Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 4089505, 11 pages
https://doi.org/10.1155/2017/4089505

Research Article

Color Distribution Pattern Metric for Person Reidentification

Yingsheng Ye, Xingming Zhang, and Wing W. Y. Ng

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

Correspondence should be addressed to Yingsheng Ye; 1543590032@qq.com

Received 18 July 2017; Accepted 27 November 2017; Published 18 December 2017

Academic Editor: Zhaolong Ning

Accompanying the growth of surveillance infrastructures, surveillance IP cameras mount up rapidly, crowding Internet of Things (IoT) with countless surveillance frames and increasing the need of person reidentification (Re-ID) in video searching for surveillance and forensic fields. In real scenarios, performance of current proposed Re-ID methods suffers from pose and viewpoint variations due to feature extraction containing background pixels and fixed feature selection strategy for pose and viewpoint variations. To deal with pose and viewpoint variations, we propose the color distribution pattern metric *(CDPM)* method, employing color distribution pattern (*CDP*) for feature representation and SVM for classification. Different from other methods, *CDP* does not extract features over a certain number of dense blocks and is free from varied pedestrian image resolutions and resizing distortion. Moreover, it provides more precise features with less background influences under different body types, severe pose variations, and viewpoint variations. Experimental results show that our *CDPM* method achieves state-of-the-art performance on both 3DPeS dataset and ImageLab Pedestrian Recognition dataset with 68.8% and 79.8% rank 1 accuracy, respectively, under the single-shot experimental setting.

1. Introduction

The person reidentification (Re-ID) task searches for a targeted person from images captured in different times and places under the assumption that the target person is wearing the same clothing as provided in source images. With wide deployments of public surveillance infrastructures and private surveillance products, people are paying more attention to Re-ID than ever before and are desperate for assistance from Re-ID methods when searching for an individual among massive amounts of surveillance videos. Although many methods have been proposed in the last decade with notable progress, Re-ID methods still face many challenges in real scenarios. Intraclass differences of pose and viewpoint variations, interclass similarities of appearance, different camera settings, and significant environment changes together make the Re-ID task much more complicated and challenging. Movements of body parts cause intraclass differences in pose and viewpoints, while interclass similarities include resemblances between 2 individuals sharing similar body types with similar or identical clothing. Furthermore, significant environmental changes like illumination changes and occlusions have great influence on a person's appearance and bring environmental background noises. To deal with these challenges, most Re-ID algorithms focus on finding robustness features and a reliable similarity classification metric. For robustness features, unlike shape, position, soft-biometry, and body models, combination features of color and texture information based on dense blocks are very popular in many approaches well summarized in [1]. Besides these features, silhouette information also has its own challenges and opportunities in Re-ID. Silhouette information comes from background substraction of surveillance videos. In IoT, it is possible to integrate background substraction and device management service like the lightweight RESTful Web service [2] into IP cameras to help ease the load of IoT and make IP cameras smarter. In our work, we traded texture information for silhouette information using combination of color and silhouette information. As for finding a reliable similarity classification metric, an annoying problem is lack of sufficiently labeled training samples, especially positive training samples. Therefore, we proposed a grouping strategy called Pure-vs-Exception to separate training images into different groups and automatically generate sufficient and balanced labeled training samples.

In this paper, we propose a color distribution pattern metric (*CDPM*) method for Re-ID. *CDPM* consists of a color distribution pattern (*CDP*) feature model for feature

extraction and a SVM [3] for training and classification, under the assumption that the target person still wears the same clothing. Our *CDP* feature model is designed to deal with background interference, pose variations, and viewpoint variations. It extracts color distribution patterns from HSV and Lab color spaces, combining with silhouette information based on leg and torso-head (TH) body parts derived from the leg-torso-head body model [4]. In similarity classification metric learning, we aim to maximize differences of interclass variations and suppress differences of intraclass variations by proposing the Pure-vs-Exception strategy. We use Pure-vs-Exception strategy to group pedestrian images and generate training samples, which simulate significant difference patterns between 2 pedestrian images of 2 different individuals and significant difference patterns between 2 pedestrian images of the same individual. Experimental results show that our *CDPM* method outperforms other comparative approaches on both 3DPeS and ImageLab Pedestrian Recognition (IPR) datasets and possesses great tolerance for pose and viewpoint variations.

The remainder of this paper is as follows. Section 2 presents a review of related works. Section 3 is a detailed introduction of the proposed *CDPM*. Experimental results on 2 public datasets are covered in Section 4. Section 5 concludes this paper.

2. Related Work

Almost all Re-ID approaches are based on 2-step framework: (i) feature extraction from pedestrian image pairs and (ii) feature similarity calculation under a prelearned classification metric. Thus, most efforts are devoted to these 2 steps. According to the Re-ID taxonomy summarized by Vezzani et al. [1], local features and pedestrian body model-based features are very common because they capture detailed and localized information for matching. Among local features, it is very popular to use combination features of color and texture histograms, for example, color histograms (from HSV color space) and scale invariant local ternary pattern (SILTP) histograms [5], color histograms (from RGB, YUV, and HSV color spaces) and local binary pattern (LBP) histograms [6, 7], color histograms (from RGB and HSV color spaces) and LBP histograms [8], color histograms (from RGB, YCbCr, and HS color spaces) and texture histograms of Gabor filters and Schmid filters [9], and color histograms (from HSV and Lab color spaces) and LBP histograms [10, 11]. The implementation details of these color and texture histograms might be different. After feature extraction, it is very important to find a reliable similarity classification metric exploiting feature representativeness. Hence, a variety of strategies are introduced when searching for a reliable classification metric.

Some approaches focus on cross-domains classification metric learning. Hu et al. [7] employed deep learning in a transfer metric, which learned hierarchical nonlinear transformations of cross-domains by mapping discriminative knowledge between a labeled source domain and unlabeled target domain. Others [12] jointly learned a transfer metric in an asymmetric way by extracting discriminant shared components through multitask modeling to enhance target interclass differences under shared latent space. Recently, Shi et al. [13] showed interest in attribute features extracted from semantic level for the Re-ID task and employed it in cross-domains transfer metric learning.

Most of these approaches focused on classification metric learning under the same domain. In [14], authors used an improved deep learning architecture with 2 novel layers to extract the relationships of the input image pair. One layer computed the cross-input neighborhood differences and the other subsequently summarized these differences. Köstinger et al. [15] presented KISS classification metric learned from equivalence constraints based on statistical inference rather than computational complex optimization. However, this classification metric learning could be unstable under a small-sized training set, mentioned in [10]. Thus, Tao et al. [10] integrated smoothing and regularization techniques in KISS for robust estimation of covariance matrices and stable performance and proposed regularized smoothing KISS. Pedagadi et al. [16] used local Fisher discriminant analysis (LFDA) to learn similarity classification metric from a set of labeled feature pairs, which consisted of features extracted from high-dimensional features by unsupervised PCA. However, unsupervised PCA undesirably compressed most of discriminative features under relatively small dataset. Hence, Xiong et al. [6] used kernel approach to replace the unsupervised dimensionality reduction by supervised dimensionality reduction and proposed kernel LDFA (kLDFA), avoiding eigenvalue decomposition of large scatter matrices and replacing kernel easily for better performance. Other researchers [11] used a reference set to learn a reference subspace by regularized canonical correlation analysis. Then, similarity was calculated by the features projected in reference subspace and reranked by saliency information. Conversely, Liao et al. [5] proposed cross-view quadratic discriminant analysis to learn similarity classification metric under a low-dimensional subspace. Furthermore, Zhang et al. [17] treated hashing learning as regularized similarity learning and used deep learning to train hashing codes and additional regularization term, which encoded adjacency consistency.

Besides approaches with a fixed classification metric for all image pairs, some approaches use multiple metrics in Re-ID. To address viewpoint variations, Wang et al. [8] proposed a data-driven distance metric method based on cross-view support consistency and cross-view projection consistency to readjust the distance metric for each image pair by exploiting the training data each time, which was quite computational for a large scale. Ma et al. [9] modeled a multitask distance metric for the Re-ID in camera networks and proposed multitask maximally collapsing metric learning method to learn distance metrics between different cameras. Others [19] proposed a spatiotemporal feature named optical flow energy image (OFEI) and a corresponding matching method named multiview relevance metric learning with listwise constraints (mvRMLLC) for video-based Re-ID. The mvRMLLC assumed that OFEI features of the same person from different views should be globally consistent with their similarities.

Different from these above 2-dimensional- (2D-) based methods, a 3-dimensional (3D) method called SARC3D [20]

(a) Different body types (b) Pose and viewpoint variations

Figure 1: Sample images from 3DPeS dataset.

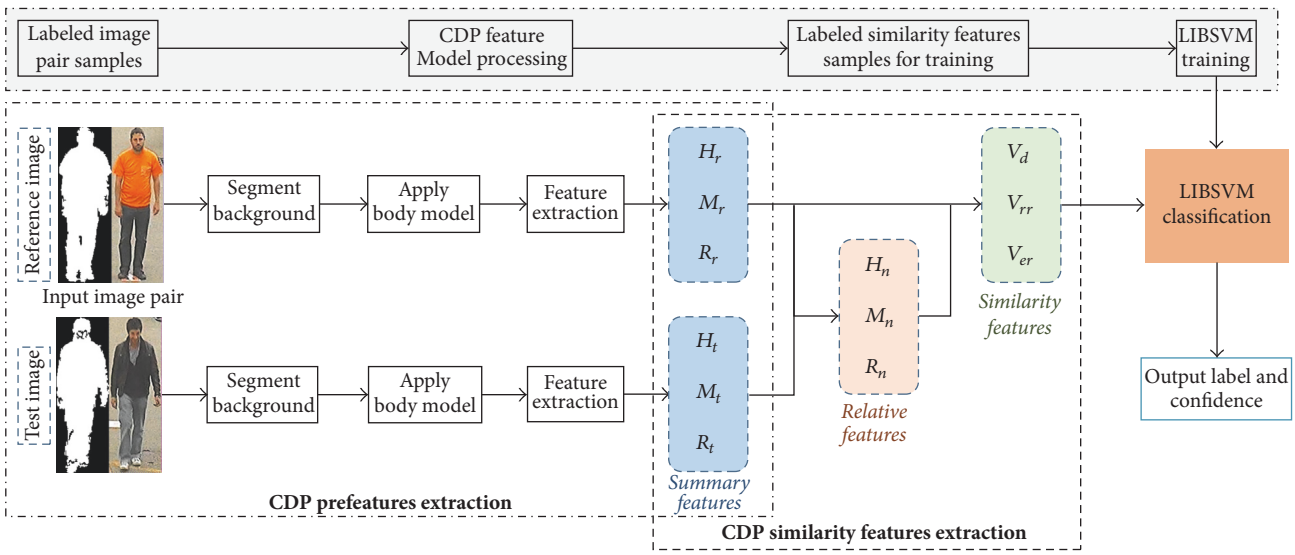

Figure 2: CDPM for person reidentification. The *CDP* feature model contains two parts: prefeatures extraction and similarity features extraction.

used averaged side, frontal, and top views silhouettes to learn an approximate graphical 3D body model, which then could compare people from even 1 image due to its precise 3D feature mapping via a selected set of vertices. This comparison not only allowed one to look for details and global features but also copped with partial data and occluded views.

However, methods based on dense block features [5–14, 16], which encode not only person pixel information but also background pixel information, cannot properly filter out background information due to pose variations and different body types. Furthermore, dense block features introduce distortion caused by resizing from different resolutions. Figure 1 shows different body types (including massive build, medium build, and slim build), pose variations, and viewpoint variations (note that viewpoint variations in this paper refer to front view, back view, left view, and right view or any state in-between but do not include top view). Matching accuracies of these aforementioned methods [5–14, 16] suffer from encoded background information due to inappropriate feature selection, in which case the fixed classification metric could not properly separate encoded person information from encoded background information, under significant pose variations and different body types. We refer to this kind of interference as background interference.

Our *CDP* feature model is designed to deal with background interference by filtering out pedestrian image background pixels using silhouette images provided or extracted from surveillance sequences and generates more precise pedestrian features than those obtained by other methods [5–11, 14]. In addition, our *CDP* feature model does not have to worry about body type variations, pose variations, and image resolution variations, which are common in real scenarios of video surveillance and video forensics, as long as the silhouette images are well provided.

3. Color Distribution Pattern Metric

Our *CDPM* is designed to deal with severe pose and viewpoint variations, as well as background influences. Furthermore, resolutions of pedestrian images are not required to be fixed in *CDPM*, different from fixed image resolutions in [5–14, 16]. This characteristic of *CDPM* avoids distortion interference brought by resizing.

Figure 2 shows the main processes of our *CDPM*, including *CDP* feature model, SVM training, and SVM classification. *CDP* feature model is used to extract similarity features from image pairs, while SVM is used for training and classification based on similarity features. We divide *CDP*

feature model into 2 phases: *CDP* prefeatures extraction and *CDP* similarity features extraction. Implementations of input images in the prefeatures extraction phase are independent from image pairs and could be preprocessed parallelly. For *CDP* prefeatures extractions in Figure 2, inputs include two pedestrian images (an image pair of test image and reference image) with corresponding silhouette mask images, which are in the form of binary pixels. Through successive implementations of background segmentation, applying body model, and features extraction, we obtain summary features for each pedestrian image at the end of *CDP* prefeatures extraction phase. In *CDP* similarity extraction phase from Figure 2, we use summary features of test and reference images to produce relative features. And then we extract similarity features from relative features and summary features of reference image only. Finally, we use similarity features as the SVM input for classification metric learning and prediction. In classification metric learning, shown in the top row of Figure 2, training samples are labeled similarity features extracted from labeled image pairs, where each image pair contains one test image and one reference image.

The following are the detailed implementations of *CDP* feature model and metric learning.

3.1. CDP Prefeature Extraction. We use background pixels ratios calculated by (1) to observe percentages of background pixels in pedestrian images.

$$R_{\text{background pixels}} = 1 - \frac{\sum_{x=0}^{r-1} \sum_{y=0}^{c-1} B(x, y)}{r * c}, \tag{1}$$

where r and c represent rows and columns of corresponding silhouette image, respectively, and B represents the binary image of corresponding silhouette image. According to 3DPeS dataset [21], background pixels ratios of pedestrian images vary from 34% to 89%, with a mean value of 59.4%. In IPR dataset [18], background pixels ratios vary from 36% to 80%, with a mean value of 59%. The average background pixels ratio is about 59.2% in these two datasets, which could bring great interference for feature robustness and classification if they are not filtered out before final classification. To avoid interference caused by background pixels, we use silhouette images to segment pedestrian from the background by setting background pixels values to 0. First, silhouette images are transformed into binary images using thresholds. Then, we multiply the binary image with corresponding original image at pixel level for each channel. We filter out background pixels of pedestrian images before feature extraction, saving a lot of work during classification metric learning by enhancing feature efficiency and robustness. So, we finish part of the work of feature selection at the beginning of feature extraction, unlike most of state-of-the-art methods, which implemented feature selection after applying dense block feature extraction.

After filtering out background pixels, we divide the pedestrian image into 2 parts, leg and torso-head (TH), based on the leg-torso-head [4] body model. This implementation aims to extract locations of body parts and enhance robustness of *CDP* model. The main reason we choose leg and TH parts is that most clothing styles around the world share a commonality that there are differences between leg and TH parts. Even though the whole outfit may share the same color, it is still essential and practical to treat body as 2 parts, while the other one from input image pair shares a different clothing style. However, due to small proportion of head part and low resolution of pedestrian images in real surveillance scenarios and datasets, it is neither necessary nor significant to treat head part alone. Besides, it is difficult to automatically and accurately divide the head part. Therefore, we choose to put torso and head parts together as TH to reduce computation cost for more practical implementation.

When these 2 aforementioned implementations are ready, we implement color extraction. We use both HSV and lab color spaces to extract color distribution histogram matrices on leg and TH parts, respectively, while ignoring the pixels with a value of 0. Details of color distribution extraction are as follows. With 2 color spaces involved, we obtain 6 channels of data for each image and group them into 3 pairs: HS, ab, and LV. We then apply 2D histogram extraction on these 3 pairs of channels to extract color distribution histogram matrices with sizes $(16, 32)$ for HS channels, $(32, 32)$ for ab channels, and $(32, 32)$ for LV channels. So, we obtain 6 histogram matrices, 3 histogram matrices each for the leg and TH parts, respectively. We mark these histogram matrices as H. For each histogram matrix, we generate a binary matrix with a threshold value of 3 for noises reduction. Then we mark these 6 binary matrices as occurrence masks M. The occurrence masks specify the domain boundaries for corresponding histogram matrices. We use occurrence masks M to update the corresponding histogram matrices H by multiplying H with M at pixel level as follows:

$$H(x, y) = H(x, y) M(x, y). \tag{2}$$

We use these 6 occurrence masks to calculate the corresponding occurrence ratios, plus mean occurrence ratios of occurrence ratios on leg and TH parts for each pair of channels, which produced 3 mean occurrence ratios. The occurrence ratio, not the mean occurrence ratio, is calculated as follows:

$$R_{\text{occurrence-ratio}} = \frac{\sum_{x=0}^{r-1} \sum_{y=0}^{c-1} M(x, y)}{r * c}. \tag{3}$$

In results, there are 9 occurrence ratios and we mark these 9 occurrence ratios as R. At this point, summary features including histogram matrices H, occurrence masks M, and occurrence ratios R are extracted, representing the color distribution of each pedestrian image without background pixels. Thus, for the reference image in Figure 2, we have its summary features marked as H_r, M_r, and R_r. As for the test image in Figure 2, we have its summary features marked as H_t, M_t, and R_t.

3.2. CDP Similarity Features Extraction. Obviously, summary features extracted above are not the final data that we want in the classification phase. There must be a pair of images in the matching phase: one as the reference pedestrian image and the other as the test pedestrian image. From last subsection,

we already get H_r, M_r, and R_r for reference pedestrian image and H_t, M_t, and R_t for test pedestrian image. Now, we use H_r and M_r to filter out some certain irrelevant patterns in H_t and M_t by multiplying them at pixel level, shown in (4). We treat the resulting features as parts of the relative features and mark them as H_n and M_n.

$$\begin{aligned} H_n(x,y) &= H_t(x,y)\, M_r(x,y), \\ M_{\mathrm{n}}(x,y) &= M_t(x,y)\, M_r(x,y). \end{aligned} \tag{4}$$

Besides H_n and M_n, relative features also include occurrence ratios R_n. We calculate R_n directly from M_n using (3) in *CDP* prefeature extraction phase. H_n enables filtering out some certain irrelevant color distribution patterns while letting similarity matching focus on the rest. M_n enables performing a domain boundary match in later implementation. In other words, H_n focuses on the fine-grain distribution matching, while M_n and R_n focus on domain boundaries matching.

Next step is to use the summary features of reference pedestrian image and relative features to generate similarity features for SVM [3] training and classification. Firstly, we normalize both H_r and H_n by dividing the cumulated sum of their own, respectively. This normalization aims to reduce fluctuation noises caused by occlusions, pose variations, and viewpoint variations, making H_r and H_n more compatible for different scales of input images while well preserving color distribution patterns. Then, we compute the absolute difference matrices of H_r and H_n. Right after that, we project these difference matrices into a subspace by computing the corresponding cumulated sums of absolute difference matrices as distance values by (5). We mark these 6 distance values calculated from the corresponding absolute difference matrices as V_d.

$$V_d = \mathrm{sum}\left(\mathrm{abs}\left(\frac{H_n}{\mathrm{sum}(H_n)} - \frac{H_r}{\mathrm{sum}(H_r)}\right)\right). \tag{5}$$

Secondly, we get 6 occurrence relative ratios R_{r-o-r} from corresponding M_r and M_n calculated by (6), plus 1 summary relative ratio R_{s-r-r} calculated by (7), resulting in 7 relative ratios marked as V_{rr}.

$$R_{r-o-r} = \frac{\sum_{x=0}^{r-1}\sum_{y=0}^{c-1}\left(M_n(x,y)\, M_r(x,y)\right)}{\sum_{x=0}^{r-1}\sum_{y=0}^{c-1} M_r(x,y)}, \tag{6}$$

$$R_{s-r-r} = \frac{\sum_{i=1}^{6}\sum_{x=0}^{r-1}\sum_{y=0}^{c-1}\left(M_{n,i}(x,y)\, M_{r,i}(x,y)\right)}{\sum_{i=1}^{6}\sum_{x=0}^{r-1}\sum_{y=0}^{c-1} M_{r,i}(x,y)}, \tag{7}$$

where i stands for different occurrence mask in both M_n and M_r. Thirdly, for R_r and R_n, we calculate absolute difference ratios by (8) and mark them as V_{er}.

$$V_{er} = \mathrm{abs}(R_n - R_r). \tag{8}$$

Finally, we concatenate V_d, V_{rr}, and V_{er} into a 22-dimensional vector forming the similarity features as input of SVM for training and classification. This 22-dimensional vector is designed with the aim of avoiding background pixels and providing robust representativeness for classification.

3.3. Classification Metric Learning. In classification metric learning, we use 435 pedestrian images, which are automatically extracted from ViSOR [22] surveillance video sequences, to learn interclass variations between 2 pedestrian images of 2 individuals and intraclass variations between 2 pedestrian images of the same individual. We propose a Pure-vs-Exception strategy to separate these 435 pedestrian images into 4 groups. Three of them are Pure Groups, only containing 3 different individuals with each individual corresponding to 1 specific Pure Group. The last group, called Exception Group, contains pedestrian images of individuals different from those individuals in Pure Groups. Pure Groups have 130, 134, and 136 pedestrian images, respectively, while the Exception Group has 35 pedestrian images of 21 individuals. Positive training samples are generated from every combination of 2 images inside each Pure Group, considering combination order which matters for reference image and test image. So, Pure Groups have 16,900 (130^2), 17,956 (134^2), and 18,496 (136^2) positive training samples, respectively, adding up to 53,352. Meanwhile, negative training samples are generated from image pairs where one is from one of the Pure Groups and the other is from any other group, resulting in an amount of 67,324. The number 67,324 comes from $(130 * (134 + 136 + 35) + 134 * (136 + 35) + 136 * 35)$. Thus, we generate a total of 120,676 labeled samples for training from these 435 pedestrian images.

With these training samples, we explore *CDPM*'s performance using the SVM provided by [3] under different kernels and different kernel degrees, while the other parameters of SVM are set as default. The kernels include linear kernel (Lin.), polynomial kernel (Poly.), radial basis kernel (Rad.), and sigmoid kernel (Sig.). The kernel degree varies from 1 to 5, which only works in polynomial kernel. We train the SVMs with different kernels and degrees on these 120,676 labeled samples.

After training, we also use 200 pedestrian images (with 50 individuals and 4 pedestrian images of different viewpoints for each individual) in IPR dataset to generate 20,400 labeled samples with 800 positives and 19,600 negatives for the classification test. Table 1 shows the test results of the trained *CDPM* methods. In Table 1, AC means overall accuracy on test samples, while PM and NM represent the false prediction rates inside positive samples and negative samples, respectively. The highest accuracies are highlighted in bold type. As the kernel degree varies from 1 to 5, test accuracy of polynomial *CDPM* descends from 94.89% to 91.93%, and test accuracy declines to 81.57% when the kernel degree reaches 10 in further tests. The classification results of all 4 kernels, where polynomial kernel's degree is 1, are very close, showing the stability of our *CDP* feature model. Even though the test samples are unbalanced on the distribution of positives and negatives, PMs and NMs are both below 9%, and the differences between them are less than 6%. This shows great and balanced classification performance of *CDPM* methods on both positives and negatives. However, it is hard to determine which kernel is better in the Re-ID ranking problem from Table 1 without any further experiments.

Table 1: Classification accuracies oF *CDPM* methods.

Degree	1			2			3			4			5		
	AC	PM	NM	AC	PM	NM	AC	PM	NM	AC	PM	NM	AC	PM	NM
CDPM*l* (Lin.)	**95.31**	8.25	4.54	-	-	-	-	-	-	-	-	-	-	-	-
CDPM*p* (Poly.)	94.89	5.88	5.08	94.43	5.25	5.59	93.78	4.25	6.3	93.08	3.25	7.07	91.93	2.5	8.3
CDPM*r* (Rad.)	94.82	6.88	5.11	-	-	-	-	-	-	-	-	-	-	-	-
CDPM*s* (Sig.)	94.81	6	5.15	-	-	-	-	-	-	-	-	-	-	-	-

Figure 3: Sample pedestrian images from 3DPeS.

Thus, in further experiments, we use these *CDPM* methods with 4 kernel types (where the kernel degree for polynomial kernel is 1) to evaluate our *CDP* feature models capabilities and compare them with other state-of-the-art methods on 2 public datasets: 3DPeS dataset [21] and IPR dataset [18].

4. Experiment

In this section, we evaluate our *CDPM* method on 2 public datasets: 3DPeS dataset [21] and IPR dataset [18]. These 2 datasets are based on real surveillance setup of University of Modena and Reggio Emilia campus. Surveillance setup contains 8 cameras at different zoom levels with a recording resolution of 704 × 576 pixels. 3DPeS and IPR are 2 of the few person Re-ID datasets that provide silhouette images required by our *CDPM* method, while the others, like VIPeR [23], iLIDS [24], and CAVIAR [25], do not provide silhouette images. Note that, due to implementation limits, 3DPeS and IPR datasets provide silhouette images that filter out most background pixels rather than all background pixels. In addition, IPR dataset is better segmented and smaller than 3DPeS, while some silhouette images in 3DPeS contain shadow areas caused by sunlight.

Here, we present averaged comparison results with the widely used cumulative match characteristic (CMC) and CMC curves. We also report the proportion of uncertainty removed (PUR) value proposed in [16]:

$$\mathrm{PUR} = \frac{\log(S) + \sum_{r}^{S} M(r) \log(M(r) + e)}{\log(S)}, \quad (9)$$

where S is the size of test data, $M(r)$ is match characteristic instead of cumulative match characteristic, and e is an extremely small constant to avoid meaningless definition in logarithm when $M(r)$ is 0. In our experiment, e is set to $1e{-}18$, which still preserves 16 precisions after decimal point for $S = 95$. The PUR value evaluates the entropy information of output rank and is invariant to the logarithm base used. Additionally, PUR value ranges from 0 to 1. A larger PUR value indicates that more values are concentrated in fewer $M(r)$s since all $M(r)$s add up to 1. For example, assume that $M(1)$ is 1 and the remaining $M(r)$'s are all 0, and PUR value is 1 according to (9). When all $M(r)$'s are equal, PUR value is 0. As mentioned in [16], PUR value describes the behaviors of all $M(r)$'s in the entire rank instead of the behavior of any single $M(r)$. To obtain a precise evaluation of different ranks performances, we must combine PUR scores with CMC scores during assessment. This is exceptionally helpful for evaluating close CMC scores at different ranks.

The following subsections explain the details of experimental settings and results.

4.1. Experiment on 3DPeS Dataset

4.1.1. Dataset and Experimental Protocol. 3DPeS [21] dataset provides 1,011 snapshots of 192 individuals with both appearance images and silhouette images. These 1,011 pedestrian images contain significant pose viewpoint variations, as well as illumination changes. Examples are shown in Figure 3.

We adopt the single-shot experiment setting [6] to randomly divide 3DPeS dataset into 2 subsets with P individuals in test subset, where P is set to 95. In the test subset, we randomly choose 1 image of each individual to form the reference set and randomly choose another image of each individual to form the test set. We repeat this partition 10 times to generate 10 groups of test data for evaluation.

4.1.2. Features and Implementation Details. In our experiment, we loosely crop out background areas, as shown in Figure 4, instead of scaling pedestrian images into a fixed resolution like [5–7, 12].

We employ 4 different kernels, linear kernel, polynomial kernel with kernel degree of 1, radial basis kernel, and sigmoid

(a) Before cropping (b) After cropping

Figure 4: Cropping out background extension area.

kernel in our *CDPM*. Thus, there are 4 corresponding *CDPM* methods: *CDPMl*, *CDPMp*, *CDPMr*, and *CDPMs*, in which the postfixes *l*, *p*, *r*, and *s* represent linear kernel, polynomial kernel, radial basis kernel, and sigmoid kernel, respectively. Our *CDP* feature model extracts 32 × 32 2D histogram matrices from 6 channels of Lab and HSV color spaces based on 2 body parts, as well as 32 × 32 occurrence mask matrices, plus a 7-bin occurrence ratio. We project their concatenation into a 22-dimensional feature vector.

In cAMT-DCA [12], color, LBP, and HOG features are extracted from 75 overlapping 16 × 16 blocks. Block step sizes in both horizontal and vertical directions are 8 pixels. The color histogram is a 16-bin histogram from 8 color channels of RGB, YCbCr, and HS. For each block, a 484-dimensional vector is generated, including color, LBP, and HOG features. Since there are 75 blocks in each pedestrian image, a 36,300-dimensional vector is generated, which is then compressed into a 100-dimensional vector by PCA before being applied in cAMT-DCA.

DSTML [7] uses features extracted from 6 nonoverlapping horizontal stripes of each pedestrian image. Features consist of 16-bin histograms from 8 channels of RGB, YUV, and HS channels, as well as uniform LBP histograms with 8 neighbors and 16 neighbors, respectively. Then, histograms of each channel are normalized and concatenated into a 2,580-dimensional feature vector. Before being applied in DSTML, PCA learned from source data is used to project the target feature vector into a low-dimensional feature vector.

PCCA [26], LDFA [16], SVMML [27], KISSME [10], rPCCA [6], kLDFA [6], and MFA [6] use the same set of features. The features used in PCCA are extracted from 32 × 32 overlapping blocks with a step size of 16 pixels in both horizontal and vertical directions, while the features used in the other 6 methods are extracted from 16 × 16 overlapping blocks with a step size of 8 pixels in both horizontal and vertical directions. Features, containing 16-bin histograms of 8 color channels (RGB, YUV, and HS) and 2 kinds of uniform LBP histograms with 8 neighbors and 16 neighbors, are extracted from these blocks. Then, histograms are normalized and concatenated into a feature vector. In PCCA, rPCCA, LFDA, and kLFDA, the concatenated feature vector space is projected into 40-dimensional space, while it is projected into 30-dimensional space in MFA and 70-dimensional space in KISSME.

REV [6] conducts ranking ensemble voting of rPCCA, kLDFA, and MFA by adding rankings in a simple voting scheme.

The first 10 methods in Table 2 use features extracted over dense blocks from resized pedestrian images with a resolution of 128 × 48 pixels. In the parentheses of first row in Table 2, the first parameter in numeric form indicates the region size of pedestrian image when extracting features, while the second parameter indicates the kernel type in which R_{x^2} means $RBF-x^2$ kernel. The parameter settings of these 10 methods are the best settings reported in [6, 7, 12].

Note that our training setting is different from these other methods. We use a training set extracted from ViSOR surveillance video sequences, which is the source of 3DPeS dataset. Our training set only contains 435 pedestrian images of 24 individuals. Among these 435 images, there are 400 pedestrian images coming from 3 individuals in the Pure Groups. In cAMT-DCA, the corresponding training set contains all images in the source dataset (VIPeR, 1,264 pedestrian images of 632 individuals) and the images of the other 97 individuals from 3DPeS dataset. DSTMLs corresponding training set consists of the label information of source dataset CAVIAR, with 1,220 pedestrian images of 72 individuals. The other methods mentioned in this section all use the rest 97 individuals in 3DPeS dataset as training set, with around half of 1,011 pedestrian images. Thus, compared to those methods,

Table 2: PUR and CMC scores on 3DPeS dataset with $P = 95$ test individuals.

r	cAMT-DCA	DSTML	SVMML (75)	PCCA (14, R_{x^2})	LFDA (75)	KISSME (75)	rPCCA (75, x^2)	kLFDA (75, x^2)	MFA (75, x^2)	REV	*CDPMl*	*CDPMp*	*CDPMr*	*CDPMs*
1	31.9	32.5	34.7	42.2	45.5	41.3	47.3	54.0	48.4	54.2	67.9	**68.8**	68.6	68.7
5	53.5	54.3	66.4	71.1	69.2	66.2	75.0	77.7	72.4	77.7	79.9	**80.5**	80.0	**80.5**
10	63.9	65.3	78.8	82.1	78.0	76.3	84.5	85.9	81.5	**86.1**	85.2	85.7	85.1	85.7
20	75.1	N/A	88.5	90.5	86.1	85.3	91.9	92.4	89.8	**92.8**	92.3	**92.8**	92.1	91.8
PUR	N/A	N/A	9.7	45.1	43.2	40.1	49.3	53.5	47.6	53.8	62.3	**63.6**	62.9	63.2

our training set is the smallest and hardest with least number in both images and individuals.

4.1.3. Experimental Results. Table 2 and Figure 5 show the average results of our 4 *CDPM* methods and several other methods. The highest CMC scores at every rank and highest PUR value are highlighted in bold type. For fair comparison, we use the results provided by the authors or by corresponding cited papers, under the single-shot experiment setting with 95 test individuals. Results in the first and second columns in Table 2 are from Wang et al. [12] and Hu et al. [7], respectively, and results of the third to tenth columns come from the fantastic work reported in literature [6]. These results are the best reported results of 3DPeS in the existing literatures [6, 7, 12, 16, 21, 28]. Table 2 shows that all 4 of our *CDPM* methods outperform the other methods at rank 1 and rank 5 CMC scores, achieving at least 13.7% and 2.2% improvement, respectively. Meanwhile, our *CDPMp* achieves 14.6% and 2.8% improvement at rank 1 and rank 5 CMC scores, respectively. At rank 10 CMC score, REV performs best with 0.4% to 1% improvement compared to our 4 methods. As for rank 20 CMC score, our *CDPM* and REV achieve the highest figure, 92.8%. Note that REV is a fusion solution of kLFDA, rPCCA, and MFA and is more complicated and computational than any of them. Although REV is slightly better than *CDPMp* in rank 10 CMC score, our methods have higher PUR scores than those of methods reported by Xiong et al. [6]. Among PUR scores, *CDPMp* possesses the highest score, with at least 9.8% improvement over the PUR scores reported in the other 8 methods. Under the consideration of both CMC and PUR, our *CDPMp* has the best performance among all methods in Table 2.

In Figure 5, our methods show great advantage with rank 1 and rank 2 matching accuracies. The other 10 methods extracted features from dense overlapping regions containing both background and person information and do not alter their feature selection strategies automatically according to different poses and different body types between test and reference pedestrian images. This means they are unable to properly filter out background features in matching phase when significant pose and body type variations occur. Our methods possess great tolerance for different poses and body types because our *CDP* feature model has already filtered out most background pixels with silhouette images before features extractions. Furthermore, it projects the extracted features into a uniform space for similarity matching. The slopes of our methods in Figure 5 decrease faster than those of other methods from rank 2 to rank 10, which means the growth of matching accuracies of our methods in top 10 ranks is smaller than those of the other methods. Normally, the slope of CMC curve decreases faster when the rank 1 accuracy is higher due to smaller growth of CMC scores in subsequent ranks. Despite lack of competitiveness after rank 5, our 4 methods remain in top ones, still holding some of the best reported results.

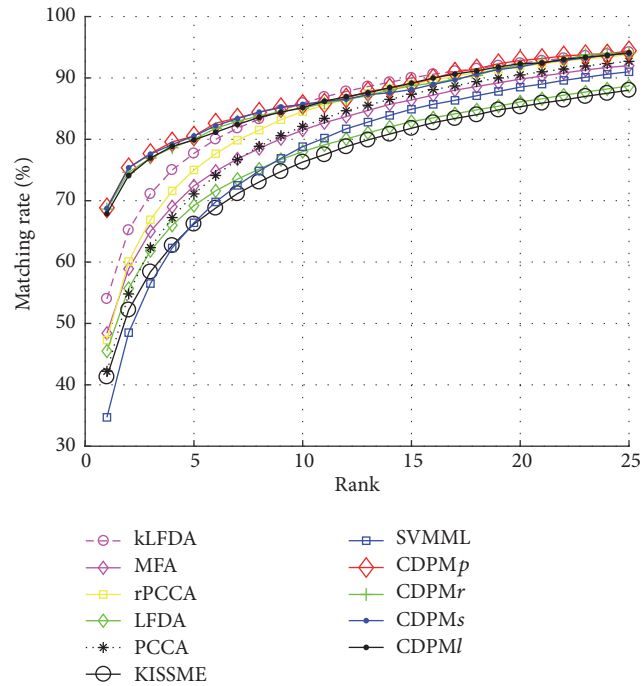

Figure 5: CMC curves on 3DPeS dataset.

The cAMT-DCA and DSTML methods are designed under the circumstance of learning a similarity classification metric for the target scenario from another existing annotated dataset, since manually annotating a large dataset of pairwise pedestrian images is costly and impractical for efficiently deploying a Re-ID system to a completely new scenario [12]. Even though the results of these 2 methods are the lowest at rank 1, rank 5, and rank 10, they are still good trials in transfer metric learning.

Our *CDPM* provides a practical alternative to automatically collecting pedestrian images from video sequences and uses Pure-vs-Exception strategy to group these images for similarity classification metric learning. The experimental results of our methods on 3DPeS dataset show the effectiveness of our similarity classification metric learning.

Table 3: PUR and CMC scores on IPR dataset with $P = 25$ test individuals.

r	Reference or model: 3 images for each individual					Reference: 2 images for each individual				Reference: 1 image for each individual			
	SARC3D	*CDPMl*	*CDPMp*	*CDPMr*	*CDPMs*	*CDPMl*	*CDPMp*	*CDPMr*	*CDPMs*	*CDPMl*	*CDPMp*	*CDPMr*	*CDPMs*
1	76	87.5	**89.5**	88	89	84.3	**86.8**	86.2	86.3	76.8	79.7	79	**79.8**
2	85	**95**	**95**	94.5	**95**	92.7	**94.3**	94	93.8	88.8	**90.8**	89.8	**90.8**
5	93	**99.5**	**99.5**	**99.5**	**99.5**	98.2	**98.3**	98.2	**98.3**	97.5	98.2	97.7	**98.3**
10	98[a]	**99.5**	**99.5**	**99.5**	**99.5**	98.8	99.2	98.8	**99.3**	98.7	**98.8**	98	98.7
PUR	68.8[b]	84.0	**85.9**	84.1	85.4	79.6	**81.8**	81.3	81.7	72.2	75	74.1	**75.5**

[a]The accuracy of rank 10 in SARC3D is read from Figure 6 in [18]. [b]The PUR score is calculated using figures read from Figure 6 in [18].

Our 4 methods, *CDPMl*, *CDPMp*, *CDPMr*, and *CDPMs*, corresponding to 4 different kernels had close results on 3DPeS dataset. Although *CDPMp* and *CDPMs* are very close, *CDPMp* is slightly better than *CDPMs*. *CDPMr* is third best, and *CDPMl* is last. Despite these rankings, such close results are due to more precise and stable features provided in our *CDP* feature model by filtering out most background information.

4.2. Experiment on IPR Dataset

4.2.1. Dataset and Experimental Protocol. IPR dataset is smaller than 3DPeS dataset and does not contain many illumination variations among images of the same individual. However, it is a suitable dataset to evaluate the robustness of viewpoint variations due to its specific settings. IPR dataset is a specialized dataset that contains 50 individuals with 4 specific images for each individual, resulting in a total of 200 pedestrian images. There are 4 different viewpoints for each individual, as shown in Figure 6: front view, back view, left view, and right view. These kinds of variations are challenging for methods based on dense blocks over fixed resolutions because of their asymmetric and nonuniform distribution of body. Furthermore, due to its specialization and small quantity of images, IPR dataset is not well known to the public. Thus, we only compared our *CDPM* method with SARC3D [18], which is intrinsically independent of viewpoint variations as mentioned in [18], and evaluated *CDPM*'s capability of dealing with these viewpoint variations.

We employ 3 different test settings with 1, 2, and 3 images of each individual to form the reference set but only 1 image of each individual to form the test set. We replicate 24, 24, and 8 partitions for these 3 experimental settings, respectively, by choosing different images forming different reference and test sets. Half partitions come from the first 25 individuals, and the other half come from the last 25 individuals. The reference set with 3 images is the easiest setting for testing, while the reference set with 1 image is the hardest for testing. Since SARC3D method requires 3 images to generate feature model, SARC3D is only evaluated in the reference setting with 3 images. Conversely, our methods are presented in all 3 reference settings.

4.2.2. Features and Implementation Details. We evaluate *CDPMl*, *CDPMp*, *CDPMr*, and *CDPMs* with the same model setting in 3DPeS on IPR dataset, since IPR dataset is also generated from ViSOR sequences.

Figure 6: Sample images from IPR dataset. Rows from up to bottom are front views, back views, left views, and right views, respectively.

SARC3D method uses 3 images to generate features through a graphical 3D body model with a selected set of vertices. Note that the 3D body model is learned prior to extract feature, which is an exclusive and advantageous setting. SARC3D extracts a 16-bin color histogram from HSV with 8 bins for H channel and 4 bins each for S and V channels, over 10×10 pixel blocks for each vertex. Besides color histogram, location, mean color, optical reliability, and saliency are also extracted for each vertex. Then, global similarity distance D_{H} and local similarity distance D_{S} are computed from weighted average of vertex-wise distances by using different weighting strategies. During matching, the final distance measure D_{HS} is computed by the product of D_{H} and D_{S}.

4.2.3. Experimental Results. Averaged results are shown in Table 3 and Figure 7. The highest CMC and PUR scores

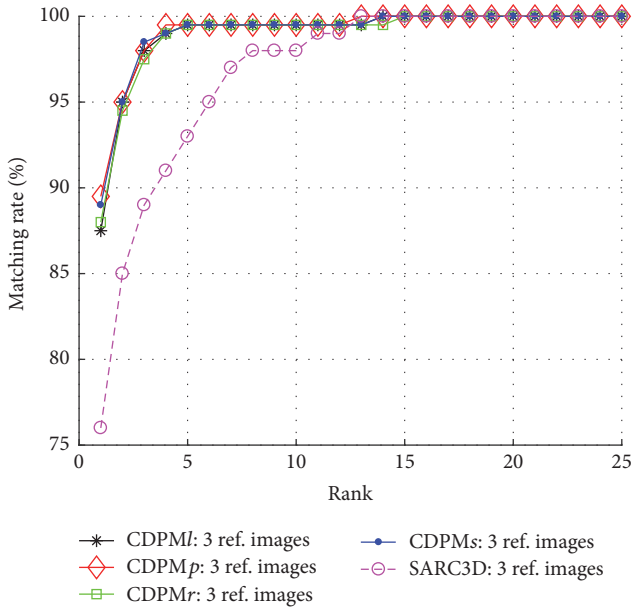

Figure 7: CMC curves on IPR dataset.

of each reference setting are highlighted in bold type in Table 3. For better observation, we only draw results of the first reference setting in Figure 7. With less test individuals and less illuminations variations, CMC scores are much higher on IPR dataset than those on 3DPeS dataset. However, IPR dataset still possesses challenges on viewpoint and pose variations, as shown in Figure 6.

In the first reference setting of Table 3, SARC3D's results represent the best setting in [18], and PUR score is calculated using reading figures with maximum error less than 0.2%. All 4 of our methods outperform SARC3D. In particular, *CDPMp* achieves 13.5%, 10%, 6.5%, and 1.5% improvements at rank 1, rank 2, rank 5, and rank 10 CMC scores, respectively, while PUR score improvement reaches 17.1%. The higher CMC and PUR scores of our *CDPM* methods on this reference setting can be explained in 3 aspects. Firstly, test set is small with only 25 individuals, and pedestrian images are well segmented with nearly no background information. Meanwhile, there is no illumination variance among the pedestrian images of the same individual. Secondly, our *CDPM* methods are free from background influences caused by pose and body type variations, especially for the well-segmented pedestrian images. Thirdly, our *CDPM* methods use similarity scores computed from 3 reference images to evaluate the final similarity measure for classification.

In the second reference setting, compared to SARC3D, improvements of *CDPMp* at rank 1, rank 2, rank 5, and rank 10 CMC scores are 10.8%, 9.3%, 5.3%, and 1.2%, respectively, and PUR score improvement is 13%. In the third reference setting, improvements of *CDPMs* are 3.8%, 5.8%, 5.3%, 0.7%, and 6.7%, respectively. All our methods in these 2 reference settings outperform SARC3D in the first reference setting according to CMC and PUR scores. The results on these two harder settings further prove that our *CDPM* methods have better tolerance for pose and viewpoint variations in IPR dataset than SARC3D. Furthermore, our experimental results prove the effectiveness of our metric learning strategy under the same domain.

Through these 3 reference settings, our experimental results show that our *CDPM* methods achieve significant improvements with multireference images. The behaviors of our 4 methods on IPR dataset are very similar to those on 3DPeS dataset. But the lower number of test individuals, well-segmented silhouette images, and multiple reference images not only make our 4 methods more different at rank 1 CMC scores and PUR scores but also swap the rankings of *CDPMs* and *CDPMp* in the hardest setting (with 1 reference image).

4.3. Analysis across Tables. Test results from Table 1 in distance metric learning are subject to a classification problem, while the experiment results from Tables 2 and 3 on 3DPeS and IPR datasets are subject to a ranking problem. Thus, the rankings of our 4 methods in Table 1 are different from those in Tables 2 and 3. However, there is a link between Table 1 and the other tables. From the first column in Table 1, the sums of PM and NM for *CDPMl*, *CDPMp*, *CDPMr*, and *CDPMs* are 12.79%, 10.96%, 11.99%, and 11.15%, respectively, and the corresponding mean variances of PM and NM are 3.44, 0.16, 0.78, and 0.18 with a scale of 10^{-4}, respectively. The ascending orders of the sums and mean variances are consistent with the rankings on 3DPeS and IPR datasets. Thus, under identical or very close ACs, *CDPM* tended to perform better on the Re-ID task with both less sum and less mean variance of PM and NM. This explains why *CDPMp* had better performance than the other 3 *CDPM* methods. Additionally, the close ACs, sums, and mean variances of PN and NM between *CDPMp* and *CDPMs* explain the close results between *CDPMp* and *CDPMs* on 3DPeS and IPR datasets.

5. Conclusion

We have proposed a *CDP* feature model and evaluated the performances of our Re-ID *CDPM* methods on 3DPeS and IPR datasets. Experimental results show that our *CDPM* methods have better performance compared to other state-of-the-art approaches and possess great tolerance for pose and viewpoint variations. In addition, we provide an effective similarity classification metric learning strategy for our *CDP* feature model to maximize interclass differences and suppress intraclass differences. Although our *CDP* feature model relies on silhouette images, which can be automatically extracted from video sequences or provided, it is compatible with varied image resolutions and is free of resizing distortion.

However, due to strong reliance on silhouette images, unfiltered background pixels can directly affect the classification results in our *CDPM* method. Thus, we intend to improve segmentations of the silhouette images provided or autoextracted, by filtering out more background pixels with less loss of person pixels in the future. In particular, we would like to automatically filter out shadow pixels cast by sunlight. In addition, we will also focus on finding integrable texture features and enriching the feature dimension of *CDP* to enhance robustness for significant illumination variations. By applying our method to surveillance systems in future work, we will find a more reliable matching strategy to exploit

CDP features in the similarity classification metric learning and multireference matching.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: a survey," *ACM Computing Surveys*, vol. 46, no. 2, 2013.

[2] Z. Sheng, H. Wang, C. Yin, X. Hu, S. Yang, and V. C. M. Leung, "Lightweight Management of Resource-Constrained Sensor Devices in Internet of Things," *IEEE Internet of Things Journal*, vol. 2, no. 5, pp. 402–411, 2015.

[3] C. Chih-Chung and L. Chih-Jen, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.

[4] L. Michel, P. Marc, and B. Robert, "Vip: Vision tool for comparing images of people," in *Proceedings of the 16th International Conference on Vision Interface*, pp. 35–42, June 2003.

[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 2197–2206, Mass, USA, June 2015.

[6] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8695, no. 7, pp. 1–16, 2014.

[7] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5576–5588, 2016.

[8] Z. Wang, R. Hu, C. Liang et al., "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.

[9] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3656–3670, 2014.

[10] D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li, "Person re-identification by regularized smoothing kiss metric learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 10, pp. 1675–1685, 2013.

[11] L. An, M. Kafai, S. Yang, and B. Bhanu, "Person Reidentification with Reference Descriptor," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 776–787, 2016.

[12] X. Wang, W.-S. Zheng, X. Li, and J. Zhang, "Cross-scenario transfer person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 8, pp. 1447–1460, 2016.

[13] Z. Shi, T. M. Hospedales, and T. Xiang, "Transferring a semantic representation for person re-identification and search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 4184–4193, USA, June 2015.

[14] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 3908–3916, USA, June 2015.

[15] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2288–2295, June 2012.

[16] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 3318–3325, USA, June 2013.

[17] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4766–4779, 2015.

[18] D. Baltieri, R. Vezzani, and R. Cucchiara, "SARC3D: a new 3D body model for people tracking and re-identification," in *Proceedings of the 16th Conference on Analysis and Processing: Part I (ICIAP '11)*, vol. 6978 of *Lecture Notes in Computer Science*, pp. 197–206, Springer, Heidelberg, Germany, 2011.

[19] J. Chen, Y. Wang, and Y. Y. Tang, "Person Re-identification by Exploiting Spatio-Temporal Cues and Multi-view Metric Learning," *IEEE Signal Processing Letters*, vol. 23, no. 7, pp. 998–1002, 2016.

[20] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1565–1573, USA, June 2015.

[21] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in *Proceedings of the the 2011 joint ACM workshop on Human Gesture and Behavior Understanding (J-HGBU '11)*, pp. 59–64, NY, USA, December 2011.

[22] http://imagelab.ing.unimore.it/visor/index.asp.

[23] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the 10th European Conference on Computer Vision: Part I (ECCV '08)*, vol. 5302 of *Lecture Notes in Computer Science*, pp. 262–275, Springer, Berlin, Germany, 2008.

[24] W. S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the 20th British Machine Vision Conference (BMVC' 09)*, September 2009.

[25] D. C. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proceedings of the 22nd British Machine Vision Conference, BMVC' 11*, pp. 1–11, September 2011.

[26] A. Mignon and F. Jurie, "PCCA: a new approach for distance learning from sparse pairwise constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2666–2672, June 2012.

[27] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 3610–3617, USA, June 2013.

[28] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1846–1855, USA, June 2015.