View-Adaptive Metric Learning for Multi-view Person Re-identification

Canxiang Yan[1,2], Shiguang Shan[1], Dan Wang[2,1], Hao Li[1,2], Xilin Chen[1]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences(CAS),Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
canxiang.yan@vipl.ict.ac.cn; sgshan@ict.ac.cn; dan.wang@vipl.ict.ac.cn; hao.li.ict@gmail.com; xlchen@ict.ac.cn

Abstract. Person re-identification is a challenging problem due to drastic variations in viewpoint, illumination and pose. Most previous works on metric learning learn a global distance metric to handle those variations. Different from them, we propose a view-adaptive metric learning (VAML) method, which adopts different metrics adaptively for different image pairs under varying views. Specifically, given a pair of images (or features extracted), VAML firstly estimates their view vectors (consisting of probabilities belonging to each view) respectively, and then adaptively generates a specific metric for these two images. To better achieve this goal, we elaborately encode the automatically estimated view vector into an augmented representation of the input feature, with which the distance can be analytically learned and simply computed. Furthermore, we also contribute a new large-scale multi-view pedestrian dataset containing 1000 subjects and 8 kinds of view-angles. Extensive experiments show that the proposed method achieves state-of-the-art performance on the public VIPeR dataset and the new dataset.

1 Introduction

Person re-identification is the technique to identify an individual across spatially disjoint cameras. It is believed to have deep potential applications such as suspect tracking and lost children finding in next-generation intelligent video surveillance. With the ever growing requirements in public security, such techniques are becoming more and more urgently required in order to automatically locate and track wanted persons, or at least dramatically reduce the workload of human operators checking the large-scale recorded surveillance videos.

However, even if it is assumed that the person does not change clothes across the network of cameras, person re-identification suffers from two technical difficulties: first, the appearance of the same person can vary dramatically in different cameras because of both intrinsic and extrinsic variations, including poses, lighting (especially in outdoor scenario), viewpoints, etc. The second difficulty is that there might be a large number of similar individuals, such

as, many people wearing dark coats of similar color in winter. Essentially, these two difficulties can be cast to the general pattern recognition challenges: large within-class variations and small between-class variations.

Because of the above-mentioned application values and theoretical challenges, person re-identification has attracted more and more research efforts in recent years. Similar to most methods for pattern recognition problems, existing technologies for person re-identification either seek good features or pursue good distance metrics. Previous methods [1–11] seeking good features attempt to extract features that are not only robust to variations, but are also discriminative for different persons. Gray and Tao [7] proposed a boosting-based approach to find the best feature representation for the viewpoint invariant person recognition. However, such selection may not be globally optimal because features are selected independently from the original feature space in which different classes can be heavily overlapped. In [2], co-occurrence metric is used to capture the spatial structure of the colors in each divided region. Farenzena et al. [3] proposed three localized features under symmetric-driven principles to achieve the robustness to pose, viewpoint and illumination variations. Cheng et al. [6] adopted a part-based model to handle pose variation. However, It is not flexible enough and has strong dependence on the performance of the pose estimators. More recently, Zhao et al. [11] proposed a saliency matching method, which used patch saliency to find the distinctive local patches and recognized same persons by minimizing the salience matching cost. These handcrafted appearance descriptors mostly worked on person matching from close views, but it is not necessarily true for large viewpoint variations, e.g., front view vs. back view. Directly feature matching in corresponding region may derive false distance when existing large view gap.

In contrast to the above feature extraction method, metric learning emphasizes the similarity/dissimilarity measurement, given a pair of images or features extracted using above methods. For instance, LMNN [12] learned a distance metric for kNN classification with the goal that k-nearest neighbors are from the same class as that of input one while instances from different classes should be separated by a large margin. Davis et al. [13] formulated metric learning problem as that of minimizing the differential relative entropy between two multivariate Gaussians distance distribution under the given constraints on the distance function. They integrated a regularization step to avoid overfitting. Zheng et al. [14] proposed Relative Distance Comparison (RDC) to deal with large appearance changes. In their model, the likelihood of image pairs of the same person having relatively smaller distance than that of different persons is maximized to obtain the optimal similarity measure. Recently, Köstinger [15] proposed a simple strategy to learn a distance metric from equivalence constraints, based on a statistical inference perspective. Pedagadi et al.[16] proposed a supervised dimensionality reduction method based on Fisher Discriminant Analysis. Li et al.[17] learned a decision function with locally adaptive thresholding rule to deal with appearance variations.

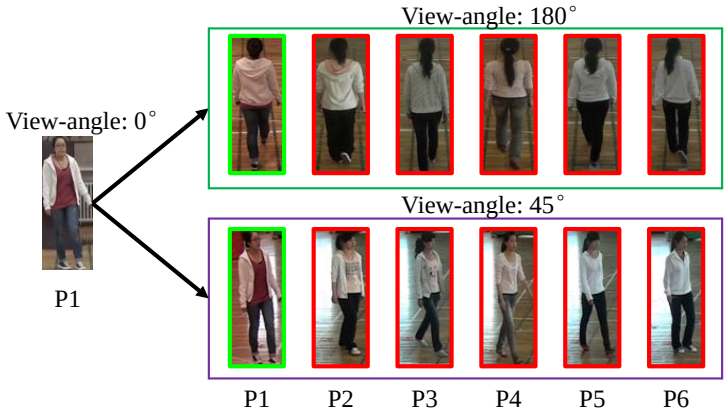


Fig. 1. An example of multi-view person re-identification. Given a probe image with view-angle 0° and two gallery sets with view-angle 45° and 180° respectively, it is easier to find the target image (with green bounding box) from the gallery set with view-angle 45° because more common appearances are shared between images with smaller view gap.

All above metric learning methods learn a global distance metric for matching across different views. However, since the discriminatory power of the input features might vary between different image pairs under varying views, learning a global metric cannot fit well the distance over the multi-view image pairs. As shown in Figure 1, given a probe image with view-angle 0°, it is more difficult to recognize its target image with larger view gap (180°) than that with smaller view gap (45°). Thus, it is necessary and reasonable to learn different metrics for different view pairs. For instance, the Multi-view CCA (MCCA) [18] obtains one common space for multiple views. In MCCA, several view-specific transforms, each for one person view-angle, are obtained by maximizing total correlations between any pair of views. However, it not only neglects discriminant information when training but also needs to know view-angle of each image when testing.

To explicitly address the multi-view person re-identification problem, we propose a view-adaptive metric learning (VAML) method. Different from traditional metric learning methods, VAML adopts different metrics adaptively for different image pairs, according to their views. Specifically, given a pair of images (or features extracted), VAML firstly estimates their view vectors (consisting of probabilities belonging to each view) respectively, and then adaptively generates a specific metric for these two images. To learn single unified discriminant common space, the view vector is encoded into an augmented representation of the input feature. Then, all the view-specific metrics are jointly optimized by maximizing between-class variations while minimizing within-class variations from both inter-view and intra-view. This optimization problem can be solved analytically by using generalized eigenvalue decomposition. Extensive

comparisons to state of the art methods on VIPeR dataset [19] show that the proposed method achieves better performance. To advance the multi-view person re-identification problem, we further collect a new large-scale **M**ulti-**V**iew pedestrian dataset (MV), simulating video surveillance scenario. In this dataset, there are 1000 subjects, each with 8 discrete view-angles quantified from the full range of 360°. To our best knowledge, this dataset is the largest one of the same type (at least in terms of the number of persons and view-angles). On this new dataset, the proposed VAML achieves higher performance than the state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 describes the proposed approach. Section 3 introduces the MV dataset. Experiments on VIPeR and MV dataset are presented in Section 4. Finally, we conclude and summarize the paper in Section 5.

2 View-Adaptive Metric Learning

We define the multi-view person re-identification problem as follows: suppose we have a gallery set G consisting of N persons, each has one or multiple images captured from any of V views. Given a probe image $\mathbf{x}_i$, our goal is to find the image of the same person in a different view from G. To make images with different views comparable, we assume that there is a common metric space. In this common space, the matching of all the image pairs can be done by applying a view-adaptive Mahalanobis metric, which is learned to maximize between-class variation while minimizing within-class variation, as shown in Figure 2. To better achieve this goal, we first extract feature and estimate view vector (consisting of the probabilities of the image belonging to each view) for any input image. Then, the estimated view vector is encoded into an augmented representation of the feature.

In the next, we first introduce the formulation of the VAML. Then, describe the process of feature augmentation in detail. Finally, we describe how to learn the metric analytically.

2.1 Formulation

Recently, metric learning methods [15, 20] have been proposed for person re-identification. They learn a global distance metric for image matching across different views. However, since the discriminatory power of the image features varies a lot between different image pairs under varying views, learning a global metric cannot fit well the distance over the multi-view image pairs. Thus, the goal of this paper is to introduce a view-adaptive metric, which adopts different metrics adaptively for different image pairs and can be derived in the following.

The most widely used approach for metric learning is Mahalanobis distance learning. Given data points $\mathbf{x}_i$ and $\mathbf{x}_j \in \mathbb{R}^D$, Mahalanobis distance metric between the two data points is

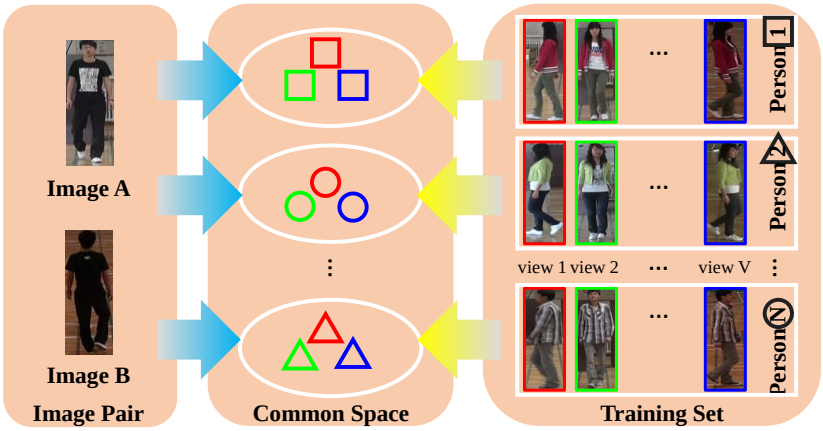


Fig. 2. The overview of VAML. Image pairs with different views are matched in a common metric space. In this common space, images in one class with different views are close to each other, while images in different classes with different views are far away from each other. According to the view information (e.g. view vector) of input image pair, a specific metric can be adaptively generated to measure the dissimilarity of the pair.

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) \tag{1}$$

where $\mathbf{M} \succeq 0$ is a positive semi-define matrix. $\mathbf{M}$ can also be decomposed to $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$. Then,

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{L}\mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j) = \left\| \mathbf{L}^\top (\mathbf{x}_i - \mathbf{x}_j) \right\|^2 \tag{2}$$

Note that Eq.(2) uses a global metric to match all image pairs with different views. Here we introduce a view-adaptive metric, which is adaptive to different image views. Suppose there are V views, a new distance between a pair of images is defined as the sum of Mahalabonis distances over all the views:

$$d_{mv}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{v=1}^{V} (\mathbf{x}_{iv} - \mathbf{x}_{jv})^\top \mathbf{M}_v (\mathbf{x}_{iv} - \mathbf{x}_{jv}) = \sum_{v=1}^{V} \left\| \mathbf{L}_v^\top (\mathbf{x}_{iv} - \mathbf{x}_{jv}) \right\|^2 \tag{3}$$

where $\mathbf{M}_v = \mathbf{L}_v \mathbf{L}_v^\top$ is positive semi-define and is the metric matrix for vth view; $\mathbf{x}_{iv}$ and $\mathbf{x}_{jv}$ are features under the vth view. However, it's hard to extract all the view-specific features $\{\mathbf{x}_{iv}\}_{v=1}^{V}$ from single image because only part of person appearances are visible. Instead, we introduce a view vector $\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{iv}, \ldots, p_{iV}]^\top$, where p_{iv} measures the ability of $\mathbf{x}_i$ to represent person appearance under the vth view, to weigh the image feature $\mathbf{x}_i$ and make

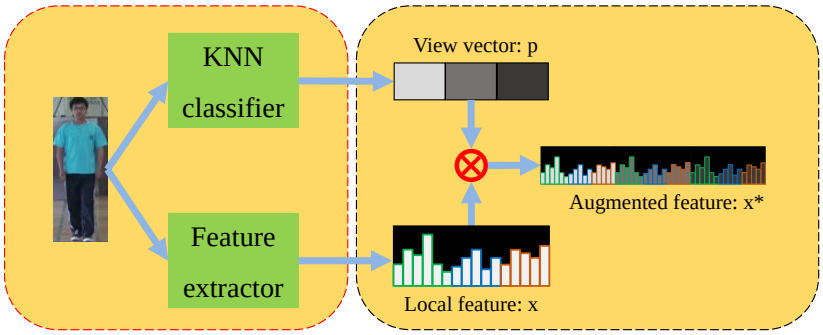


Fig. 3. Feature augmentation. The view vector of input image is estimated using the kNN-classifier. Then the view vector is encoded into an augmented representation of the image feature using Kronecker multiplication operation.

it view-specific. Thus, let $\mathbf{x}_{iv} = p_{iv}\mathbf{x}_i$, the Eq.(3) can be re-written as

$$d_{mv}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{v=1}^{V} \left\| \mathbf{L}_v^\top (p_{iv}\mathbf{x}_i - p_{jv}\mathbf{x}_j) \right\|^2 \tag{4}$$

By expanding Eq.(4), we can get the view-adaptive metric as follows:

$$\begin{aligned} d_{mv}(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{p}_i \otimes \mathbf{x}_i - \mathbf{p}_j \otimes \mathbf{x}_j)^\top \mathbf{M}_{mv} (\mathbf{p}_i \otimes \mathbf{x}_i - \mathbf{p}_j \otimes \mathbf{x}_j) \\ &= (\mathbf{x}_i^* - \mathbf{x}_j^*)^\top \mathbf{M}_{mv} (\mathbf{x}_i^* - \mathbf{x}_j^*) \end{aligned} \tag{5}$$

where $\mathbf{M}_{mv}$ is a new positive semi-define matrix, which contains V positive semi-define matrices; '$\otimes$' is a Kronecker multiplication operator, and $\mathbf{x}^* = \mathbf{p} \otimes \mathbf{x}$ is an augmented representation of the original feature $\mathbf{x}$ (see Section 2.2 for details). Thus, Eq.(5) is a kind of parameter metric learning methods. The view adaptive property is achieved by the parameter vector $\mathbf{p}_i$ and $\mathbf{p}_j$.

2.2 Feature Augmentation

Figure 3 illustrates the process of feature augmentation. To obtain the augmented feature, we extract texture and color features to generate $\mathbf{x}$ and use kNN-based view estimation to get the view vector $\mathbf{p}$. We describe the details of feature extraction and view estimation in the following:

Feature Extraction We use texture and color features to represent the input image. Haralick et al. [21] proposed gray level co-occurrence matrix (GLCM) as the distribution of co-ocurring values at a given offset vector (angle and distance) and extracted texture features based on it. In our method, we first separate the images into horizontal strips of size 8×48 and control the overlapping stride to be 4 in the vertical direction. Then local GLCMs are calculated from each strip with 4 offset vectors: $[0°; 1]$, $[45°; 1]$, $[90°; 1]$ and

[135°; 1]. We also calculate GLCMs between any two strips, which describe frequencies of co-ocurring color pair. For each GLCM, entropy and homogeneity [22] are used to generate the texture features. Then in each strip, HSV histogram is extracted from three color channels with (8,8,4) bins respectively. Finally, texture and color features extracted from all strips are concatenated to generate the representation $\mathbf{x}$ for the input image.

View Estimation We treat the view estimation problem as a multi-class classification problem by grouping data with different view-angles into different classes. Then kNN classifier is learned from a training set with labeled view and used to estimate the possibilities of input image belonging to each view in order to form the corresponding view vector $\mathbf{p}$. Considering that only part of person appearances are visible in single image, thus some entries corresponding to invisible views in the view vector are insignificant. To reduce the effects of those noises, we only keep few values by thresholding very small values in $\mathbf{p}$ to be zero. We set the threshold to be 0.2 empirically.

2.3 Metric Learning

Given a set of n training data points $\chi = \{\mathbf{x}_i\}_{i=1}^n$, and the corresponding class label $\mathcal{L} = \{l_i\}_{i=1}^n$, where $l_i \in \{1, 2, ..., C\}$, and corresponding binary view vector $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^n$, in which $\mathbf{p}_i$ consists of V binary values and only one of them that corresponds to the labeled view-angle is 1, we describe the process of view-adaptive metric learning in the following.

Denote that $\mathbf{L}_{mv} = [\mathbf{L}_1^\top \mathbf{L}_2^\top \cdots \mathbf{L}_V^\top]^\top$, $\mathbf{M}_{mv}$ in Eq.(5) can be decomposed to $\mathbf{M}_{mv} = \mathbf{L}_{mv}\mathbf{L}_{mv}^\top$. Thus, Eq.(5) can be equivalently written as:

$$\begin{aligned} d_{mv}(\mathbf{x}_i, \mathbf{x}_j) &= (\mathbf{x}_i^* - \mathbf{x}_j^*)^\top \mathbf{L}_{mv}\mathbf{L}_{mv}^\top (\mathbf{x}_i^* - \mathbf{x}_j^*) \\ &= \left\| \mathbf{L}_{mv}^\top \mathbf{x}_i^* - \mathbf{L}_{mv}^\top \mathbf{x}_j^* \right\|^2 \end{aligned} \tag{6}$$

Based on the derivation of view-adaptive Mahalanobis distance above, we then define our objective function by considering two aspects in the new metric space: the separability of distances between images from different classes and the compactness of distances between images from the same class.

The separability, which describes the between-class variation, is defined as

$$J_S = \sum_{i=1}^{C} \frac{n_i}{n} d_{mv}(\boldsymbol{\mu}_i, \boldsymbol{\mu}) = \mathrm{Tr}(\mathbf{L}_{mv}^\top \mathbf{S}_b \mathbf{L}_{mv}), \tag{7}$$

where $\boldsymbol{\mu}_i$ and n_i are the mean and the number of the data points belonging to the *ith* class, and $\boldsymbol{\mu}$ is the mean of all the data points in the transformed space $\mathbf{L}_{mv}$. $\mathrm{Tr}(\cdot)$ is trace operator; $S_b = \sum_{i=1}^{C}(n_i)/(n)(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top$ is the between-class covariance matrix.

From another aspect, we use compactness to represent the intra-class variation. Let J_C denoted the compactness, which can be calculated as the sum

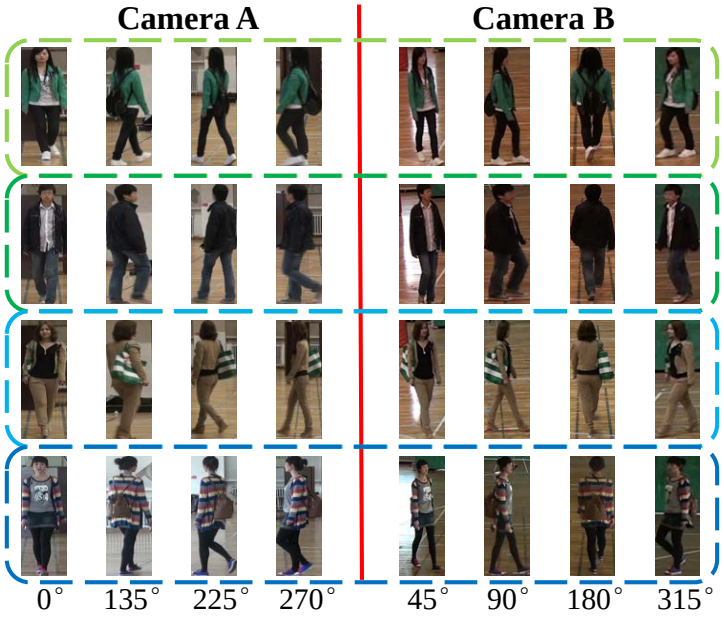


Fig. 4. Examples of multi-view images from MV dataset. Pose, viewpoint and illuminance variations can be observed across camera views.

of distances of images from the same class:

$$J_C = \sum_{i=1}^{n} \frac{n_{l_i}}{n} d_{\mathbf{L}}(\mathbf{x}_i^*, \boldsymbol{\mu}_{l_i}) = \mathrm{Tr}(\mathbf{L}_{mv}^\top \mathbf{S}_w \mathbf{L}_{mv}), \tag{8}$$

where the within-class covariance matrix $S_w = \sum_{i=1}^{n}(n_{l_i})/(n)(\mathbf{x}_i^* - \boldsymbol{\mu}_{l_i})(\mathbf{x}_i^* - \boldsymbol{\mu}_{l_i})^\top$.

To obtain the optimal $\mathbf{L}_{mv}$, the following objective function should be maximized:

$$\begin{aligned} \mathbf{L}_{mv}^* &= \underset{\mathbf{L}_{mv}}{\mathrm{argmax}}\, \mathrm{Tr}(\mathbf{L}_{mv}^\top \mathbf{S}_b \mathbf{L}_{mv}) \\ s.t. \quad & \mathbf{L}_{mv}^\top \mathbf{S}_w \mathbf{L}_{mv} = \mathbf{I} \end{aligned} \tag{9}$$

This problem can be efficiently solved by generalized eigenvalue decomposition $\mathbf{S}_b\theta_k = \beta_k \mathbf{S}_w \theta_k$,where β_k is the kth largest generalized eigenvalue. The matrix $\mathbf{L}_{mv}^*$ is then constituted of the corresponding eigenvalues θ_k, $k = 1, 2, \ldots, d$.

3 MV Dataset and Evaluation Protocol

MV is a new multi-view pedestrian dataset we constructed for the research on the multi-view person re-identification problem. To our best knowledge, MV is the largest dataset in terms of the number of persons and annotated view-angles. The following subsection will describe the construction and evaluation protocol of the MV dataset.

3.1 Construction of MV Dataset

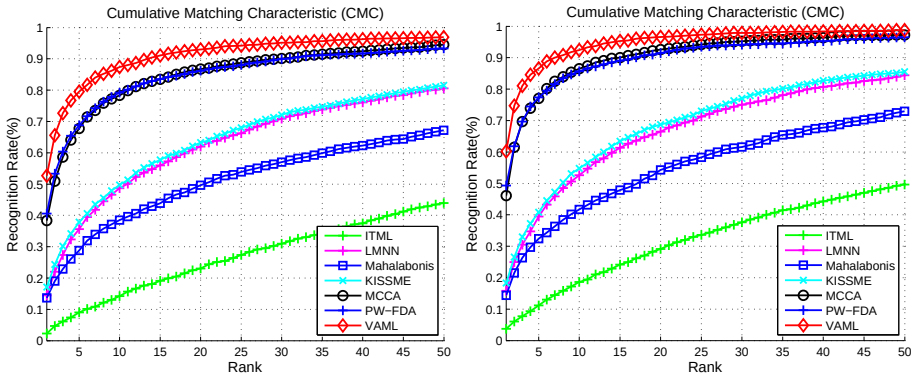
The dataset is collected from two HD (1920×1080) cameras in different locations of a sport square. 1000 participants from the local university or residents walk along the same 'S'-type route in the sport square. We record video clips for each person using the two cameras simultaneously. Then we perform background substraction [23] to locate the person. With estimation of the walking direction at each location, we can get full range of views. We quantify the range and define 8 discrete view-angles: $0°$, $45°$, $90°$, $135°$, $180°$, $225°$, $270°$ and $315°$. To cover the variations of an individual, we sampled 5 frames for each person in each of the 8 different view-angles and cropped them out from background with resolution of 48×128. Fig. 4 shows examples with different view-angles from the two cameras.

3.2 Evaluation Protocol

To allow consistent comparison of different methods, we define a standard evaluation protocol about dataset splitting and evaluation. We randomly split the dataset into two sets of 500 persons each, one for training and one for testing. This process is carried out 10 times. For each splitting, there are two testing scenarios:

- S2S (single-shot vs single-shot). In the testing set, we select one view from **Camera A** as gallery set P and another view from **Camera B** as probe set G. Totally, there are 4 combinations of P and G: $(0°, 180°)$, $(135°, 315°)$, $(225°, 45°)$ and $(270°, 90°)$. In each combination, P and G both have size of 500 images, each represents a different individual and its corresponding view-angle is considered to be unknown. This is a general setting which can be found in [7, 24, 2].
- M2M (multi-shot vs multi-shot). The only difference from S2S scenario is that each person is described by multiple images in both gallery set G and probe set P, following previous work[3, 25]. In this scenario, the number of images of each person are set to 3.

There are several established evaluation methods for evaluating person re-identification system. Among them, cumulative matching characteristic (CMC) curve is used to indicate performance of various methods. In our evaluation protocol, we average the multi-view recognition results of the 4 combinations of probe set and gallery set to present **average CMC curve**. To measure the



(a) S2S on MV dataset (b) M2M on MV dataset

Fig. 5. Comparisons to metric learning method on MV dataset under S2S and M2M protocol using CMC curve. The rank-1 ARR of VAML is much higher than others. It indicates that VAML achieves state-of-the-art performance.

performance of multi-view person re-identification, we also propose **average recognition rate(ARR)** which results from averaging recognition rates of all combinations.

4 Experiments

We evaluate our approach on MV dataset and the public VIPeR dataset. The reason we choose VIPeR is that it is the most widely used dataset for evaluation and it provides most of the challenges faced in real-world person re-identification applications, e.g., viewpoint, pose, different background, illumination variation, low resolution, occlusions, etc. Experimental results are shown in terms of recognition rate, by the Cumulative Matching Characteristic (CMC) curve.

4.1 MV Dataset

We randomly split the dataset as described in the protocol. The color and texture features are extracted from images with resolution of 48×128. Then, we use PCA to reduce the feature dimension by keeping 90% energy for all the metric learning methods. For view estimation, we set the parameter k, the number of nearest neighbors in kNN Classifier, to be 100. This process repeats 10 times.

Cross-view experiments are conducted under the S2S and M2M scenarios. We select images from one view as probe set and images from another view as gallery set(totally 4 combinations). In Figure 5 we report the average CMC curve under S2S and M2M scenarios for LMNN [12], ITML[13], KISSME [15], MCCA [18], our method(VAML), the Mahalabonis distance of the similar pairs and pairwise Fisher Discriminant Analysis (PW-FDA) [26] as baseline. Note that PW-FDA and MCCA use the view-angle of input image when testing. It

Table 1. Rank-1 recognition rates in % on MV dataset under S2S and M2M scenarios respectively.

(a) Rank-1 recognition rate in % on MV dataset under S2S scenario

Method	0° ⇒ 180°	135° ⇒ 315°	225° ⇒ 45°	270° ⇒ 90°	ARR
VAML	**53.4**	**48.5**	**47.1**	**55.3**	**51.1**
MCCA[18]	40.6	36.4	33.4	43.0	38.4
PW-FDA	39.8	37.2	39.8	45.4	40.6
KISSME[15]	22.2	16.4	15.0	14.8	17.1
LMNN[12]	18.2	13.2	14.6	11.8	14.5
ITML[13]	2.0	1.8	2.0	3.4	2.3
Mahalabonis	16.8	11.2	15.6	11.0	13.7
SDALF[3]	7.4	5.2	5.8	6.2	6.2
LLADF[17]	21.0	12.2	13.2	17.8	16.1
LFDA[16]	18.2	11.2	11.6	18.0	14.8
SM[11]	9.2	5.0	6.2	8.5	7.2
P-VAML	53.9	47.9	47.2	55.0	51.0

(b) Rank-1 recognition rate in % on MV dataset under M2M scenario

Method	0° ⇒ 180°	135° ⇒ 315°	225° ⇒ 45°	270° ⇒ 90°	ARR
VAML	**60.6**	**57.4**	**58.0**	**65.0**	**60.3**
MCCA[18]	49.2	42.2	41.8	51.4	46.2
PW-FDA	48.8	45.4	49.0	54.0	49.3
KISSME[15]	21.8	16.6	20.0	14.8	18.3
LMNN[12]	20.2	14.8	16.6	11.8	15.9
ITML[13]	4.6	3.2	3.4	3.6	3.7
Mahalabonis	19.2	11.0	17.0	10.4	14.4
SDALF[3]	8.8	5.4	7.8	7.2	7.3
P-VAML	61.2	56.8	58.2	64.6	60.2

is obvious that using the proposed VAML metric leads to a large performance gain over traditional metric learning methods and that VAML also outperforms the two methods using labeled view.

Moreover, in Table 1 we show the result of rank-1 recognition rate on each cross-view combination. It can be seen that our VAML under two scenarios is significantly better than other methods reported results on MV dataset. Specifically under S2S scenario, rank-1 ARR is 51.1% for VAML, versus 38.4% for MCCA [18], 16.1% for LLADF[17], 14.8% for LFDA[16], 7.2% for SM[11], 17.1% for KISSME [15], 14.5% for LMNN [12], 2.3% for ITML [13], 13.7% for Mahalabonis and 6.2% for SDALF[3]. In particular, VAML outperforms the rank-1 ARR of the second best PW-FDA [26] by 10.5%. This improvement is due to our view-adaptive strategy, which can make full use of multi-view information from the same class and is robust to viewpoint change and pose variations. In M2M, set-to-set distance is introduced because each person in probe and gallery set contains 3 images. To recognize one person, we compare

the distance of each possible pair from different persons, associating the person to the one from gallery set with lowest distance. In Table 1(b), all the methods have performance gain under M2M compared to that under S2S. Specifically, rank-1 ARR of VAML is improved by 9.2%, versus 7.8% for MCCA [18] and 8.7% for PW-FDA [26]. In particular, the rank-1 ARR difference between our method and KISSME is increased to 42.0%. By comparing the results of all methods on different combinations, it also can be observed that (135°,315°) and (225°,45°) are the most challenge combinations because worst performance of most methods are reported on them.

Based on this analysis, VAML outperforms all other metric learning methods significantly. The main reason for VAML to obtain the best performance is that latent relationship between different views of the same person is learned successfully and robustness to large viewpoint variations is improved by exploiting the view-adaptive metric.

4.2 VIPeR Dataset

VIPeR [19] is the first publicly available dataset for person re-identification consisting of 632 people captured outdoor with two images for each person with size at 128×48 pixels. The biggest challenges in VIPeR are viewpoint and illuminance variations, which may cause the change of appearance largely. For each person, corresponding individuals have viewpoint change up to 90 degrees.

Our setting for the splitting of training/testing set is same to SDALF[3], by which VIPeR dataset is splitted into two set with equal size (316 persons), one as training set and another as testing set. Then we estimate the view vector of each image by kNN classifier using 20,000 images from MV dataset as training data. Different from the setting of MV, only three most dominant estimated view-angles are kept. Thus, the view vector of each data in VIPeR have only 3 dimensions. The whole evaluation procedure is carried out 10 times.

We compare the performance of proposed VAML in the ranging of first 50 ranks to various state of the arts, as illustrated in Table 2. It is noted that our method outperforms all other appearance-based methods. Specifically, SM[11] achieves second best results compared to the other appearance-based methods, like CPS[6], SDALF [3], ELF [7], DDC [25] and ERSVM [8]. However its recognition rate is 1% lower than ours at rank-1 and have a difference of 11%, 10%, 5% at rank-10, rank-25 and rank-50 respectively. It shows that our method can handle the appearance variations caused by viewpoint change better than traditional appearance-based methods. Moreover, we also analyze the performance of popular metric learning methods[15, 12–14, 17, 16]. Our VAML has much better performance compared to LMNN[12], ITML[13], KISSME[15],LFDA[16] and RDC[14], and shows comparable performance with that of LLADF[17]. The performance of top-25 ranks is also represented with CMC curve in Figure 6.

The effect of view estimation. One factor affecting the performance is the accuracy of view estimation. The average accuracy of kNN-based view classification is 70% on VIPeR dataset and 90% on MV dataset repectively. In

Table 2. Recognition rates in [%] at different ranks r on VIPeR dataset.

Method	r=1	10	25	50
VAML	**26**	**63**	**82**	**92**
ELF[7]	12	43	66	81
SDALF[3]	20	49	70	83
CPS[6]	22	57	76	87
DDC[25]	19	52	69	80
ERSVM[8]	13	50	71	85
SM[11]	25	52	72	87
RDC[14]	16	54	76	87
KISSME[15]	21	60	81	92
LMNN[12]	19	53	74	87
ITML[13]	16	51	77	90
LLADF[17]	29	78	92	97
LFDA[16]	23	66	84	93
Mahalabonis	17	49	69	82
P-VAML	29	68	84	94

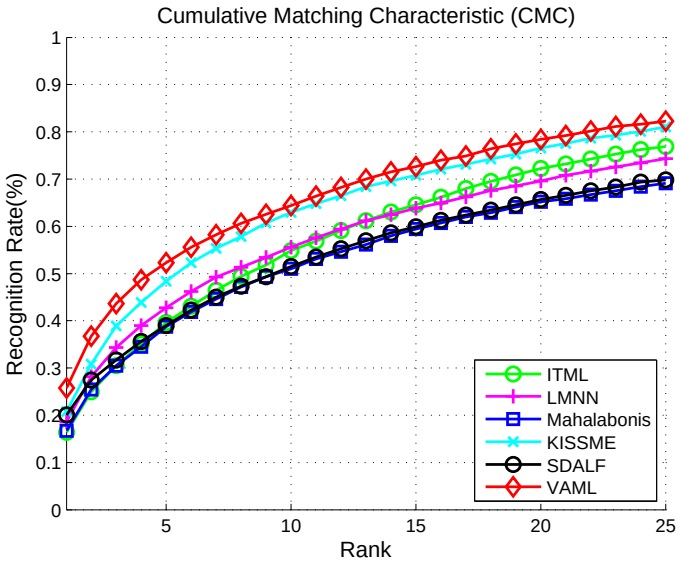


Fig. 6. Average Cumulative Matching Characteristic (CMC) curves of metric learning methods.

order to evaluate the contribution of view estimation, we propose a 'perfect' VAML (P-VAML), in which we use labeled view vector (100 % accuracy) to form the augmented feature. Table 1 and Table 2 show the results of P-VAML on VIPeR and MV datasets respectively. It is observed that better performance can be achieved along with the increasement of classification accuracy on VIPeR while the improvement is not obvious on MV dataset.

5 Conclusion

In this paper, we have proposed view-adaptive metric learning to learn a metric which can be adaptive to the views of matching pair for multi-view person re-identification. Both separability of instances from different classes and compactness of instances with different views from same class are exploited in our method. Meanwhile, a multi-view dataset MV, which consists of 1000 persons and has explicit annotation of view-angles, have been released with our expectation to advance the research of multi-view person re-identification. Compared with existing competitive methods, the extensive experiments show that our approach achieves the state-of-the-art results over MV and VIPeR datasets. In the future, we would like to develop the VAML by considering the symmetry of views.

Acknowledgement. The work is partially supported by Natural Science Foundation of China(NSFC) under contracts Nos. 61222211, 61272321, 61402430 and 61025010; and the China Postdoctoral Science Foundation 133366.

References

1. Gheissari, N., Sebastian, T., Hartley, R.: Person reidentification using spatiotemporal appearance. In: CVPR. (2006) 1528–1535
2. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: ICCV. (2007) 1–8
3. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR. (2010) 2360–2367
4. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: AVSS. (2011) 179–184
5. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: ICPR. (2010) 1413–1416
6. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: BMVC. (2011)
7. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensamble of localized features. In: ECCV. (2008) 262–275
8. Prosser, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: BMVC. (2010)
9. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: BMVC. (2012)

10. Schwartz, W., Davis, L.: Learning discriminative appearance-based models using partial least squares. In: SIBGRAPI. (2009) 322–329
11. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: ICCV. (2013) 2528–2535
12. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: NIPS. (2006)
13. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML. (2007) 209–216
14. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. PAMI (2013) 653–668
15. Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR. (2012)
16. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: CVPR. (2013) 3318–3325
17. Li, Z., Chang, S., Liang, F., Huang, T., Cao, L., Smith, J.: Learning locally-adaptive decision functions for person verification. In: CVPR. (2013) 3610–3617
18. Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. In: SiKDD. (2010) 1–4
19. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS. (2007)
20. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR. (2011)
21. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. SMC (1973) 610–621
22. Howarth, P., Rüger, S.: Evaluation of texture features for content-based image retrieval. In: IVR. (2004) 326–334
23. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: ICPR. (2004) 28–31
24. Lin, Z., Davis, L.: Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In: AVC. (2008) 23–24
25. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: IA. Lecture Notes in Computer Science (2011) 91–102
26. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. PAMI (1997) 711–720