Locally Aligned Feature Transforms across Views

Wei Li and Xiaogang Wang
Electronic Engineering Department
The Chinese University of Hong Kong, Shatin, Hong Kong
{liwei, xgwang}@ee.cuhk.edu.hk

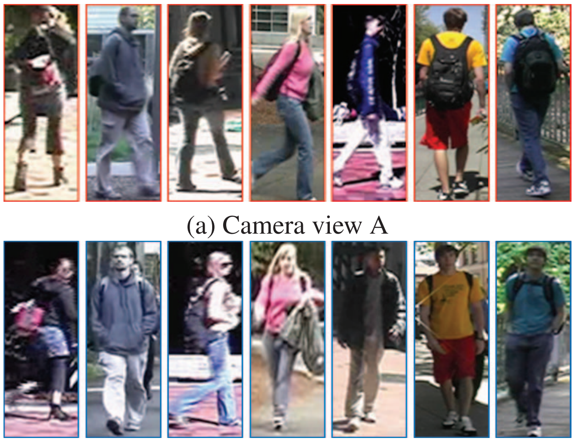
Abstract

In this paper, we propose a new approach for matching images observed in different camera views with complex cross-view transforms and apply it to person re-identification. It jointly partitions the image spaces of two camera views into different configurations according to the similarity of cross-view transforms. The visual features of an image pair from different views are first locally aligned by being projected to a common feature space and then matched with softly assigned metrics which are locally optimized. The features optimal for recognizing identities are different from those for clustering cross-view transforms. They are jointly learned by utilizing sparsity-inducing norm and information theoretical regularization. This approach can be generalized to the settings where test images are from new camera views, not the same as those in the training set. Extensive experiments are conducted on public datasets and our own dataset. Comparisons with the state-of-the-art metric learning and person re-identification methods show the superior performance of our approach.

1. Introduction

Person re-identification is to match the snapshots of pedestrians observed in non-overlapping camera views with visual features. It has drawn a lot of attentions in recent years [11, 23, 32, 27] because of its important applications in video surveillance [16], such as cross-camera tracking, multi-camera behavior analysis and pedestrian search. However, this problem is extremely challenging, because it is difficult to match the visual features of pedestrians captured in different camera views due to the large variations of lightings, poses, viewpoints, image resolutions, photometric settings of cameras, and backgrounds. Accurate human parsing[18] will benefit person re-identification, but it is a hard problems to solve.

Existing works solve this challenge in two possible ways: (1) learning the photometric or geometric transforms between two camera views, if the photometric/geometric models can be assumed [24]; (2) learning a distance metric or projecting visual features from different views into a common feature space for matching in order to suppress inter-camera variations. The approaches from both categories assume two fixed camera views with a uni-model inter-camera transform and labeled training samples from the two views are available. However, in practice the configurations (which are the combinations of view points, poses, image resolutions, lightings and photometric settings) of pedestrian images are multi-modal even if they are observed in the same camera views. Some examples are shown in Figure 1. Therefore, the inter-camera variations cannot be well learned with a single transform or metric. Moreover, given a large camera network in video surveillance, it is impossible to label training samples for every pair of camera views. It is highly desirable to develop an algorithm which can match images from two new camera views given training samples collected from other camera views.


(a) Camera view A

(b) Camera view B

Figure 1. Examples of pedestrians captured in two camera views in the VIPeR dataset[10]. Two images in the same column belong to the same person. Images have different poses, lightings and background even if they are captured in the same camera view.

We propose a new approach of learning locally aligned feature transforms across multiple views and apply it to person re-identification. The contribution of this work can be

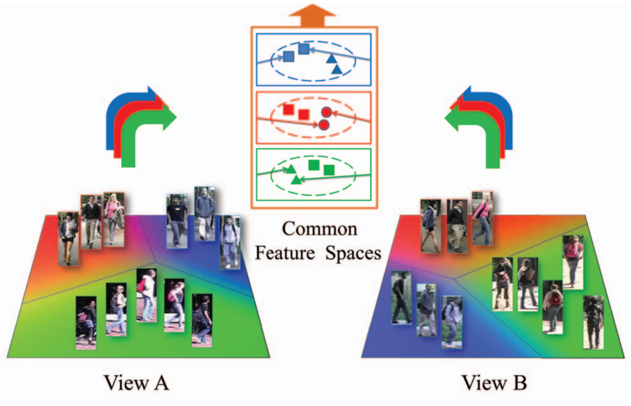


Figure 2. Person re-identification in locally aligned feature transformations. The image spaces of two camera views are jointly partitioned based on the similarity of cross-view transforms. Sample pairs with similar transforms are projected to a common feature space for matching.

summarized as following. (1) As illustrated in Figure 2, the proposed approach automatically partitions the image spaces of two camera views into subregions which correspond to different configurations, and learns a different feature transform for a pair of configurations. Given a pair of images to be matched, they are softly assigned to configuration types with a gating network and their visual features are projected to a common feature space and matched by a local expert. Therefore, it well handles the multi-modal transform problem discussed above. (2) The features optimal for configuration estimation and identity matching are different. They are jointly learned and selected in our approach with sparsity and log-determinant divergence regularizations. (3) The image spaces of the two camera views are jointly partitioned instead of separately, to avoid some combinations of configurations rarely appearing in the two views. The local experts of these rare combinations cannot be well learned given few, if any, training samples. (4) Besides suppressing cross-view variations, the discriminative power of local experts is further increased by locally magnifying inter-person variations. (5) This approach is extended to the case when test images are from new camera views not existing in the training set. Extensive experimental evaluations on public datasets and our own dataset and comparison with state-of-the-art methods show the effectiveness of the proposed approach.

2. Related Work

Metric learning and feature selection have been widely used to reduce cross-view variations and to increase the discriminative power in person re-identification. Some approaches [26, 23] assume that all the persons to be identified have samples in the training set. In [26], Partial Least Square reduction was used to reduce the dimensionality of visual features and to weight features according to their discriminative power under a one-against-all scheme. Lin and Davis [23] assumed that a feature optimal for distinguishing a pair of persons might not be effective for others, and therefore learned the dissimilarity profiles under a pairwise scheme. In order to identify persons outside the training set, Zheng *et al.* [32] formulated person re-identification as a distance learning problem by maximizing the probability that a pair of true match has a smaller distance than a wrong match. A relaxed distance metric learning is used to address this problem in [19]. In [17], Jurie *et al.* learned a metric specially designed for identification tasks under pairwise constraints and further kernelized it to overcome the linearity. In [11, 25] boosting and RankSVM were used to select features to compute the distance between images observed across camera views. In [22], Li *et al.* proposed a transferred metric learning framework for learning specific metric for different query-candidate combinations. Metric learning loses important information when directly computing the difference between two feature vectors without aligning them first. Although not being widely applied to person re-identification yet, Canonical Correlation Analysis (CCA)[14] has been used to match data from different views or in different modalities in the applications of face recognition [21, 33] and image-to-text matching [13]. All the approaches discussed above assume a single global model or a generic metric, which cannot well handle multiple types of transforms between two views. It is also hard for these learning-based approaches to be generalized to new views without re-labeling training data.

Localized learning methods work more effectively on datasets with complex distributions. They learn different classifiers or metrics for different clusters of images or even individual samples. Weinberger *et al.* [28] extended their metric learning framework named Large Margin Nearest Neighbor (LMNN) to learn multiple localized metrics for different image clusters. In mixture of experts [15], a gating network classified test samples into different clusters and samples within one cluster were classified with the same local expert. In [31], a local classifier was learned for every training sample. A test sample found its classifier from its nearest neighbor in the training set. Similarly, Zhan *et al.* [30] learned a different metric for each training sample. In [6, 7], each training sample had a different metric and all the metrics were aligned with global constraints. All the approaches discussed above matched images in the same feature space and did not consider cross-view transforms. They clustered images or identified per-instance metric/classifier by comparing visual similarities. Differently, we jointly partition the image spaces of two views based on the similarity of cross-view transform. Since the possible transforms is much less than the total visual diversity, it leads to a smaller number of local experts which can be well learned

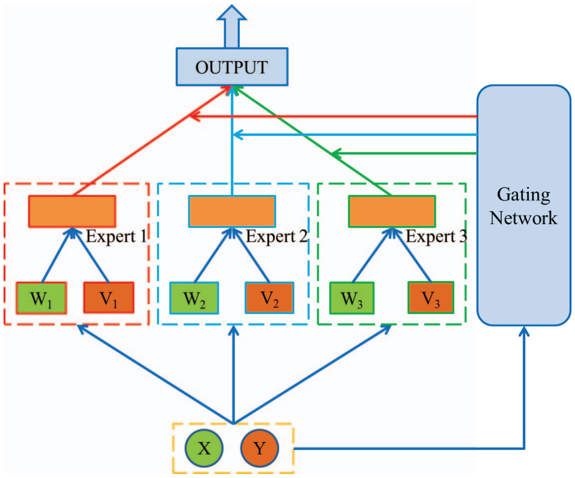


Figure 3. Graphical illustration of our model.

from a smaller training set. Moreover, features optimal for identifying cross-view transforms (*i.e.* identifying proper local metrics) are different from those for identity matching. Our approach automatically separates the two types of features with a proposed sparse gating network.

3. Model

A graphical illustration of our model is shown in Figure 3. $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^m$ are the visual feature vectors of a pair of images observed in two camera views. $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ are the training sets from the two views. There are K local experts to be learned. A pair of test samples to be matched are input to a gating network to choose local experts in a soft way, and matched with the selected experts. The details are given below.

Gating network. One way of designing the gating network by following traditional approaches with a single image space is to partition each of the two image spaces separately into K regions, and then learn K^2 experts for all the combinations. The gating functions in two image spaces are independent, *i.e.* $\eta(s_x = k_1, s_y = k_2) = \eta_x(s_x = k_1|x)\eta_y(s_y = k_2|y)$. This leads to a large number of experts. Since some configurations (s_x, s_y) rarely co-exist in both views, not enough training samples can be found to train the experts. Instead, we assume the two image samples are correlated and compute the gating function as[1]

$$p(s = k|\mathbf{x}, \mathbf{y}) = \frac{\exp(\boldsymbol{\phi}_k^T \mathbf{x}) \exp(\boldsymbol{\psi}_k^T \mathbf{y})}{\sum_{k'=1}^{K} \exp(\boldsymbol{\phi}_{k'}^T \mathbf{x}) \exp(\boldsymbol{\psi}_{k'}^T \mathbf{y})}, \quad \boldsymbol{\phi}_k, \boldsymbol{\psi}_k \in \mathbb{R}^m. \tag{1}$$

Local experts. Each local expert k does the alignment by projecting the two samples $(\mathbf{x}, \mathbf{y})$ to be matched into a common feature subspace with linear projections $\mathbf{W}_k$ and $\mathbf{V}_k$, and compare their Euclidean distance in this subspace. Let $z = 1$ indicates that $\mathbf{x}$ and $\mathbf{y}$ belong to the same person; otherwise it is zero. Then the expert computes the conditional probability,

$$p(z = 1|\mathbf{x}, \mathbf{y}, s = k) \propto e^{-||\mathbf{W}_k\mathbf{x} - \mathbf{V}_k\mathbf{y}||_2^2}, \quad \mathbf{W}_k, \mathbf{V}_k \in \mathbb{R}^{d \times m}. \tag{2}$$

$(\mathbf{W}_k,\ \mathbf{V}_k)$ is the alignment matrix pair for expert k. The decision function of two samples being the same identity is

$$p(z = 1|\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} p(z = 1|\mathbf{x}, \mathbf{y}, s = k) p(s = k|\mathbf{x}, \mathbf{y}). \tag{3}$$

4. Learning

4.1. Priors

The gating function parameters $\{\boldsymbol{\phi}_k, \boldsymbol{\psi}_k\}_{k=1}^K$ and expert parameters $\{\mathbf{W}_k, \mathbf{V}_k\}_{k=1}^K$ are to be learned from training data. They are all in very high dimensional spaces, and therefore need to be properly regularized with priors at the training stage. In order to select features more effectively for comparing cross-view transforms, a Laplacian prior is added to $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$,

$$p(\boldsymbol{\phi}_k) \propto \exp\left(-\frac{||\boldsymbol{\phi}_k||_1}{\lambda}\right), \quad p(\boldsymbol{\psi}_k) \propto \exp\left(-\frac{||\boldsymbol{\psi}_k||_1}{\lambda}\right). \tag{4}$$

ℓ_1 regularization requires the selected features to be sparse.

For each pair of $(\mathbf{W}_k, \mathbf{V}_k)$, if they are rotated by the same projection matrix $\mathbf{P}$ satisfying $\mathbf{P}\mathbf{P}^T = \mathbf{I}$, the projected common feature space and the conditional probability in Eq (2) remains the same. Therefore, the solution of $\mathbf{W}_k$ and $\mathbf{V}_k$ is not unique. Moreover, $\mathbf{W}_k$ and $\mathbf{V}_k$ may degenerate to zero and map all the points to a small region around the origin without proper regularization. Therefore, we employ a log-determinant divergence [20, 3] between $\mathbf{W}_k(\mathbf{V}_k)$ and a prior $\mathbf{W}_0(\mathbf{V}_0)$, which are learned from regularized CCA.

$$\mathcal{D}_{ld}(\mathbf{W}_k, \mathbf{W}_0) = \mathbf{trace}(\mathbf{W}_k^T(\mathbf{W}_0\mathbf{W}_0^T)^{-1}\mathbf{W}_k) - \mathbf{logdet}(\mathbf{W}_k\mathbf{W}_k^T(\mathbf{W}_0\mathbf{W}_0^T)^{-1}) - d. \tag{5}$$

We choose this prior for several reasons. $(\mathbf{W}_0, \mathbf{V}_0)$ is a reasonable regularization for local experts, since CCA provides a global solution of reducing cross-view variations. Second, we wish that the information loss is minimized after projecting feature vectors $\mathbf{x}$ to the common feature space through $\mathbf{W}_k$. It is known that the dimensionality of the space spanned with $\{\mathbf{x}_i\}$ is not reduced after being projected to $\{\mathbf{W}_0\mathbf{x}_i\}$ with CCA (*i.e.* feature space is not shrunk after CCA), and its covariance matrix becomes an identity matrix $\mathbf{I}$. Eq (5) measures the differential relative entropy between two equal mean multivariate Gaussian with variance

[1] Bias term can be added padding vector $\mathbf{x}, \mathbf{y}$ with 1 and thus omitted here for simplicity

$\mathbf{W}_0\mathbf{W}_0^T$ and $\mathbf{W}_k\mathbf{W}_k^T$[3]. This actually indicates the information loss when projecting $\mathbf{x}$ into the common feature space through $\mathbf{W}_k$ compared with the CCA feature space. Also, if the dimensionality of the space spanned by $\{\mathbf{x}_k\}$ gets reduced after projection, $\mathbf{det}(\mathbf{W}_k\mathbf{W}_k^T) = 0$, leading Eq (5) to be infinity. Therefore, the prior of Eq (5) prevents shrinkage of the feature space[2].

$$\begin{aligned} p(\mathbf{W}_k) &\propto \exp(-\mu\mathcal{D}_{ld}(\mathbf{W}_k, \mathbf{W}_0)), \\ p(\mathbf{V}_k) &\propto \exp(-\mu\mathcal{D}_{ld}(\mathbf{V}_k, \mathbf{V}_0)). \end{aligned} \quad (6)$$

4.2. Objective function

The objective function on the training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $(\mathbf{x}_i, \mathbf{y}_i)$ is a pair of samples with the same identity but observed in different views, is written as following

$$\begin{aligned} &\prod_{i=1}^N p(z_i = 1|(\mathbf{x}_i, \mathbf{y}_i), \{\boldsymbol{\phi}_k, \boldsymbol{\psi}_k\}, \{\mathbf{W}_k, \mathbf{V}_k\}) \\ &\prod_{k=1}^K p(\boldsymbol{\phi}_k)p(\boldsymbol{\psi}_k)p(\mathbf{W}_k)p(\mathbf{V}_k) \quad (7) \\ \propto &\prod_{i=1}^N \sum_{k=1}^K \frac{\exp(\boldsymbol{\phi}_k^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_k^T\mathbf{y}_i)}{\sum_{k'=1}^K \exp(\boldsymbol{\phi}_{k'}^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_{k'}^T\mathbf{y}_i)} e^{-||\mathbf{W}_k\mathbf{x}_i - \mathbf{V}_k\mathbf{y}_i||_2^2} \\ &\prod_{k=1}^K \exp\left(-\frac{||\boldsymbol{\phi}_k||_1 + ||\boldsymbol{\psi}_k||_1}{\lambda}\right) \\ &\prod_{k=1}^K \exp\big(-\mu(\mathcal{D}_{ld}(\mathbf{W}_k, \mathbf{W}_0) + \mathcal{D}_{ld}(\mathbf{V}_k, \mathbf{V}_0))\big). \quad (8) \end{aligned}$$

In Eq (7), the first row is the data likelihood and the second row is prior.

4.3. Optimization

Optimizing $\mathbf{W}_k, \mathbf{V}_k$

We iteratively optimize $\boldsymbol{\phi}_k$, $\boldsymbol{\psi}_k$, $\mathbf{W}_k$, and $\mathbf{V}_k$. Fixing $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_{\boldsymbol{k}}$, we apply a block coordinate descent to optimize $\mathbf{W}_k$ and $\mathbf{V}_k$. The gradient for the negative log of Eq (8), named f, can be calculated as follows, and similar results can be obtained for $\mathbf{V_k}$:

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{W}_k} =& 2\mathbf{W}_k\boldsymbol{\Sigma}_{xx}^k - 2\mathbf{V}_k\boldsymbol{\Sigma}_{yx}^k \\ &+ 2\mu\left((\mathbf{W}_0\mathbf{W}_0^T)^{-1} - (\mathbf{W}_k\mathbf{W}_k^T)^{-1}\right)\mathbf{W}_k, \quad (9) \end{aligned}$$

$$\boldsymbol{\Sigma}_{xx}^k = \sum_i \frac{\exp(\boldsymbol{\phi}_k^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_k^T\mathbf{y}_i)e^{-||\mathbf{W}_k\mathbf{x}_i - \mathbf{V}_k\mathbf{y}_i||_2^2}}{\sum_{k'} \exp(\boldsymbol{\phi}_{k'}^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_{k'}^T\mathbf{y}_i)e^{-||\mathbf{W}_{k'}\mathbf{x}_i - \mathbf{V}_{k'}\mathbf{y}_i||_2^2}} \mathbf{x}_i\mathbf{x}_i^T,$$

$$\boldsymbol{\Sigma}_{yx}^k = \sum_i \frac{\exp(\boldsymbol{\phi}_k^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_k^T\mathbf{y}_i)e^{-||\mathbf{W}_k\mathbf{x}_i - \mathbf{V}_k\mathbf{y}_i||_2^2}}{\sum_{k'} \exp(\boldsymbol{\phi}_{k'}^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_{k'}^T\mathbf{y}_i)e^{-||\mathbf{W}_{k'}\mathbf{x}_i - \mathbf{V}_{k'}\mathbf{y}_i||_2^2}} \mathbf{y}_i\mathbf{x}_i^T.$$

We update $\mathbf{W}_k^{t+1} = \mathbf{W}_k^t - \gamma\frac{\partial f}{\partial \mathbf{W}_k}$.

Optimizing $\boldsymbol{\phi}_k, \boldsymbol{\psi}_k$

Fixing $\mathbf{W}_k, \mathbf{V}_k$, the objective function becomes a generalized multinomial logistic regression with ℓ_1 regularization. It is the summation of a smooth negative log-likelihood function (g) and a non-smooth ℓ_1 regularizer in Eq (4). We employ a modified version of the cyclic coordinate descent method [1] which is simple and effective without calculating the Hessian. We take the derivative of the smooth loss term w.r.t each element j of $\boldsymbol{\phi}_k$.

$$\begin{aligned} &\frac{\partial g}{\partial \boldsymbol{\phi}_{kj}} = \\ &-\sum_i \left[\frac{e^{-||\mathbf{W}_k\mathbf{x}_i - \mathbf{V}_k\mathbf{y}_i||_2^2}}{\sum_{k'} \pi_{k'}(\mathbf{x}_i, \mathbf{y}_i)e^{-||\mathbf{W}_{k'}\mathbf{x}_i - \mathbf{V}_{k'}\mathbf{y}_i||_2^2}} - 1\right]\pi_k(\mathbf{x}_i, \mathbf{y}_i)\mathbf{x}_{ij}, \end{aligned}$$

$$\text{where} \quad \pi_k(\mathbf{x}_i, \mathbf{y}_i) = \frac{\exp(\boldsymbol{\phi}_k^T\mathbf{x}_i)\exp(\boldsymbol{\psi}_k^T\mathbf{y}_i)}{\sum_{k'}^K \exp(\boldsymbol{\phi}_{k'}\mathbf{x}_i)\exp(\boldsymbol{\psi}_{k'}\mathbf{y}_i)}.$$

Then we can define the *effective* gradient [1] to handle the case that when $\boldsymbol{\phi}_{kj} = 0$, where ℓ_1 regularizer is non-smooth

$$\frac{\partial f}{\partial \boldsymbol{\phi}_{kj}} = \begin{cases} \frac{\partial g}{\phi_{kj}} + \text{sign}(\boldsymbol{\phi}_{kj})/\lambda & \boldsymbol{\phi}_{kj} \neq 0 \\ \frac{\partial g}{\phi_{kj}} + 1/\lambda & \boldsymbol{\phi}_{kj} = 0, \frac{\partial g}{\boldsymbol{\phi}_{kj}} + 1/\lambda < 0 \\ \frac{\partial g}{\phi_{kj}} - 1/\lambda & \boldsymbol{\phi}_{kj} = 0, \frac{\partial g}{\boldsymbol{\phi}_{kj}} - 1/\lambda > 0 \\ 0 & \text{Otherwise} \end{cases}$$

The ℓ_1-regularizer is locally a hyperplane, thus $\frac{\partial^2||\boldsymbol{\phi}_{kj}||_1}{\partial\boldsymbol{\phi}_{kj}^2} = 0$. Then each $\boldsymbol{\phi}_{kj}$ can be optimized using the Newton's method as $\boldsymbol{\phi}_{kj}^{t+1} = \boldsymbol{\phi}_{kj}^t - \frac{\partial f}{\partial\boldsymbol{\phi}_{kj}}[\frac{\partial^2 g}{\partial\boldsymbol{\phi}_{kj}^2}]^{-1}$ when $\frac{\partial^2 g}{\partial\boldsymbol{\phi}_{kj}^2} > 0$ and a line search is carried out otherwise. The remaining question is how to choose j. Observing the geometric correlation of each feature, at the first few steps we choose j with steepest gradient and then we choose features that are geometrically close for the next coordinate descent step.

4.4. Multi-Shot extension

All the descriptions so far assume single-shot person re-identification, *i.e.* for a query sample $\mathbf{x}_i$ in view A, there is only one sample $\mathbf{y}_i$ with the same identity in the gallery of view B. Multi-shot person re-identification occurs when there are more than one samples $\mathcal{Y}_i$ with the same identity as $\mathbf{x}_i$ in view B. The identification is successful if there is at least one $\mathbf{y}_j \in \mathcal{Y}_i$ matched with $\mathbf{x}_i$. Our learning can be extended to the multi-shot scenario by modifying the data likelihood term in Eq (8) as,

$$\prod_i \max_{\mathbf{y}_j \in \mathcal{Y}_i} p(z_{ij} = 1|(\mathbf{x}_i, \mathbf{y}_j), \{\phi_k, \psi_k\}, \{\mathbf{W}_k, \mathbf{V}_k\}). \quad (10)$$

As the max operation is non-smooth, we use the log-sum-of-exponentials as a smooth approximate. Such function works better for a larger range rather than $[0, 1]$. Therefore a re-scaling factor $\eta = 0.1$ is used as follows,

$$\prod_i \eta\log\Big(\sum_{\mathbf{y}_j \in \mathcal{C}_i} \exp\big(\frac{p(z_{ij} = 1|(\mathbf{x}_i, \mathbf{y}_j), \{\mathbf{W}_k, \mathbf{V}_k\}, \{\phi_k, \psi_k\})}{\eta}\big)\Big).$$

[2] For numerical stability, we use a surrogate loss function $\log\det(\mathbf{W}_k^T(\mathbf{W}_0\mathbf{W}_0^T)^{-1}\mathbf{W}_k + \epsilon I)$ where $\epsilon = 10^{-15}$ instead.

Our multi-shot extension makes training easier, since it does not have to match every training pair when learning the cross-view transforms. It only needs to minimize the distance of best matched pairs and effectively reduces the number of cross-view transforms in consideration.

4.5. Discriminative metric learning

The proposed locally aligned feature transform only reduces the cross-view variations without considering how to discriminate different persons. In order to increase the discriminative power, we further learn a low-rank Manhalanobis $\mathbf{M}_k$ in each aligned common feature space.

$$d(\mathbf{x},\mathbf{y}|\{\mathbf{M}_k\}) = \sum_k \alpha_k \exp(-||\mathbf{W}_k\mathbf{x} - \mathbf{V}_k\mathbf{y}||_{\mathbf{M}_k}),$$

$$\alpha_k = \frac{\exp(\boldsymbol{\phi}_k\mathbf{x})\exp(\boldsymbol{\psi}_k\mathbf{y})}{\sum_{k'}\exp(\boldsymbol{\phi}_{k'}\mathbf{x})\exp(\boldsymbol{\psi}_{k'}\mathbf{y})},$$

$$||\mathbf{W}_k\mathbf{x} - \mathbf{V}_k\mathbf{y}||_{\mathbf{M}_k} = (\mathbf{W}_k\mathbf{x} - \mathbf{V}_k\mathbf{y})^T\mathbf{M}_k(\mathbf{W}_k\mathbf{x} - \mathbf{V}_k\mathbf{y}).$$

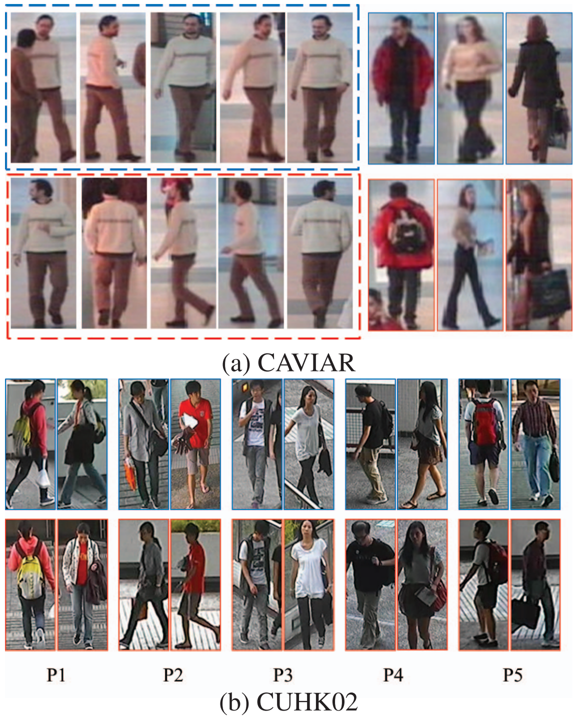
At the training stage, we learn $\mathbf{M}_k$ by maximizing the expected rank-1 accuracy within top ranks as proposed in [9]. The objective function is

$$\theta\sum_k \mathbf{tr}(\mathbf{M}_k) - \sum_{\mathbf{x}_i}\log\left(\frac{\sum_{\mathbf{y}_j\in\mathcal{Y}_i} d(\mathbf{x}_i,\mathbf{y}_j)}{\sum_{\mathbf{y}_j\in\mathcal{T}_i} d(\mathbf{x}_i,\mathbf{y}_j)}\right). \quad (11)$$

$\mathcal{Y}_i$ is the set of samples in view B with the same identity as $\mathbf{x}_i$. $\mathcal{T}_i$ is the set of top ranked samples in view B with $\mathbf{x}_i$ as query and under the distance $||\mathbf{W}_k\mathbf{x} - \mathbf{V}_k\mathbf{y}||_2^2$. The goal is to well distinguish the true identities with other top ranked persons, who are easily confused with the query person. To optimize $\mathbf{M}_k$ we take a block coordinate descent method w.r.t each $\mathbf{M}_k$ and project it to $\{\mathbf{M}_k \succeq 0\}$ by spectral decomposition. $\mathbf{tr}(\mathbf{M}_k)$ equals to the nuclear norm of metric $\mathbf{M}_k$ and induces low-rank.

5. Experiment

Datasets. We evaluate our approach on two widely used public benchmark datasets, VIPeR[10] and CAVIAR[2], and our own dataset(CUHK02)[3]. VIPeR contains 632 persons and two camera views. Each person has one image per camera view. CAVIAR contains 72 persons and two views. 50 persons appear in both views and 22 persons appear only in one view. Each person has 5 images per view. These two datasets are used to evaluate person re-identification given two fixed camera views. CUHK02 contains 1,816 persons and five pairs of camera views (P1-P5, ten camera views). They have 971, 306, 107, 193 and 239 persons respectively. Each person has two images in each camera view. This dataset is used to evaluate the performance when camera views in test are different than those in training. Samples from those datasets can be found in Figure 1 and 4.


(a) CAVIAR

(b) CUHK02

Figure 4. Sample images from CAVIAR and our dataset. (a) The five columns on the left show five images of the same person in each of the two camera views. The three columns on the right show another three persons. Only two images in different views are shown for each person. CAVIAR not only has large variations on poses, but also on sizes and aspect ratios. (b) CUHK02 has five pairs of camera views denoted with P1-P5. Two exemplar persons are shown for each pair of views.

Methods in comparison. We compare with several state-of-the-art methods which are popularly used to reduce cross-view variations, including CCA[14], Kernel CCA[33], and three metric learning methods: Information Theoretical Metric Learning (ITML)[3], Multiple Largest Margin Nearest Neighbor (mLMNN)[28] and Localized Distance Metric Learning (LDM)[29]. mLMNN and LDM are localized learning methods. Many state-of-the-art person re-identification methods have published their results on VIPeR and CAVIER. We also compare with them on these two datasets by using the same evaluation protocol.

Features. We combine four types of visual features. LBP, HSV color histogram, and Gabor features are extracted from a 16×8 dense grid with 25% overlap between adjacent patches. Gabor filters have 8 orientations and 5 scales. HOG features have 9 orientations and each block contains 3×3 cells. Each type of feature is normalized to a vector with unit norm and multiple features are stacked to form a single feature vector. This combined feature vector is further normalized to zero mean. The three metric learning methods ITML, mLMNN, and LDM require reducing the dimensionality of visual features as a pre-processing step. We use regularized CCA for dimension reduction, because

[3] The dataset can be downloaded from http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html

our empirical study shows that it is better than using PCA.

Parameter settings. We set parameters as $\lambda = \sqrt{m}$ in Eq(4), $\mu = \frac{1}{d}$ in Eq(6), $\theta = \frac{1}{d}$ in Eq(11). The dimensionality of projected common feature space in local experts (*i.e.* number of rows of $\mathbf{W}_k$ and $\mathbf{V}_k$) is changeable and we set it as the number of CCA components with non-zero canonical correlation. The number of local experts $K = 5$.

5.1. Identification with two fixed camera views

It is assumed that all the training and test samples come from the same pair of camera views. Following existing protocols, both single-shot and multi-shot identification are evaluated. Each query person has one query image. Single-shot assumes each person has one image in the gallery, while multi-shot assumes M gallery images per person.

5.1.1 Single-Shot results

Figure 5 (a) shows the results of single-shot test on VIPeR, CAVIAR, and CUHK02. All the random partitions described below repeat for 100 times. Two protocols on VIPeR were used in the past: randomly splitting the whole dataset into 316 persons for training and the remaining 316 for test; and randomly splitting into 100 persons for training and 532 for test. We evaluate both. Table 1 and 2 compare with results previously published on VIPeR with the same protocol. For CUHK02, we choose view pair P1 for evaluation. It has 971 persons, which are split to 485 for training and 486 for test. Each person has two images in each view. They are also randomly selected. CAVIRR has a small number of persons, so we did not split the persons. It is also to be consistent with existing protocol. If a person has images in both camera views, we randomly select two pairs of images in different views for training. One query image and one gallery image are randomly selected from the remaining images per person. Table 3 compares with results previously published on CAVIAR. Experimental results show that our method significantly outperforms other learning approaches and achieves the best results on the two public datasets. CCA does not work very well since it assumes the feature transforms to be uni-modal while the three datasets are much more complicated. Kernel CCA alleviates the problem, but its performance is still not good as ours after tuning the kernel. Metric learning methods do not align features. Much information has been lost after the first step of computing pointwise difference $\delta_i = \mathbf{x}_i - \mathbf{y}_i$ due to misalignment between features. For our method, we also compare with the case when discriminative metric learning described in Section 4.5 is not used. This technique is effective. Its improvement is even larger for multi-shot test.

Methods	Top 1	Top 10	Top 25	Top 50
Ours	**29.6**	**69.3**	**88.7**	**96.8**
KISSME[19]	19.6	62.2	80.7	91.8
PS[2]	21.8	57.2	76.0	88.1
SDALF[5]	19.9	49.4	70.5	84.8
PRDC [32]	15.7	53.9	76	87
LDML[12][19]	10.4	31.3	44.6	60.4
LMNN-R [4]	23.7	68	84	93
MCC[8][32]	15.2	57.6	80	91
PCCAχ^2_{rbf} [17]	19.3	64.9	83	**96**

Table 1. Compare rank-n identification rates (%) with other published single-shot results on VIPeR. The gallery size is 316.

Methods	Top 1	Top 5	Top 10	Top 20
Ours	**12.90**	**30.30**	**42.73**	**58.02**
PRDC[32]	9.12	24.19	34.40	48.55
PCCAχ^2_{RBF}[17]	9.27	24.89	37.43	52.89
MCC[32]	5.00	16.32	25.92	39.64

Table 2. Compare rank-n identification rates (%) with other published single-shot results on VIPeR. The gallery size is 512.

Figure 6 shows the rank-1 and rank-10 rates on VIPeR and CUHK02 with different numbers (K) of local experts. Our method works in a large range of $K = 4 - 10$, because the gating network softly splits the feature space and output is the weighted sum of all experts. With a large number of experts, some adjacent regions may share training samples without degenerating each expert. Figure 7 shows an exemplar pair for each local expert, according to the largest responses of the gating network. These pairs show different transforms caused by poses, lightings and backgrounds.

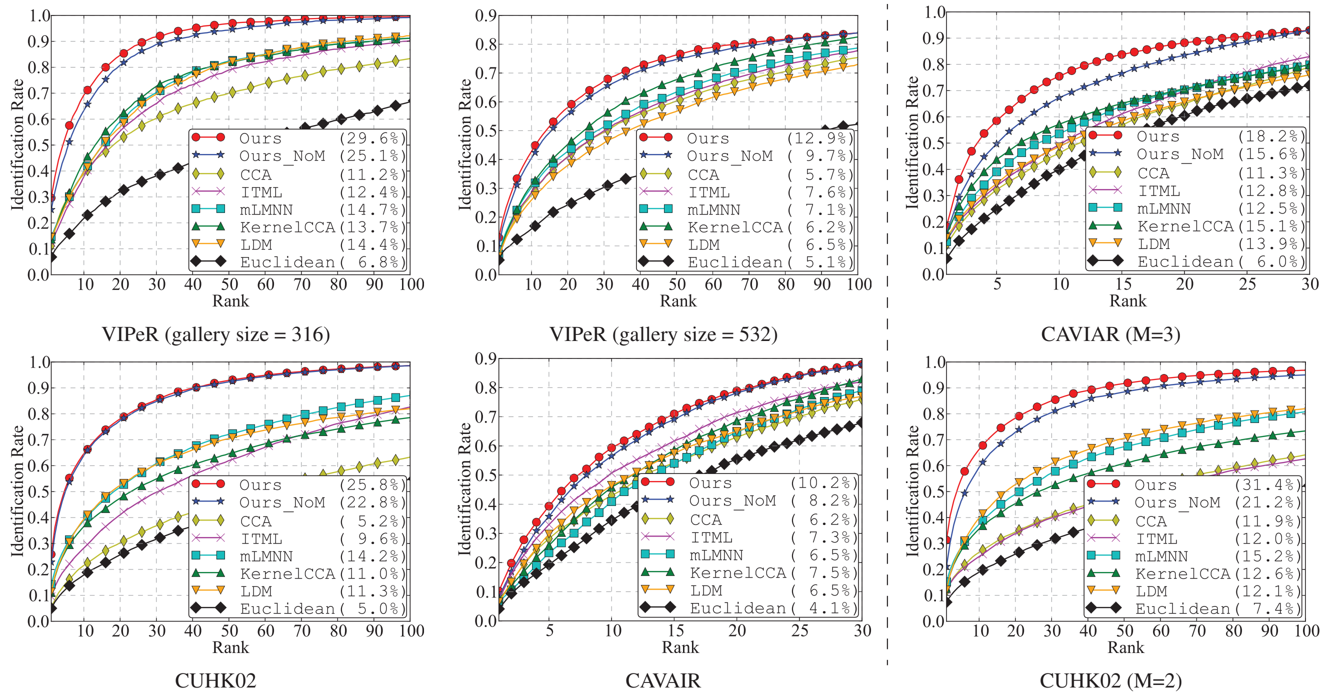
5.1.2 Multi-Shot results

Figure 5 (b) shows the results of multi-shot test on CAVIAR and CUHK02 P1, since VIPeR does not have multiple images per view. The dataset partition is similar to Section 5.1.1. On CAVIAR, each person has $M = 3$ gallery images following the protocol in [2]. On CUHK02 P1, each person has $M = 2$ gallery images. Table 4 compares with other multi-shot results published on CAVIAR[4]. Our method again shows superior performance. Its success is also due to the fact that at the training stage it does not try to reduce cross-view transforms for every pair of images, which is difficult, but instead uses a smoothed max function to select the best matches from multi-shots for learning the feature transforms. Thus it makes the training easier.

5.2. More general camera settings

Our method can be easily extended to more general settings when camera views in test are not the same as those in training. But when learning the discriminative metric in

[4] Notice that PS [2] and SDALF [5] are the only published results on CAVIAR. But they both rely on features specially design for person identification according to prior knowledge but without any learning methods. So they did not use any training samples, but we do.



VIPeR (gallery size = 316) VIPeR (gallery size = 532) CAVIAR (M=3)

CUHK02 CAVAIR CUHK02 (M=2)

(a) Single-Shot Results (b) Multi-Shot Results

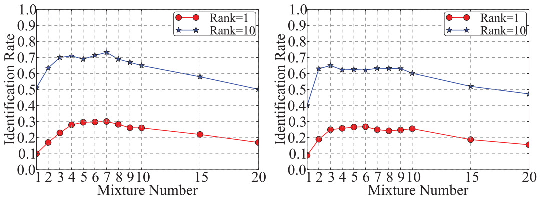
Figure 5. Rank-n identification rates give two fixed camera views. (a) Single-shot results on VIPeR, CAVIAR and CUHK02. (b) Multi-shot results on CAVIAR and CUHK02. Rank-1 rates are shown in parentheses. Ours_NoM denotes our method but without discriminative metric learning in Section 4.5. Euclidean is to directly match features.

Methods	Top 1	Top 5	Top 10	Top 30
Ours	**10.2**	**39**	**59**	**88**
PS[2]	8.5	32	48	86
SDALF[2]	6.8	25	45	83

Table 3. Compare rank-n identification rates (%) with other published single-shot results on CAVIAR.

Methods	Top 1	Top 5	Top 10	Top 30
Ours	**18.2**	**58**	**75**	92
PS[2]	13.5	44	64	**93**
SDALF[2]	9.0	34	53	88

Table 4. Compare rank-n identification rates (%) with other published multi-shot results (M=3) on CAVIAR.



(a) VIPeR (b) CUHK02

Figure 6. Identification rates with different number of experts.

Section 4.5, we have to assume each view in the training set could be a query view or a gallery view. If the training set contain multiple view pairs, we simply put their training samples together. Our training set contains three view pairs (P1, P2 and P3) with $1,384$ persons. View pairs P4 and P5 are selected for test. To make results stable, we randomly select a gallery set of 100 persons for 100 times. Figure 8 plots the multi-shot (M=2) test results. Our method is still effective, because it has the ability to find the best cross-view transforms from a complicated training set with combined view pairs. Table 5 reports the rank-1 rates when P4 is in test and the training set has different combination of view pairs. In CUHK02, the cross-view transforms in P3 have larger difference than those in P4. When P3 is added to the training set, the performance of other learning methods (LDM, LMNN and ITML) drops significantly, because it makes the feature transforms in the training set more complicate to learn and there is a larger mismatch between the training set and camera views in test. See the results of columns (P1) and (P1, P3). Our method and mLMNN are much more robust to this change.

Training	(P1, P2, P3)	P1	(P1, P2)	(P1, P3)
Ours	**28.2**	**26.2**	**28.1**	**25.9**
mLMNN	21.1	22.1	22.5	20.9
LMNN	13.9	15.8	17.3	13.2
ITML	15.7	17.9	18.8	11.3
LDM	12.7	13.6	16.8	10.1

Table 5. Rank-1 rates (%) when P4 is in test and the training set has different combinations of view pairs.

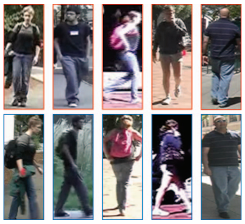


Figure 7. Exemplars sampled from each experts for VIPeR.

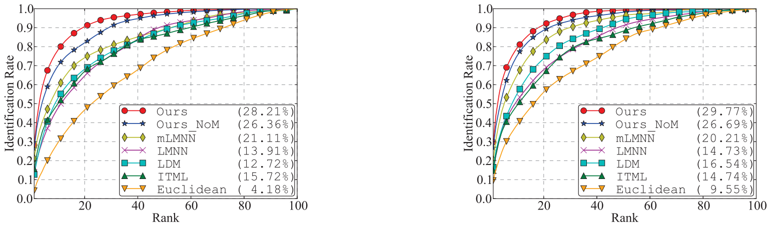


Figure 8. Rank-n identification rates under more general camera settings. Training set includes samples from P1 - P3. The left is the result of P4 in test and the right is the result of P5 in test.

6. Conclusions

We propose locally aligned feature transforms for matching pedestrians across camera views with complex cross-view variations. Images to be matched are softly assigned to different local experts according to the similarity of cross-view transforms, then they are projected to a common feature space and matched with a locally learned discriminative metric. It outperforms the state-of-the-art under the setting when two fixed camera views are given. Experiments on a small camera network with five pairs of camera views show its good potential of being generalized to generic camera settings. In the future, we will further explore its generalization capability by creating a much larger camera network with more diversified cross-view variations.

7. Acknowledgment

This work is supported by the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project No. CUHK 417110 and CUHK 417011) and National Natural Science Foundation of China (Project No. 61005057).

References

[1] G. C. Cawley, N. L. C. Talbot, and M. Girolami. Sparse multinomial logistic regression via bayesian l1 regularisation. In *NIPS*, 2007.

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

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[4] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*. 2011.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[6] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*. 2007.

[7] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007.

[8] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005.

[9] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2004.

[10] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. 2007.

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[12] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[13] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 2004.

[14] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 1936.

[15] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 1991.

[16] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *CVPR*, 2005.

[17] F. Jurie and A. Mignon. Pcca: A new approach for distance learning from sparse pairwise constraints. *CVPR*, 2012.

[18] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *CVPR*, 2007.

[19] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[20] B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with bregman matrix divergences. *JMLR*, 2009.

[21] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *CVPR*, 2009.

[22] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012.

[23] Z. Lin and L. Davis. Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. In *Proc. Int'l Symposium on Advances in Visual Computing*, 2008.

[24] B. Prosser, S. Gong, and T. Xiang. Multi-camera matching using bi-directional cumulative brightness transfer functions. In *BMVC*, 2008.

[25] B. Prosser, W. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *BMVC*, 2010.

[26] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least sqaures. In *SIBGRAPI*, 2009.

[27] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[28] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *ICML*, 2008.

[29] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI*, 2006.

[30] D.-C. Zhan, M. Li, Y.-F. Li, and Z.-H. Zhou. Learning instance specific distances using metric propagation. In *ICML*, 2009.

[31] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.

[32] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

[33] W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial expression recognition using kernel canonical correlation analysis (kcca). *Neural Networks, IEEE Transactions on*, 2006.