions, e.g., the same jacket. In this section, we propose a new feature utilizing these human perception characteristics to solve this problem. Human identifies the same person of different poses through matching dominant color compositions, which sometimes occupy only a small portion of body [6]. More specifically, pose-invariant color feature is proposed to describe the dominant color information of each identity by using GMMs. To calculate a distance between two GMMs, we do not use conventional distance metric, such as Kullback–Leibler (KL) divergence [35], but propose a new fitting model.

1) 2WGMMF With Concatenated Histogram: In Fig. 5, there are three pairs of unmasked and masked images. Masked images on the right-hand side of each pair are the results of Section IV-A, i.e., torso and legs parts. Fig. 5(a) and (b) is the same person with ID 57 on the VIPeR data set, captured in different views, and Fig. 5(c) is another person with ID 554. Although Fig. 5(a) and (b) are the same person, and Fig. 5(c) is a different person, torso parts of Fig. 5(a) and (c) look more similar with respect to color composition, since the torso part mainly shows a shirt and a jacket in Fig. 5(a), but only jacket is visible in Fig. 5(b). Our goal is to match Fig. 5(a) and (b) as the same identity with a color appearance model. In Table I, we show the performance of several approaches in Fig. 5. The Chi-squared distance of 32-bin red, green and blue (RGB) histogram is 1.29 between Fig. 5(a) and (b), and 1.07 between Fig. 5(a) and (c). The EMD distance also shows that interpersonal variation [Fig. 5(a)–(c)] is smaller than intra-personal variation [Fig. 5(a) and (b)]. Intuitively, to solve this problem with color features, we need to extract main color modes from each image and utilize them to identify a correct pair of images. In case of Fig. 5(a), the main color mode of the jacket region should be exploited to match with Fig. 5(b) and the shirt region's color mode needs to be ignored. For extracting key information, [6] and [17] tried to find salient patches and [23] proposed to adopt PSs. However, the saliency matching method needs to build dense correspondence in advance with additional training data sets, which have to be a part of the target data set. In PS, decomposing a body into six parts, i.e., chest, head, thighs, and legs, is not easy, because pedestrian image, as captured by a surveillance camera, is normally low resolution and too small to detect each body part exactly. Even though body parts are partitioned perfectly, we should still deal with mixed color components, i.e., a shirt region merged with a jacket in the torso part of Fig. 5(a). Naturally, color histogram includes main color modes and the other color components of body parts as well. Instead of cropping the salient part in an image domain, we propose to find main color modes in color histogram domain. We contribute to extract main color modes among several color modes by using GMM and fitting model. In [34], GMM is used as a parametric technique for representing the color distribution in a person's clothing for person re-id. Furthermore, the number of Gaussians is fixed as only one to fit Gaussian distribution to the top and the bottom parts of the target, respectively, and calculate the distance simply between the overall appearance descriptions using the KL divergence [35]. On the contrary, our proposed method systematically determines the number of Gaussian components and successfully distinguishes main color modes from the others to achieve the re-id.

We construct a color histogram from torso and legs parts of the probe and the gallery images separately as

$$\mathbf{H}^p = \left[\mathbf{h}^p_{\text{torso}}, \mathbf{h}^p_{\text{legs}}\right], \quad \mathbf{H}^g = \left[\mathbf{h}^g_{\text{torso}}, \mathbf{h}^g_{\text{legs}}\right] \tag{5}$$

where $\mathbf{h}$ denotes a concatenated histogram, which contains m bins and n channels, $\mathbf{h} \in \mathbb{R}^{mn\times 1}$. To build color appearance model of an image, GMM is employed, because we believe that dominant color modes can be universally represented as a mixture of Gaussian components in the color histogram. Thus, we have to estimate Gaussian parameters from the color histogram. We assume that each histogram, $\mathbf{h}$, in $\mathbf{H}$ is a mixture of K Gaussian probability density functions with heteroscedastic components, that is

$$p(\mathbf{h}|\theta_1, \dots, \theta_K) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{6}$$

where $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}$ denotes the set of parameters for component k, π_k denotes the mixing proportion where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k$ denotes the mean vector for component k, $\boldsymbol{\Sigma}_k$ denotes the covariance matrix, and $\mathcal{N}(\cdot)$ denotes the Gaussian distribution. K denotes the number of Gaussian components, i.e., the number of dominant color modes.

In general, GMM tries to find θ_k that maximize $p(\mathbf{h})$; this is equivalent to minimizing the negative log-likelihood function

$$-\ln p(\mathbf{h}) = -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right). \tag{7}$$

Then, color appearance model can be formulated as

$$\min_{\boldsymbol{\mu},\boldsymbol{\Sigma},\pi} -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

$$\text{s.t. } \pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1. \tag{8}$$

This model fitting problem can be efficiently solved via the expectation–maximization algorithm [36]. In the finite mixture models, determining the number of components K is an important but very critical problem, which has not been completely resolved. Akaike information criterion and Bayesian information criterion (BIC) have been commonly used for choosing the number of components for a suitable density estimation [36]. Both of them are effective measures of the relative quality of statistical models for a given set of data with the incorporation of a penalty term to discourage overfitting. Since the density estimate that uses BIC to select the number of components in the mixture has performed very well in previous simulation studies [37]–[39], BIC is employed in this paper. The preferred model is the one with the minimum BIC value, $-2\log\mathcal{L}(\theta) + K\log n$, where $\mathcal{L}(\theta)$ denotes the maximized value of likelihood function, which is equal to the inverse of final prediction error for the estimated model, K is the total number of parameters, and n is the sample size [36].

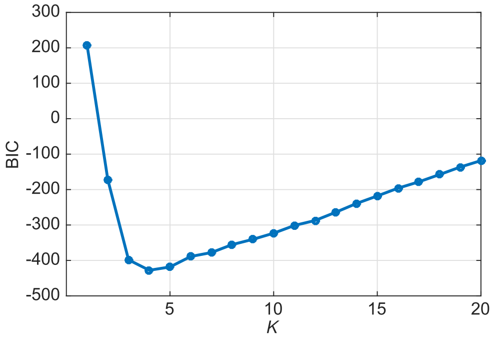
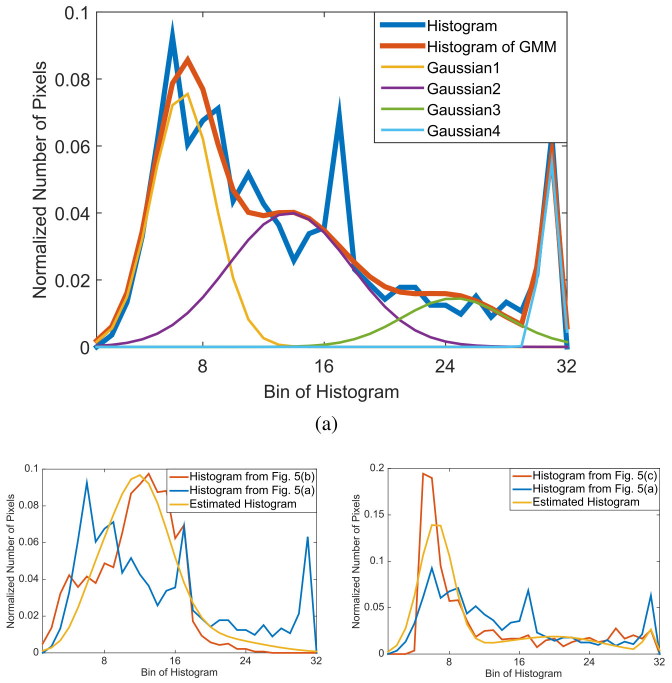


Fig. 6. BIC curve for GMM component estimation of the torso in Fig. 5(a).



(a)

(b)

(c)

Fig. 7. Example of the 2WGMMF application. (a) Result of (8) with Fig. 5(a). (b) Result of (9) with Fig. 5(b). (c) Result of (9) with Fig. 5(c).

In Fig. 6, BIC score is shown with respect to K from 1 to 20 and the preferred model is chosen when K equals 4 in case of the torso in Fig. 5(a). Fig. 7(a) shows the result of (8), where the blue curve is the color histogram of Fig. 5(a), the red curve is the histogram of GMM, and the other curves are the components of GMM on the blue channel.

After modeling GMMs from a torso and a legs part of both a probe and a gallery image, respectively, we utilize chosen models to measure the difference between each part. The mean vectors and covariance matrices associated with the Gaussian components of the GMM are exploited to evaluate mixing weights, $\boldsymbol{\pi}$, by solving the least-squares curve fitting problem as

$$\hat{\boldsymbol{\pi}} = \arg\min_{\boldsymbol{\pi}} \left\| G\left(\mathbf{h}_i^p\right) \cdot \boldsymbol{\pi} - \mathbf{h}_i^g \right\|_2^2 \quad \text{s.t. } 0 \leq \pi_k,\ k = 1, \dots, K \tag{9}$$

where $\mathbf{h}_i$ is the normalized color histogram and $G(\mathbf{h}_i^p)$ denotes a GMM of a probe's ith part and is defined as

$G(\mathbf{h}_i^p) = [\mathcal{N}(\mathbf{h}_i^p|\mu_1, \Sigma_1), \dots, \mathcal{N}(\mathbf{h}_i^p|\mu_K, \Sigma_K)]$. Note that $\mathcal{N}(\mathbf{h}|\mu_k, \Sigma_k) \in \mathbb{R}^{mn\times 1}$ for $i = \{\text{torso}, \text{legs}\}$. The mixing weights are $\boldsymbol{\pi} \in \mathbb{R}^{K\times 1}$ and $||\cdot||_2^2$ denotes the squared L_2 norm. Equation (9) is a formulation of applying the GMM of a gallery on a probe. In this step, a different pose of the same identity can be compensated with changeable mixing proportion, $\boldsymbol{\pi}$. In Fig. 7(b) and (c), the red curves denote the 32-bin blue channel histogram of the torso part of Fig. 5(b) and (c), respectively. The yellow curves show the fitting result of (9) with the GMM components in Fig. 7(a). In other words, blue curves are transformed to yellow curves by fitting GMM in (9). More specifically, we find that the second component, purple curve, in Fig. 7(a) increases its scale and the fourth component, sky blue curve, in Fig. 7(a) decreases to 0 in Fig. 7(b). The residual calculated based on (10) decreases large enough for us to conclude that these are the same person. In contrast, the first component, yellow curve, in Fig. 7(a) increases its scale π_1 and the second component, purple curve, in Fig. 7(a) decreases its scale π_2 in Fig. 7(c). Thus, we can conclude that the second peak represents the unique jacket region of ID 57 and the fourth peak represents the shirt region. The residual between a target color histogram and a summed distribution of GMM is computed as a distance

$$r(\mathbf{h}_i^g) = d_{\text{Chisq}}\big(G\big(\mathbf{h}_i^p\big)\cdot\hat{\boldsymbol{\pi}}, \mathbf{h}_i^g\big). \tag{10}$$

We also need to proceed (8)–(10) on the Gaussian mixture components of Fig. 5(b) and (c) to (a). That is why this feature is named as two-way GMM fitting (2WGMMF), since we apply the same method on both ways, i.e., probe-to-gallery and gallery-to-probe. The sum of residual is used as the final distance for this probe–gallery pair as

$$d_{\text{2WGMMF}}^{\text{concatenate}}(\mathbf{H}^p, \mathbf{H}^g) = r\big(\mathbf{h}_{\text{torso}}^p\big) + r\big(\mathbf{h}_{\text{legs}}^p\big) + r\big(\mathbf{h}_{\text{torso}}^g\big) + r\big(\mathbf{h}_{\text{legs}}^g\big). \tag{11}$$

By using 2WGMMF, the distance between Fig. 5(a) and (b) is 1.03 and the distance between Fig. 5(a) and (c) is 1.33, while KL method gives 4.12 between Fig. 5(a) and (b), and 4.00 between Fig. 5(a) and (c) in Table I. When we consider d_{2WGMMF} to be a two-way distance, both distances decrease, because 2WGMMF utilizes the Gaussian mixture components on histogram fitting, resulting in the better matching of jackets with dominant color components. Note that intra-personal variation becomes smaller than inter personal's through 2WGMMF. Algorithm 1 summarizes the complete 2WGMMF procedure with concatenated histograms.

```
Algorithm 1 2WGMMF With Concatenated Histogram
Input: a probe and a gallery image with silhouette masks
Output: a feature distance
1: Construct the concatenated color histogram
       $\mathbf{H}^p = \big[\mathbf{h}^p_{\text{torso}}, \mathbf{h}^p_{\text{legs}}\big], \quad \mathbf{H}^g = \big[\mathbf{h}^g_{\text{torso}}, \mathbf{h}^g_{\text{legs}}\big]$
2: Solve the problem of model fit:
       $\min_{\boldsymbol{\mu},\boldsymbol{\Sigma},\pi} -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$
       $s.t.\ \ \pi_k \geq 0,\ \sum_{k=1}^{K} \pi_k = 1.$
3: Solve the least-squares curve fitting problem:
       $\min_{\boldsymbol{\pi}} \left\| G(\mathbf{h}_i^p)\cdot\boldsymbol{\pi} - \mathbf{h}_i^g \right\|_2^2$
       $s.t.\ 0 \leq \pi_k,\ k = 1, \dots, K$
4: Compute the residuals
       $r(\mathbf{h}_i^g) = d_{\text{Chisq}}(G(\mathbf{h}_i^p)\cdot\hat{\boldsymbol{\pi}}, \mathbf{h}_i^g)$
5: Compute the feature distance
       $d_{\text{2WGMMF}}^{\text{concatenated}}(\mathbf{H}^p, \mathbf{H}^g) = r(\mathbf{h}^p_{\text{torso}}) + r(\mathbf{h}^p_{\text{legs}}) + r(\mathbf{h}^g_{\text{torso}}) + r(\mathbf{h}^g_{\text{legs}})$
```

2) 2WGMMF With Joint Histogram: Unlike the concatenated color histogram (we refer to this as 1D histogram), which is constructed and exploited to build GMM in Section IV-B1, we also generate 3D joint color histogram (we refer to this as 3D histogram), which takes three channels jointly. For example, RGB 3D histogram takes red, green, and blue channel jointly, so it can offer unique and distinct information from 1D histogram. The procedure of feature extraction is similar to Algorithm 1. The difference is the dimension of color histogram $\mathbf{h} \in \mathbb{R}^{m^n}$, instead of $m \times n$, where m denotes the number of bins and n denotes the number of channels. The method for calculating distance is different as well. To compute the difference between joint color histograms of a probe and a GMM of gallery, we utilize negative log likelihood as

$$\begin{aligned} d_{NL}\big(\mathbf{h}_i^g, G\big(\mathbf{h}_i^p\big)\big) &= -\ln p\big(\mathbf{h}_i^g|\theta_1^p, \dots, \theta_K^p\big) \\ &= -\ln\left(\sum_{k=1}^{K} \pi_k^p \mathcal{N}\big(\mathbf{h}_i^g|\boldsymbol{\mu}_k^p, \Sigma_k^p\big)\right) \end{aligned} \tag{12}$$

where $G(\cdot)$ denotes GMM from inside color histogram and the parameters of GMM, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and π_k, are obtained from color histogram of a part of a gallery where $i = \{\text{torso}, \text{legs}\}$. Equation (12) computes the likelihood function of how much a GMM of a probe in response to the data of a gallery. The result from (12) is regarded as one-way distance of i-part and a small value indicates that they are likely to be the same identity. The final distance can be represented as

$$\begin{aligned} &d_{\text{2WGMMF}}^{\text{joint}}(\mathbf{H}^p, \mathbf{H}^g) \\ &\quad = d_{NL}\big(\mathbf{h}^p_{\text{torso}}, G\big(\mathbf{h}^g_{\text{torso}}\big)\big) + d_{NL}\big(\mathbf{h}^p_{\text{legs}}, G\big(\mathbf{h}^g_{\text{legs}}\big)\big) \\ &\qquad + d_{NL}\big(\mathbf{h}^g_{\text{torso}}, G\big(\mathbf{h}^p_{\text{torso}}\big)\big) + d_{NL}\big(\mathbf{h}^g_{\text{legs}}, G\big(\mathbf{h}^p_{\text{legs}}\big)\big). \end{aligned} \tag{13}$$

The complete 2WGMMF procedure with joint histogram is summarized in Algorithm 2.

V. Aggregation of Features

We propose mainly three representations to describe person appearance, a holistic signature based on the DCNN, regional color with concatenate histogram (2WGMMF$_{\text{1D}}$), and joint histogram (2WGMMF$_{\text{3D}}$). To perform the person re-id, we should combine these three features effectively. Since the dynamic ranges of three distances are quite different (see Table I), it is not so meaningful to add them directly. Some form of normalization is necessary to balance the contributions of individual features.

In case of d_{2WGMMF}, it can integrate several kinds of color space, e.g., RGB, HSV, and YCbCr. When more than two color spaces are exploited, we collect them according to

Algorithm 2 2WGMMF With Joint Histogram

```
Input: a probe and a gallery image with silhouette masks
Output: a feature distance
Construct the joint color histogram
    $\mathbf{H}^p = \left[\mathbf{h}^p_{\text{torso}}, \mathbf{h}^p_{\text{legs}}\right], \quad \mathbf{H}^g = \left[\mathbf{h}^g_{\text{torso}}, \mathbf{h}^g_{\text{legs}}\right]$
Solve the problem of model fit:
    $\min_{\boldsymbol{\mu},\boldsymbol{\Sigma},\pi} -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{h}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$
    $s.t. \;\; \pi_k \geq 0, \; \sum_{k=1}^{K} \pi_k = 1.$
Compute the negative log likelihood
    $d_{NL}(\mathbf{h}^g_i, G(\mathbf{h}^p_i)) = -\ln\left(\sum_{k=1}^{K} \pi^p_k \mathcal{N}\left(\mathbf{h}^g_i|\boldsymbol{\mu}^p_k, \Sigma^p_k\right)\right)$
Compute the feature distance
    $d^{\text{joint}}_{\text{2WGMMF}}(\mathbf{H}^p, \mathbf{H}^g)$
        $= d_{NL}\left(\mathbf{h}^p_{\text{torso}}, G(\mathbf{h}^g_{\text{torso}})\right) + d_{NL}\left(\mathbf{h}^p_{\text{legs}}, G(\mathbf{h}^g_{\text{legs}})\right)$
        $+ d_{NL}\left(\mathbf{h}^g_{\text{torso}}, G(\mathbf{h}^p_{\text{torso}})\right) + d_{NL}\left(\mathbf{h}^g_{\text{legs}}, G(\mathbf{h}^p_{\text{legs}})\right)$
```

their dimensions. For example, RGB, HSV, and YCbCr color spaces are used in 2WGMMF, they are combined as

$$\begin{aligned} d^{\text{1D}}_{\text{2WGMMF}} &= d^{\text{1D RGB}}_{\text{2WGMMF}} + d^{\text{1D HSV}}_{\text{2WGMMF}} + d^{\text{1D YCbCr}}_{\text{2WGMMF}} \\ d^{\text{3D}}_{\text{2WGMMF}} &= d^{\text{3D RGB}}_{\text{2WGMMF}} + d^{\text{3D HSV}}_{\text{2WGMMF}} + d^{\text{3D YCbCr}}_{\text{2WGMMF}}. \end{aligned} \tag{14}$$

Aggregating similarity scores promises more convincing result than minimizing accumulated distances [40], [41]. Thus, the distances are converted into the similarity as in inverse proportion, respectively

$$sim\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = 1/d\left(\mathbf{H}^p, \mathbf{H}^g_j\right) \quad \text{for } j = 1, \ldots, N \tag{15}$$

where N denotes the size of gallery set. In case of 2WGMMF with joint histogram, their feature distance can be negative, because $d^{\text{joint}}_{\text{2WGMMF}}$ denotes the sum of negative log likelihood. To make them positive, we add minimum value of them before converting into similarity scores.

In order to balance the contributions of separate features, min–max normalization is performed to transform the feature distances into 0-to-1

$$\widehat{sim}\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = \frac{sim\left(\mathbf{H}^p, \mathbf{H}^g_j\right) - \min S}{\max S - \min S} \tag{16}$$

where $S = \{sim(\mathbf{H}^p, \mathbf{H}^g_1), \ldots, sim(\mathbf{H}^p, \mathbf{H}^g_N)\}$.

To stick to direct methods and to show the intrinsic quality of the proposed descriptors, we have set equal weights as

$$sim_{\text{Final}}\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = \frac{1}{3}\sum_{n=1}^{3} \widehat{sim}_n\left(\mathbf{H}^p, \mathbf{H}^g_j\right) \tag{17}$$

where $j = 1, \ldots, N$. Finally, we choose the identity of a probe image to be the one with maximum similarity among N gallery images as

$$identity(\mathbf{H}^p) = \arg\max_j S_{\text{Final}} \tag{18}$$

where $S_{\text{Final}} = \{sim_{\text{Final}}(\mathbf{H}^p, \mathbf{H}^g_j)\}$ for $j = 1, \ldots, N$. Algorithm 3 shows all the steps of classifying the same identity of a probe in the gallery based on calculating the distance scores.

Algorithm 3 Person Reidentification With Aggregation of Distance Scores

```
Input: distance scores ($d_{\text{DCNN}}, d^{\text{1D}}_{\text{2WGMMF}}, d^{\text{3D}}_{\text{2WGMMF}}$)
Output: identity
Add the distances of the same dimensional 2WGMMF feature together
    $d^{n\text{D}}_{\text{2WGMMF}} = d^{n\text{D color}_1}_{\text{2WGMMF}} + d^{n\text{D color}_2}_{\text{2WGMMF}} + \ldots + d^{n\text{D color}_\text{M}}_{\text{2WGMMF}}$
Transform the distance into the similarity:
    $sim\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = 1/d\left(\mathbf{H}^p, \mathbf{H}^g_j\right)$ for $j = 1, \ldots, N$
    where $N$ denotes the size of gallery set.
Min-max normalization
    $\widehat{sim}\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = \frac{sim\left(\mathbf{H}^p, \mathbf{H}^g_j\right) - \min S}{\max S - \min S}$
    where $S = \left\{sim\left(\mathbf{H}^q, \mathbf{H}^g_1\right), \ldots, sim\left(\mathbf{H}^p, \mathbf{H}^g_N\right)\right\}$
Compute the final similarity
    $sim_{\text{Final}}\left(\mathbf{H}^p, \mathbf{H}^g_j\right) = \frac{1}{3}\sum_{n=1}^{3} \widehat{sim}_n\left(\mathbf{H}^p, \mathbf{H}^g_j\right)$
    for $j = 1, \ldots, N$
Classify the same identity
    $identity(\mathbf{H}^p) = \arg\max_j S_{\text{Final}}$
    where $S_{\text{Final}} = \left\{sim_{\text{Final}}\left(\mathbf{H}^p, \mathbf{H}^g_j\right)\right\}$ for $j = 1, \ldots, N$
```

VI. Experimental Results

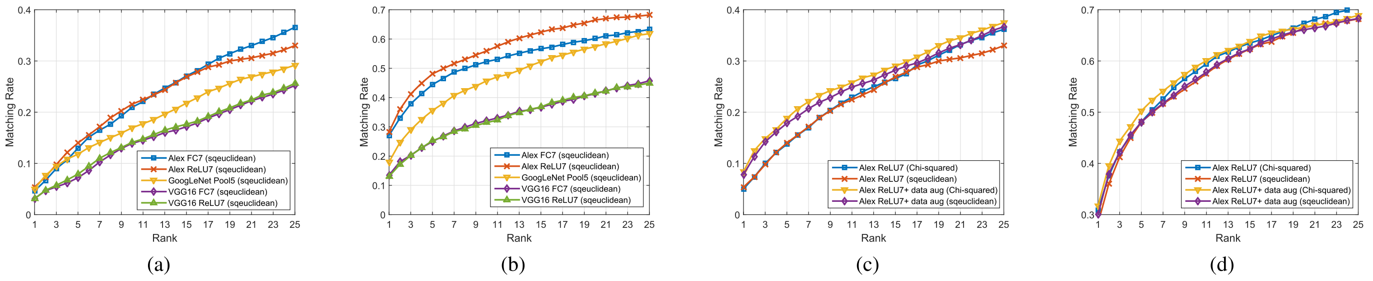
This section presents the evaluation results of our approach on the challenging benchmark data sets, VIPeR[1] [13] and 3DPeS[2] [7], which are video surveillance data sets designed for person re-id between multiple cameras with nonoverlapping field of views. We follow the same cross validation protocol as reported in [3] and [8], where experiments are performed with ten different random sets for probe and gallery sets. There is no overlapped image between the probe and gallery sets. The results are obtained by averaging ten corresponding outcomes. The performances are represented using the cumulative matching characteristic (CMC) curve and the normalized area under curve (nAUC). CMC curve represents where the rank-k recognition rate is the expectation of correct matches within the top k ranks. The nAUC is the area under the CMC curve and is exploited to summarize the overall performance. The higher nAUC indicates the better performance.

A. Analysis of Individual Features

For the pretrained DCNNs, we apply three popular large-scale convolutional neural networks: AlexNet [24], VGG [32], and GoogLeNet [33]. Since the higher layers can provide more global information with better discriminative powers [25], [26], we select activations of top layer and compare the performance with/without employing the unit rectifier, ReLU, for the AlexNet [24] and VGG [32]. As summarized in

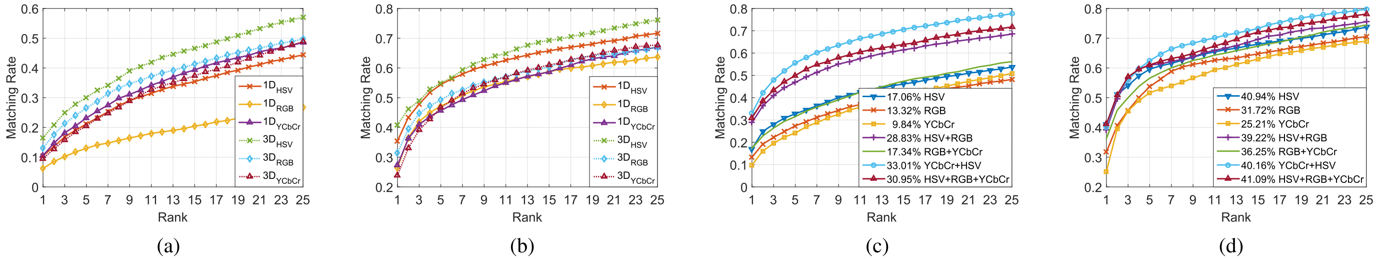
[1] VIPeR data set is available at https://vision.soe.ucsc.edu/node/178

[2] 3DPeS data set is available at http://www.openvisor.org/3dpes.asp



(a) (b) (c) (d)

Fig. 8. DCNN comparisons on (a) VIPeR and (b) 3DPeS data sets. The features from ReLU7 layer in AlexNet [24] have a better discriminative power than the other pretrained net models. The data augmentation with Chi-squared distance metric can improve about 4% on rank-1 accuracy.



(a) (b) (c) (d)

Fig. 9. (a) and (b) Comparison of individual 2WGMMF feature with a various of color spaces and dimensions. (c) and (d) Comparison of color space with 2WGMMF feature (marked values are rank-1 accuracies).

TABLE II

Selected Pretrain Models and Layer for Comparison

Selected Layer	Feature Dimension
AlexNet FC7 [24]	4,096
AlexNet ReLU7 [24]	4,096
VGG FC7 [32]	4,096
VGG ReLU7 [32]	4,096
GoogLeNet Pool5 [33]	1,024

Table II, we evaluate the features 4096 dimensional features from FC7 and ReLU7 of both AlexNet [24] and VGG [32], and 1024 dimensional features from Pool5 representing the Average-Pooling after inception 5 in GoogLeNet [33]. To compare the performance of different features, we exploit (3) based on the same squared L_2 distance ($metric =$ sqeuclidean) metric to find the best matched r instances (top-rank r) in the gallery set. The CMC curves for VIPeR and 3DPeS data sets are shown in Fig. 8(a) and (b), and the performance using features from ReLU7 in AlexNet [24] outperforms than other forms of VGG [32] and GoogLeNet [33]. Furthermore, based on the promising features of the ReLU7 in AlexNet [24], Fig. 8(c) and (d) show that the data augmentation allows to improve about 4% on rank-1 accuracy and it performs better using Chi-squared distance than the squared L_2 or L_1 distance on both VIPeR and 3DPeS data sets. Therefore, we decide to adopt the feature vector of AlexNet ReLU7 with Chi-squared distance metric to generate the holistic invariant features. This holistic invariant features will combine with the regional invariant features for further comparison.

To investigate the characteristics of 2WGMMF feature, several kinds of experiments are performed and analyzed. In Fig. 9, we show the CMC curve of several 2WGMMF features on the two data sets. In particular, Fig. 9(a) and (b) show the performances of individual 2WGMMF features with varied color spaces and dimensions. Fig. 9(c) and (d) shows the results of 2WGMMF features with both concatenated and joint color histograms on the marked color space. On both data sets, 2WGMMF features with HSV joint histogram perform the best consistently, as shown in Fig. 9(a) and (b). However, the other results are not consistent. In Fig. 9(c) and (d), the best color space is not the same with respect to rank-1 accuracy. 2WGMMF features with the combination of YCbCr and HSV color spaces outperform the others on the VIPeR. In contrast, when three color spaces are combined, rank-1 accuracy is the highest on the 3DPeS in Fig. 9(d). To sum up, the performance of features on each color space and dimension is not consistent on two different data sets, because 2WGMMF features are designed to deal with varied viewpoints and poses, and three kinds of color spaces are not illumination-invariant. Thus, we expect that the performance can be improved further when 2WGMMF features are integrated with invariant color space.

B. VIPeR Data Set

The VIPeR data set contains 632 pedestrian image pairs taken from two different cameras. Most of the image pairs contain a viewpoint change of 90° and are taken under varying illumination conditions. Each pair is randomly split into two sets. Images from one set are considered as the probe set and images from the other set are regarded as gallery set. All the images of VIPeR are scaled to the size of 128 × 48 pixels for experiments. For each image on the data set, a silhouette mask is automatically extracted by using the structure element model [43] and provided in [3].

To evaluate the performance, we compare the proposed features with the state-of-the-art methods [3]–[5], [12], [42], [44]. Farenzena *et al.* [3] divide a human body into head/torso/legs based on the horizontal and vertical asymmetries of a human

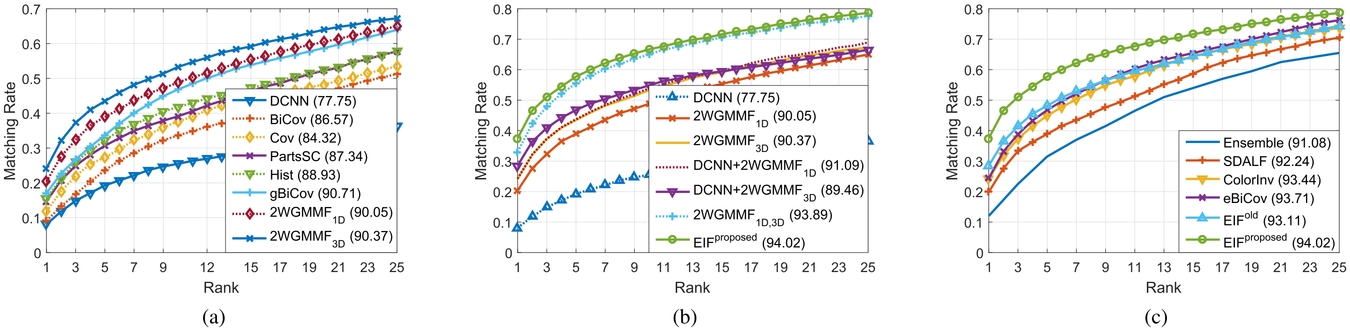


Fig. 10. VIPeR data set. (a) Comparison of the state-of-the-art single set of features. (b) Comparison of DCNN, 2WGMMF, and their combinations. (c) Comparison to Ensemble [4], SDALF [3], ColorInv [5], eBiCov [42], and EIF^{old} [12].

silhouette. They extract three complementary descriptors, including weighted HSV histogram, the spatial arrangement of colors into stable regions, maximally stable color regions (MSCRs), and recurrent high-structured patches. Similar to our proposed scheme, the Ensemble method by Gray and Tao [4] emphasizes the viewpoint invariant features. They propose an ensemble of color-based features (RGB, HSV, and YCbCr histograms) and gradient-based features (the histogram generated by Gabor filter response). The ensemble features are computed for each of the three fixed size stripes of a person's silhouette. For fair comparison, a distance between feature vectors is computed through the distance metric between the descriptors rather than applying the additional metric learning as proposed in [4]. Kviatkovsky *et al.* [5] propose illumination-invariant color features. They combine their features with several standard signatures, the parts shape context descriptor (PartsSC), color histogram (Hist), and the region covariance descriptor (Cov). BiCov [44] is a representation relied on the combination of biologically inspired features (BIFs) and covariance descriptors used to compute the similarity of the BIF features at neighboring scale. gBiCov [42] is the extended work of BiCov, combining covariance similarity with BIF. Furthermore, Ma *et al.* [42] propose eBiCov (enriched gBiCov) by combining gBiCov with weighted color histograms and MSCR defined in [3]. In [12], we proposed EIF by combining 2WGMMF with concatenated histogram, DCNN, and completed local binary pattern, which represents texture information.

In Fig. 10(a), the proposed features are compared with BiCov, Cov, PartsSC, Hist, and gBiCov for analyzing the sensitivity of the single set of features. These results are obtained from authors' reported materials [45], [46]. The 2WGMMF features with joint histogram show the best performance among the state-of-the-art single set of features both on CMC curve and nAUC, as marked in the legend. On the VIPeR data set, we exploit HSV and YCbCr for 2WGMMF features, since they show the best performance with respect to rank-1 accuracy [see Fig. 9(c)]. In Fig. 10(b), we compared DCNN, 2WGMMF, and their combinations. Among all combination of features, 2WGMMF features with concatenated and joint histogram outperform all others. In Fig. 10(c), the proposed method is compared with the state of the art, eBiCov, ColorInv, SDALF, Ensemble, and EIF^{old}. In case of ColorInv, we adopt the best signature that combines PartsSC, Hist, and Cov features together as mentioned in [5]. Our proposed EIF achieves 37.47% at rank-1, which is 13% higher than eBiCov [42] and 9% higher than the old EIF [12]. EIF^{old} has texture information, which is not effective in case of viewpoint change. In contrast, EIF^{proposed} exploits joint histogram in 2WGMMF, which is the most accurate single set of features [see Fig. 10(a)].

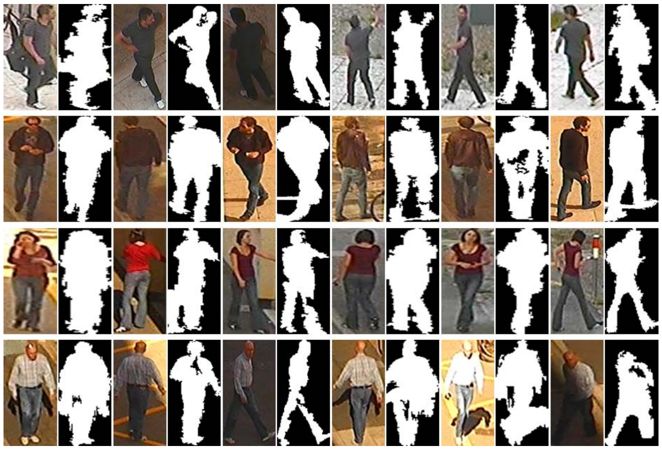
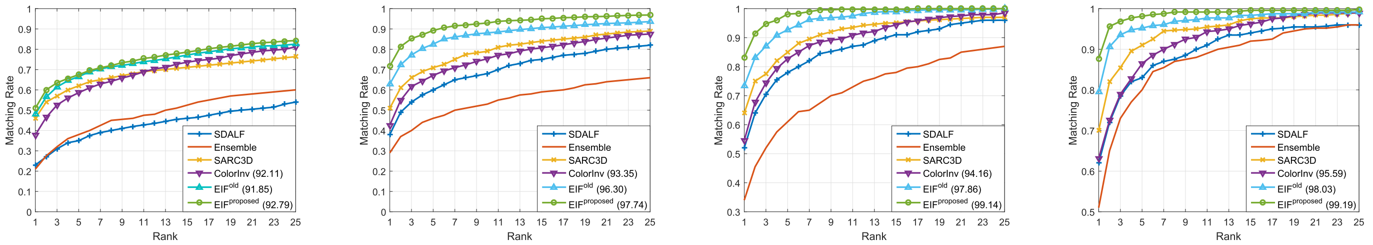


Fig. 11. Image and mask examples from 3DPeS data set.

C. 3DPeS Data Set

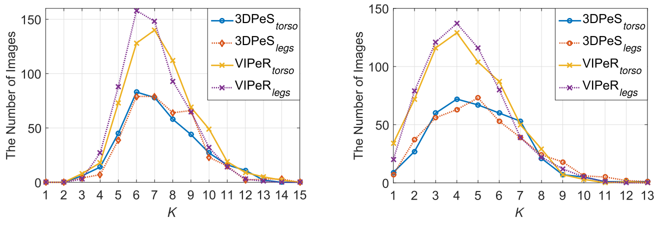
The 3DPeS data set contains short video sequences instead of still images, which increase the diversity of person poses and camera viewpoints. A collection of snapshots (six different shots) are collected for four different persons (one person per row), as shown in Fig. 11. The complete data set is composed of 1012 snapshots of 192 different people. 3DPeS images are normalized to the size of 128 × 64 pixels. The silhouette mask is automatically extracted by [47] and provided in [7]. In contrast to VIPeR data set, 3DPeS contains multishots for each person. We evaluate the capability to integrate multishots on the 3DPeS.

We compare CMC performance with five approaches, SDALF, Ensemble, SARC3D [8], ColorInv, and EIF^{old} on the 3DPeS. Baltieri *et al.* [8] propose a simplified nonarticulated 3D body model to spatially map appearance descriptors (color



(a) (b) (c) (d)

Fig. 12. Comparison to the SDALF [3], Ensemble [4], SARC3D [8], ColorInv [5], and EIF$^{\text{old}}$ [12] on the 3DPeS data set ($N$vs$M$: N gallery shots versus M probe shots). (a) 1vs1. (b) 3vs1. (c) 5vs1. (d) 3vs3.



(a) (b)

Fig. 13. Distribution of the number of components K by 2WGMMF. (a) Concatenated histogram. (b) Joint histogram.

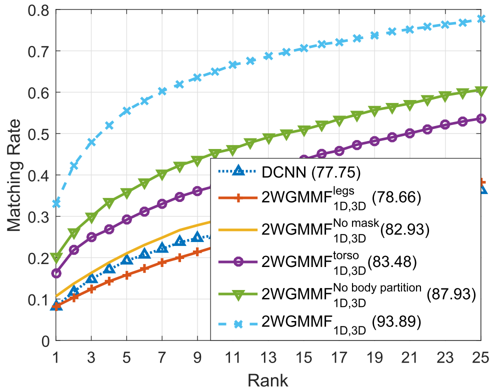


Fig. 14. Comparative results of holistic and regional features.

and gradient histograms) into a vertex-based 3D body surface. Due to the 3D information, they can directly handle the issues of occlusions, partial views, or pose changes, which normally cause performance degradation by using 2D descriptors. However, the image-to-model mapping needs the perspective projection matrix (intrinsic parameters) and extrinsic calibration matrix to estimate the rotation and translation information between the world reference and the model coordinates. Fig. 12 shows the top-rank results of our proposed algorithm, in comparison with five other state-of-the-art methods, where 2WGMMF features are applied on HSV, RGB, and YCbCr color spaces [see Fig. 9(d)]. For multishot integration, we take a pair of images, which has maximum similarity, as matching identity. Overall, the performance is improved by the proposed method about 3%, 9%, 10%, and 8% better accuracies at rank-1 with 1vs1, 3vs1, 5vs1, and 3vs3 data sets, respectively. N versus M means that a test data set has N gallery images and M probe images for an identity. It is noteworthy that the proposed method achieves a better result than SARC3D, even though our method is purely based on 2D features without requiring any additional information, e.g., intrinsic parameters, extrinsic calibration matrix, and so on, as needed in SARC3D. Specifically, our method achieves 51% rank-1 recognition rate and SARC3D obtains 46% in 1vs1 scenarios, as shown in Fig. 12(a). In the experiments of multishot, the performance of our algorithm is still superior. At rank-1, our proposed scheme achieves about 72%, 83%, and 88%, while SARC3D achieves 51%, 64%, and 70% with 3vs1, 5vs1, and 3vs3, respectively. For the 3DPeS, we only mark nAUC of ColorInv, since implementation software codes or feature representations of the other methods are not available. The CMC curves of the rest are acquired in [8].

D. Discussions

In 2WGMMF, there is no manual tweaking of parameters for experiments. The number of components K is automatically decided by BIC. We present the results of K on two data sets in Fig. 13. Each figure shows the distributions of K on each body part with both concatenated RGB-based color histogram and joint color histogram, respectively. On the VIPeR and 3DPeS data sets, 632 images of 316 identities and 384 images of 192 identities are exploited to derive K, respectively. Most of the concatenated color histogram have six or seven components in both body parts, and the distribution of two parts is very similar on both data sets. In case of joint histogram, the number of components is less than that of concatenated histogram and the distributions of two parts are comparable on both data sets. The only difference between the distributions of VIPeR and 3DPeS is the magnitude due to the difference of data set sizes [e.g., in case of $K = 7$, the number of images is 140 and 78 on the VIPeR and the 3DPeS data sets, respectively, in Fig. 13(a)]. In other words, the distribution of K is consistent on the disparate data sets of pedestrian images. Since person appearance, i.e., composition of clothes, is similar on the street, these results are reasonable. Thus, we can conclude that our proposed method, 2WGMMF, is robust to extract the dominant color modes.

We have also conducted experiments to compare the separate importance of holistic and regional features on the VIPeR. In Fig. 14, the use of regional features alone shows better performance than that of holistic features. More specifically, 2WGMMF with torso part gives better result than legs part.

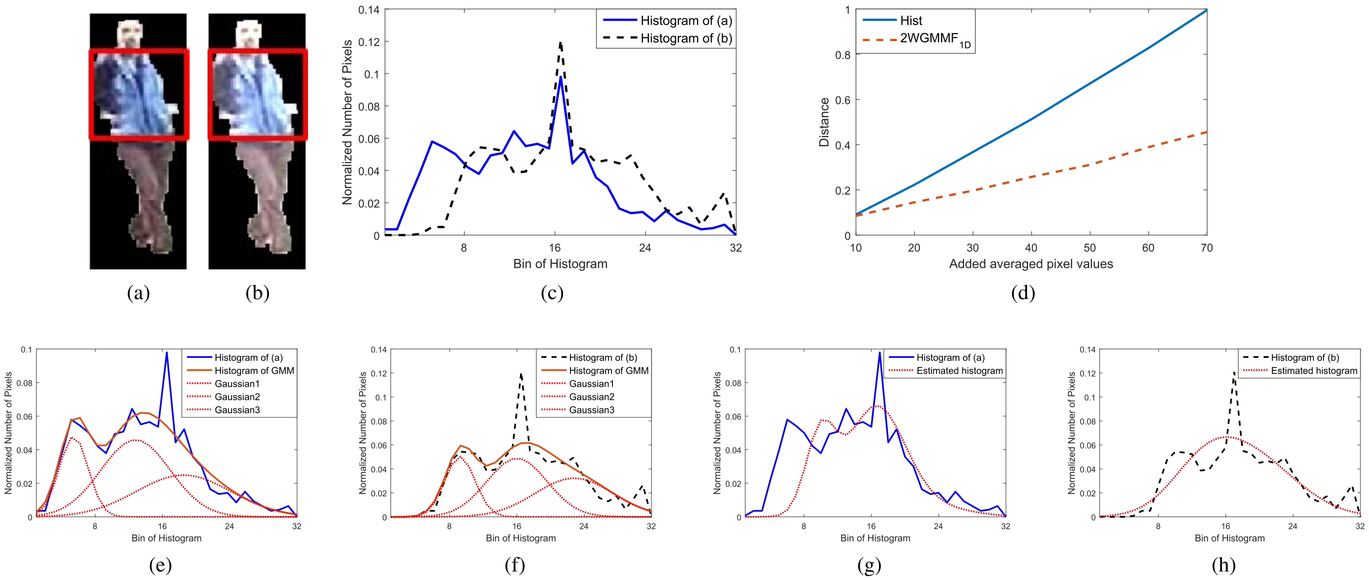


Fig. 15. (a) ID 81 foreground image of VIPeR. (b) Averaged 30 pixel value added image of (a). (c) Red channel histograms of torsos of (a) and (b). (d) Comparative results of histogram and 2WGMMF distances with concatenated histogram. (e) GMM of (a). (f) GMM of (b). (g) GMM fit of (a) with (b). (h) GMM fit of (b) with (a).

That means torso part gives more discriminative information for re-id. In addition, green curve in Fig. 14 shows the results of no body partition. When we compare green curve and sky blue curve, we can find that body partition indeed increases the rank-1 score about 13%. We conjecture that in the process of building histograms without body partition, pixel values from different colored clothes are overlapped and mixed in the histogram domain. That prevents us from extracting genuine color components, resulting in degraded accuracy. To find out the impact of background removal, we have evaluated the performance with no masks. Yellow curve shows the performance of 2WGMMF without mask. As expected, the performance without masks degrades significantly compared with that with mask, which is shown in sky blue curve. However, 2WGMMF still outperforms DCNN when no silhouette masks are available. In the proposed method, masks are exploited in body partition (Section IV-A) and feature extraction (Section IV-B). If background appearance inside the detected bounding box is included, it can result in less accurate boundary lines between head–torso and torso–legs in body partition, and disturb dominant color modes extraction in feature extraction. In other words, color components from background can be regarded as dominant color modes of person appearance and that creates more mismatches.

We also conduct some experiments to show some extent of tolerance on illumination changes of our method. In person re-id, illumination and viewpoint changes of two nonoverlapping cameras are the most difficult problems to solve. For example, strong light makes image in Fig. 15(a) brighter, which shown in Fig. 15(b) where red rectangles represent torso parts, and color histogram shifts to the right-hand side in Fig. 15(c). Fig. 15(c) shows the examples of red channel histograms of torso parts when we increase the value by 30 on average to each pixel. Even though color histogram is shifted, Gaussian components are almost retained in Fig. 15(f) compared with Fig. 15(e). Fig. 15(d) shows distance curve with respect to several added averaged values on Chi-squared distance of histogram and 2WGMMF with concatenated histogram. It can be seen that the slope of 2WGMMF is smaller than histogram distance, which implies 2WGMMF can tolerate more significant color changes caused by lighting change rather than histogram distance. 2WGMMF fits Gaussian components to probe-to-gallery and vice versa. Then, the scale change of each Gaussian is able to mitigate the effect of shifted histogram [see Fig. 15(g) and (h)], so that 2WGMMF tolerates more significant light changes.

In this paper, since we assume that we do not have any additional training set to train/update our model, it is difficult to increase the model adaptation. But once we are capable of collecting some training data from a specific camera network, we can improve from two aspects. First is improving ensemble/combination strategy and second is enhancing the similarity measurement. The effective feature ensemble or combination has still not been conclusively addressed in the existing studies [48]. Most reported methods concatenate all the feature vectors from different cues, which may cause some issues. First, the total feature dimension is increased when adding new cues, resulting in higher computational load. Second, the direct concatenation does not consider the importance among different image cues, so high-dimensional features will probably dominate the low-dimensional ones. In fact, different features may carry complementary information, e.g., our proposed pretrained DCNN features and 2WGMMF, and they require effective fusion instead of direct concatenation, which may lose some useful information for re-id. To tackle these problems, ensemble learning methods could be applied to get better predictive performance, for example, the linear combination of similarity measurements scheme where the weight parameters w are learned by the SVM, AdaBoost, or nonlinear combination by Bagging (e.g., random forest algorithm).

VII. Conclusion

We propose a novel enhanced and integrated person re-id framework, which consists of three important techniques: holistic invariant feature extraction, regional invariant feature extraction, and aggregation. A pretrained DCNN is used for extracting features describing holistic person appearance, including color, texture, shape, and other visual cues. Furthermore, we propose a two-way GMM fitting scheme to model dominant color modes of target image as GMM in color histogram domain with the partitioned body parts. We also propose the integration scheme to combine three feature distances effectively using min–max normalization. In the experiments, we show that the new framework exceeds the state-of-the-art methods on the challenging benchmarks.

References

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds., *Person Re-Identification*, vol. 1. London, U.K.: Springer, 2014.

[2] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, “Person re-identification using Haar-based and DCD-based signature,” in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug./Sep. 2010, pp. 1–8.

[3] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2360–2367.

[4] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2008, pp. 262–275.

[5] I. Kviatkovsky, A. Adam, and E. Rivlin, “Color invariants for person reidentification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.

[6] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by salience matching,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2528–2535.

[7] D. Baltieri, R. Vezzani, and R. Cucchiara, “3DPeS: 3D people dataset for surveillance and forensics,” in *Proc. Joint ACM Workshop Human Gesture Behavior Understand.*, 2011, pp. 59–64.

[8] D. Baltieri, R. Vezzani, and R. Cucchiara, “Mapping appearance descriptors on 3D body models for people re-identification,” *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 345–364, Feb. 2015.

[9] T. Gandhi and M. M. Trivedi, “Person tracking and reidentification: Introducing panoramic appearance map (PAM) for feature representation,” *Mach. Vis. Appl.*, vol. 18, no. 3, pp. 207–220, Aug. 2007.

[10] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux, “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences,” in *Proc. ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Sep. 2008, pp. 1–6.

[11] W. Hu, M. Hu, X. Zhou, T. Tan, J. Luo, and S. Maybank, “Principal axis-based correspondence between multiple cameras for people tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.

[12] S.-C. Chen, Y.-G. Lee, J.-N. Hwang, Y.-P. Hung, and J.-H. Yoo, “An ensemble of invariant features for person re-identification,” in *Proc. IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2016, pp. 1–6.

[13] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” in *Proc. IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, vol. 3. Sep. 2007, pp. 1–5.

[14] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in *Image Analysis*. Berlin, Germany: Springer, 2011, pp. 91–102.

[15] T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by video ranking,” in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2014, pp. 688–703.

[16] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. “Person re-identification by support vector ranking,” *BMVC*, vol. 2, no. 5, pp. 1–11, 2010.

[17] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3586–3593.

[18] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, “Pedestrian recognition with a learned metric,” in *Computer Vision—ACCV*. Berlin, Germany: Springer, 2011, pp. 501–512.

[19] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2288–2295.

[20] A. Mignon and F. Jurie, “PCCA: A new approach for distance learning from sparse pairwise constraints,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2666–2672.

[21] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 152–159.

[22] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[23] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” *BMVC*, vol. 1, no. 2, pp. 1–11, 2011.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[26] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2014, pp. 818–833.

[27] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, “Pedestrian detection with unsupervised multi-stage feature learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3626–3633.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1717–1724.

[29] O. Russakovsky *et al.*, “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[30] P.-A. Savalle, S. Tsogkas, G. Papandreou, and I. Kokkinos, “Deformable part models with CNN features,” in *Proc. Parts Attributes Workshop Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 1–5.

[31] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, “Hierarchical convolutional features for visual tracking,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[32] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). “Return of the devil in the details: Delving deep into convolutional nets.” arXiv preprint arXiv:1405.3531.

[33] C. Szegedy *et al.* (2014). “Going deeper with convolutions.” [Online]. Available: https://arxiv.org/abs/1409.4842

[34] K. Jeong and C. Jaynes, “Object matching in disjoint cameras using a color transfer approach,” *Mach. Vis. Appl.*, vol. 19, no. 5, pp. 443–455, Oct. 2008.

[35] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[36] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2004.

[37] K. Roeder and L. Wasserman, “Practical Bayesian density estimation using mixtures of normals,” *J. Amer. Statist. Assoc.*, vol. 92, no. 439, pp. 894–902, Sep. 1997.

[38] J. G. Campbell, C. Fraley, F. Murtagh, and A. E. Raftery, “Linear flaw detection in woven textiles using model-based clustering,” *Pattern Recognit. Lett.*, vol. 18, no. 14, pp. 1539–1548, Dec. 1997.

[39] A. Dasgupta and A. E. Raftery, “Detecting features in spatial point processes with clutter via model-based clustering,” *J. Amer. Statist. Assoc.*, vol. 93, no. 441, pp. 294–302, Mar. 1998.

[40] S. Franzini and J. Ben-Arie, “Speech recognition by indexing and sequencing,” in *Proc. Int. Conf. Soft Comput. Pattern Recognit. (SoCPaR)*, Dec. 2010, pp. 93–98.

[41] K. Ma and J. Ben-Arie, “Vector array based multi-view face detection with compound exemplars,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3186–3193.

[42] B. Ma, Y. Su, and F. Jurie, “Covariance descriptor based on bio-inspired features for person re-identification and face verification,” *Image Vis. Comput.*, vol. 32, nos. 6–7, pp. 379–390, Jun./Jul. 2014.

[43] N. Jojic, A. Perina, M. Cristani, V. Murino, and B. Frey, "Stel component analysis: Modeling spatial correlations in image class structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2044–2051.

[44] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2012, pp. 1–11.

[45] Evaluation Results of Bicov, Cov and Gbicov. [Online]. Available: http://vipl.ict.ac.cn/members/bpma

[46] Evaluation Results of Partssc and Hist. [Online]. Available: http://www.cs.technion.ac.il/~kviat/colorReid.htm

[47] R. Vezzani, C. Grana, and R. Cucchiara, "Probabilistic people tracking with appearance models and occlusion classification: The AD-HOC system," *Pattern Recognit. Lett.*, vol. 32, no. 6, pp. 867–877, Apr. 2011.

[48] X. Liu, H. Wang, Y. Wu, J. Yang, and M.-H. Yang, "An ensemble color model for human re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 868–875.

Young-Gun Lee (S'07) received the B.S. degree in chemistry and physics from Republic of Korea Air Force Academy, Cheongwon, South Korea, in 2005 and the M.S. degree in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2009. He is currently working toward the Ph.D. degree with University of Washington, Seattle, WA, USA.

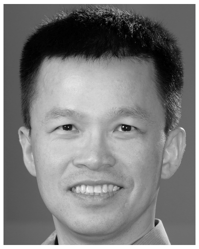
His research interests include computer vision, image processing, and video surveillance.

Shen-Chi Chen received the B.S. degree in computer science from National Chengchi University, Taipei, Taiwan, in 2007; the M.S. degree in biomedical engineering from College of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, in 2009; and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, in 2016. He is the Senior Staff Engineer with Huawei Research and Development, Santa Clara, CA, USA. His research interests include computer vision, pattern recognition, video surveillance, deep learning, and distributed cloud computing.

Jenq-Neng Hwang (F'01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from University of Southern California, Los Angeles, CA, USA.

In 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, WA, USA, where he was promoted to Full Professor in 1999. He has authored over 300 journal, conference papers, and book chapters in the areas of multimedia signal processing, and multimedia system integration and networking. He has authored the book *Multimedia Networking: From Theory to Practice* (Cambridge University Press). His research interests include industry on multimedia signal processing and multimedia networking.

Dr. Hwang is a Founding Member of the Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society, where he is also a member. He is a member of the Multimedia Technical Committee of the IEEE Communication Society. He received the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He served as the Program Co-Chair of the IEEE ICME 2016 and was the Program Co-Chair of the ICASSP 1998 and the ISCAS 2009. He served as the Associate Chair for Research from 2003 to 2005, and from 2011 to 2015. He is currently the Associate Chair for Global Affairs and International Development in the EE Department, University of Washington. He was the Society's representative to the IEEE Neural Network Council from 1996 to 2000. He served as an Associate Editor of IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON IMAGE PROCESSING, and *IEEE Signal Processing Magazine*. He currently serves on the Editorial Board of ZTE Communications, *Electronics and Telecommunications Research Institute*, *International Journal of Digital Multimedia Broadcasting*, and *Journal of Signal Processing Systems*.



Yi-Ping Hung (M'87) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1982; the M.S. degree in engineering and applied mathematics from the Division of Engineering and the Division of Applied Mathematics, Brown University, Providence, RI, USA, in 1987 and 1988; and the Ph.D. degree in engineering from the Division of Engineering, Brown University, in 1990.

From 1990 to 2002, he was with the Institute of Information Science, Academia Sinica, Taipei, where he was a Tenured Research Fellow in 1997. He served as the Deputy Director of the Institute of Information Science from 1996 to 1997, and the Director of the Graduate Institute of Networking and Multimedia, National Taiwan University, from 2007 to 2013, where he is currently a Professor with the Department of Computer Science and Information Engineering. He is also a Joint Research Fellow with the Institute of Information Science, Academia Sinica. His research interests include computer vision, pattern recognition, image processing, virtual reality, multimedia, and human-computer interaction.

Dr. Hung was the Program Co-Chair of ACCV'00 and ICAT'00, and the Workshop Co-Chair of ICCV'03. He has been an Editorial Board Member of *International Journal of Computer Vision* since 2004. He will be the General Chair of ACCV'16.