Kernelized View Adaptive Subspace Learning for Person Re-identification

Qin Zhou[12]
zhou.qin.190@sjtu.edu.cn
Shibao Zheng[12]
sbzh@sjtu.edu.cn
Hang Su[34]
suhangss@gmail.com
Hua Yang[12]
hyang@sjtu.edu.cn
Yu Wang[12]
txtxs@sjtu.edu.cn
Shuang Wu[12]
shuangwu@sjtu.edu.cn

[1] Department of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2] Shanghai Key Lab of Digital Media Processing and Transmission
[3] Department of Computer Science and Technology, Tsinghua Univelrsity
[4] Robotics Institute, Carnegie Mellon University

Abstract

Person re-identification refers to the task of recognizing the same person under different non-overlapping camera views and across different time and places. Many successful methods exploit complex feature representations or sophisticated learners. A recent trend to tackle this problem is to learn a suitable distance metric, the aim of which is to minimize the distance between true matches while maximize the distance between mismatched pairs. However, most of the existing metric learning algorithms directly take the difference of pairwise features in the original feature space as input. By doing so, they implicitly assume that there exists a projection matrix which can map feature vectors in two different subspaces into an identical subspace where desired feature distribution (features of the same person come closely and faraway otherwise) can be achieved. In this paper, we propose to learn different projection matrices for different camera views, thereby the learned matrices are adaptive to different camera views and a common subspace satisfying the desired feature distribution is more likely to be pursued. To better adapt to the different variations encountered by different views, the kernel trick is adopted to catch more information such that nonlinear transformation is possible. During test phase, the features under different camera views are projected into the learned subspace and a simple nearest neighbor classification is performed. Extensive experiments on four challenging datasets (VIPeR, iLIDS, CAVIAR4REID and ETHZ) demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Person re-identification is an important problem with many applications. Modern long-term tracking systems often need to verify whether two tracklets under different camera views

Pages 26.1-26.12
DOI: https://dx.doi.org/10.5244/C.29.26

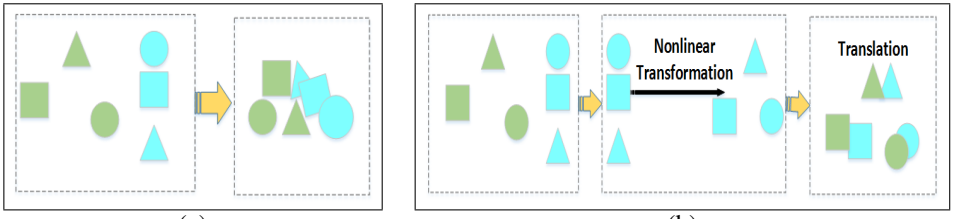


Figure 1: A conceptual illustration of how different projection functions can be more flexible than the same ones (Same shape represents same person, different colors correspond to different views). (a)The same transformation: the same rotation and scaling applied to both views (sky blue and green), cannot separate classes; (b) Different transformations: a nonlinear transformation applied to sky blue view and translation applied to green view, the different transformations successfully adapt to different view variation.

belong to the same person, which is especially important in smart video surveillance systems. Besides, with more and more surveillance cameras in our city collecting large amount of surveillance videos every day, we can not merely rely on human labors to recognize people across cameras, making the developing of an automatic person re-identification system vitally imperative. Although the depicted system is fascinating, person re-identification is confronted with great challenges in real world scenarios. The illumination and camera settings often bear great variations across cameras, in which case the appearance of different people can be much more alike than appearance of the same person across different views. Besides, in public space, people tend to dress similarly and a group of pedestrians are likely to walk together, which may bring severe occlusions among individuals, making the appearance clue less reliable.

In this paper, we aim to learn different projection functions for different camera views (see Figure 1 for a motivating example). We assume that different camera views should be equipped with different projection functions to reduce the influence of inter-camera variations and we can just focus on the identity difference. Besides different projection functions, we utilize the kernel trick to achieve nonlinear transformation such that high-order information of the image features can be catched. To that end, a novel algorithm coined as Kernelized View Adaptive Subspace Learning (KVASL) is proposed.

The key contributions of our paper are as follows:

- We propose to learn different projection functions for different views to minimize the distortion caused by different camera settings and view changes, which not only allows us to focus on the difference in identity, but also can be applied to a significantly wider range of realistic scenarios since the feature type and dimensionality are not restricted to be the same in our algorithm.

- We propose to learn the nonlinear projection functions by kernelizing the formulation which can exploit the high-order information. Also, we utilize the alternately iterative optimization algorithm to solve the proposed subspace learning problem.

1.1 Related Work

Most of the existing algorithms try to tackle the person re-identification problem by seeking descriptive and robust representations of the human appearance. For instance, Farenzena et al. [6] model three complementary aspects of the human appearance: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy. They take into account the symmetry and asymmetry structure of the human appearance, which proves to be effective in modeling human appearance. Ma et al. [12] try to aggregate simple local descriptors with Fisher Vector to build discriminative yet robust feature representation for each individual, which performs well on the available public datasets. In [3] Cheng et al. apply Pictorial Structures to person re-identification. They fit a body configuration composed of chest, head, thighs, and legs on pedestrian images and extract per-part color information as well as color displacement within the whole body.

The above-mentioned algorithms need careful consideration of the challenges posed by real world scenarios, to avoid this, some algorithms try to learn discriminative feature models based on sophisticated learners. In [8] Hirzer et al. try to combine feature designing with feature selection to exploit complementary information of the two methods. They first perform feature similarity comparisons based on the region covariance descriptors. Then human interaction is required to ensure that the true matched image is among the top ranked images, otherwise, a discriminative model is learned to refine the ranking results. However, the boosting based feature selection is instance specific, which is too time-consuming and infeasible in real scenarios. Gray and Tao [7] also present an algorithm to learn sophisticated feature model based on adaboost, which tries to weigh different local features according to their performance on the training set. However they treat each feature channel independently, which leads to sub-optimal results. Apart from feature selection, some algorithms aim to learn a perfect subspace where features of the same person distribute tightly (or are more correlated) than features of different people. For example, Pedagadi et al. [15] try to learn a transformation based on Local Fisher Discriminant Analysis, which can maximize the between-class separability and preserve the multi-class modality. The transformation has a closed-form solution by representing the objective as a generalized eigenvalue problem. An et al. [1] propose a reference-based method for across camera person re-identification. They perform Regularized Canonical Correlation Analysis to maximize the correlations between the related data pairs, then the test data is projected into the learned subspace and the reference descriptors of the test images are constructed by measuring the similarity between them and the reference data.

A new trend that has recently been explored is metric learning, which aims to learn appropriate distance/similarity measure based on the provided label information. In [18], the authors propose to learn optimal distance metric by maximizing the probability of a relevant image pair having smaller distance than a related irrelevant one. Kostinger et al. [10] present an efficient closed-form solution to the distance metric from a statistical inference perspective. The algorithm proposed by [9] leverages relaxed pairwise constraints to learn the transition from one camera to the other. In fact, explained from another way, the above-mentioned metric learning methods can all be interpreted as subspace learning algorithms, which try to learn a common subspace for two cameras with some constraints proposed by the authors.

2 Kernelized View Adaptive Subspace Learning (KVASL)

In this section, we first elaborate on the formulation of the proposed algorithm, which can be considered as an extension of the traditional Mahalanobis metric learning algorithms. Then the kernel trick is adopted to handle the linearly non-separable situations. In this case, high-order information of the features is considered and a more discriminative subspace is supposed to be learned. Finally, we utilize the alternately iterative optimization algorithm to solve the proposed subspace learning problem.

2.1 Problem Formulation

One prominent approach for metric learning is Mahalanobis distance learning, the goal of which is to adapt some pairwise real-valued metric function to the problem of interest using the information brought by training examples. In general, the Mahalanobis distance metric measures the squared distance between two data points x_i, x_j:

$$d_M^2(x_i,x_j) = (x_i - x_j)^T M(x_i - x_j) \tag{1}$$

where $M \geq 0$ is a positive semi-definite matrix and $x_i, x_j \in R^d$ is a pair of image samples. The existing metric learning methods differ in the objective functions they adopt.

An alternative formulation of Eq.(1), which is more intuitive, can be expressed as follows:

$$d_L^2(x_i,x_j) = ||L(x_i - x_j)||^2 \tag{2}$$

From Eq.(1) and Eq.(2) we can derive that $M = L^T L$, $L \in {}^{k\times d}$, where k is the rank of M. L can be considered as a projection matrix that transforms data in the original feature space to another subspace. Starting from this point, we formulate our algorithm as follows.

First, we introduce some notations that are commonly used in the following part. Let $camA, camB$ represent camera A and B. We assume that there are N different people in both camera views and each view contains D image of a specific person. We assume $D = 1$ for simplification. For convenience, we denote the features in camera A as $A_{train} = \{x_i^{train}, i \in \{1,\cdots,N_{train}\}\}$ and the features in camera B is denoted as $B_{train} = \{y_i^{train}, i \in \{1,\cdots,N_{train}\}\}$. The image pair $(x_i^{train} \in R^{d_a}, y_i^{train} \in R^{d_b})$ correspond to images of the same person i under different camera views, and d_a,d_b are the dimensions of the original feature subspaces. In our algorithm, d_a,d_b are not restricted to be the same. The motivation of our algorithm is to learn transform matrices for each camera view. To simplify the problem, we let $L_A \in R^{k\times d_a}, L_B \in R^{k\times d_b}$. Since our algorithm is based on pairwise instances, we first introduce two data sets according to the label constraints of their elements.

- Must-link set S: $S = \{(x_i^{train}, y_i^{train})\}, \quad i = 1,...,N_{train}$
- Cannot-link set D: $D = \{(x_i^{train}, y_j^{train})\}, i, j \in \{1,...,N_{train}\}, j \neq i$

Our algorithm is based on the intuition that distances between features of the same person should be smaller than distances of features belonging to different people. The empirical assumption is also utilized by many other metric learning algorithms [9][18]. In practice, regularization is used to avoid over-fitting problem especially when the sample size is mall. Various constraints can be imposed on the learned Mahanalobis matrix, including positive

semi-definity, low rank, sparsity, etc. We simply add the Frobenius norm as the regularization term. Therefore the final loss function can be formulated as:

$$\ell(L_A,L_B) = \frac{1}{|S|}\sum_{i=1}^{N_{train}} ||L_A x_i^{train} - L_B y_i^{train}||^2 - \frac{\lambda}{|D|}\sum_{i=1}^{N_{train}}\sum_{j=1,j\neq i}^{N_{train}} ||L_A x_i^{train} - L_B y_j^{train}||^2 + \mu_A||L_A||_F^2 + \mu_B||L_B||_F^2 \tag{3}$$

where S, D are the two data sets mentioned above, λ is a compromise parameter between the effect of the true matched image pairs and mismatched pairs, $|S|, |D|$ correspond to the cardinality of the related data set, and μ_A,μ_B are parameters that balance between the performance and complexity of the learned model, respectively, it degenerates into the traditional loss function as in [9] when $L_A = L_B$.

2.2 Kernelization

It is possible to apply the "kernel trick" to capture more information of the features and improve their separability when the data is not linearly separable [16]. In this section, we elaborate on how the "kernel trick" is applied in our algorithm. To leverage the benefits of kernel trick, we re-write the squared Frobenius norm of the distance in the kernel space as:

$$d^2_{L_A^K,L_B^K}(x_i,y_j) = ||L_A^K\phi(x_i) - L_B^K\phi(y_j)||^2 \tag{4}$$

where $\phi(x) \in R^m$ corresponds to the feature vector in the kernel feature space. Note that the dimension of the transform matrices (L_A^K, L_B^K) is $d' \times m$ ($d' << m$), where d' is the rank of the projection matrices in the kernel space and m indicates the dimension of the kernel space. In this case, transform matrices $Q^A, Q^B \in R^{d'\times N}$ are introduced such that the projection matrices in the kernel feature space can be expressed as follows:

$$\begin{aligned} L_A^K &= Q_A\phi(A_{train})^T \\ L_B^K &= Q_B\phi(B_{train})^T \end{aligned} \tag{5}$$

where N is the number of the people in the training set, and $\phi(A_{train}) = \left[\phi(x_{1A}^{train}),\cdots,\phi(x_{NA}^{train})\right]$, $\phi(B_{train}) = \left[\phi(y_{1B}^{train}),\cdots,\phi(y_{NB}^{train})\right] \in R^{m\times N}$ are the matrices formed by feature vectors in the kernel space of the related camera view. Through thorough derivation (trace cyclic permutation is utilized), the kernelized version of Eq.(3) can be formulated as:

$$\begin{aligned} \ell_K(L_A^K,L_B^K) = &\frac{1}{|S|}tr(K_A{}^T K_A Q_A^T Q_A - 2K_B{}^T K_A Q_A^T Q_B + K_B{}^T K_B Q_B^T Q_B) + \mu_A||L_A^K||_F^2 + \mu_B||L_B^K||_F^2 \\ &- \frac{\lambda}{|D|}tr((|S|-1)K_A{}^T K_A Q_A^T Q_A - 2K_A{}^T X K_B Q_B^T Q_A + (|S|-1)K_B{}^T K_B Q_B^T Q_B) \end{aligned} \tag{6}$$

where $K_A = \phi(A_{train})^T\phi(A_{train}) \in R^{N\times N}$, $K_B = \phi(B_{train})^T\phi(B_{train}) \in R^{N\times N}$ are symmetry matrices, $tr(\cdot)$ indicates the trace of the matrix and X corresponds to the matrix whose diagonal elements are all zeros and the other elements are all ones.

2.3 Alternately Iterative Optimization Algorithm

In fact, iteratively optimizing over Q_A with Q_B fixed and vice versa is a quadratic matrix programming problem, as shown in [2], it is ensured to be convex if and only if the quadratic

Algorithm 1 The optimization procedure for learning Q_A, Q_B

Input and Initialization:

A_{train}, B_{train} for training, Q_A, Q_B are initialized to identity matrices of size $N \times N$. N is the number of persons in the training set, ε is set to $\varepsilon = 10^{-5}$ in our experiment.

Train:

Compute initial loss ℓ according to Eq. (6).
if $\ell > \varepsilon$ **then**
  Compute $\frac{\partial \ell}{\partial Q_A}, \frac{\partial \ell}{\partial Q_B}$ according to Eq. (7);
  Update Q_A, Q_B according to Eq. (8);
  Update the loss ℓ according to Eq. (6);
else
  Break;
end if
return Q_A, Q_B;

matrix coefficient is positive semi-definite, which can be immediately deduced in our case since ${K_A}^T K_A$ and ${K_B}^T K_B$ are positive semi-definite matrices. In this case, we adopt an alternately iterative gradient descent method to optimize our loss function.

The gradients of Q_A,Q_B in Eq.(6) can be derived as follows:

$$\begin{aligned}\frac{\partial \ell}{\partial Q_A} &= \frac{2}{|S|}(Q_A K_A K_A - Q_B K_B K_A) - \frac{2\lambda}{|D|}[(|S|-1)Q_A K_A K_A - Q_B K_B X K_A] + 2\mu_A Q_A K_A \\ \frac{\partial \ell}{\partial Q_B} &= \frac{2}{|S|}(Q_B K_B K_B - Q_A K_A K_B) - \frac{2\lambda}{|D|}[(|S|-1)Q_B K_B K_B - Q_A K_A X K_B] + 2\mu_B Q_B K_B\end{aligned} \tag{7}$$

where X is the same matrix as in Eq.(6). In this case, the matrices to be learned can be updated in the following way:

$$Q_A(t+1) = Q_A(t) - \eta_{Q_A}\frac{\partial \ell}{\partial Q_A}, \quad Q_B(t+1) = Q_B(t) - \eta_{Q_B}\frac{\partial \ell}{\partial Q_B} \tag{8}$$

where $\eta_{L_A}, \eta_{L_B}, \eta_{Q_A}, \eta_{Q_B}$ are the learning rate corresponding to each transform matrix. Once Q_A, Q_B are learned, L_A^K, L_B^K can be obtained by calculating Eq.(5), but it's not necessary to explicitly knowing L_A^K, L_B^K as explained in Section 2.4. The detailed optimization procedure is presented in Algorithm 1.

2.4 Person Re-identification

After obtaining the subspace projection matrices, during the testing phase, the feature vectors in the original feature space are projected into the learned kernel subspace, then the nearest neighbor classification is performed to rank the gallery images according to their distance with respect to a specific probe image in the test set.

We present how to perform distance calculation in the projected feature space here. Assume there are total N_{test} people in the test set, and each individual has two images captured in two non-overlapping cameras (camera A, B). In this case,we have two feature sets $A_{test} = \{(x_{iA}^{test}), i \in \{1,2,...,N_{test}\}\}, B_{test} = \{(y_{jB}^{test}), j \in \{1,2,...,N_{test}\}\}$, each correspond to feature vectors in the related camera view. Then the distance between x_{iA} and y_{jB} in the projected feature space can be calculated as in Eq.(4), which can not be directly calculated since the dimension of the kernel space is unknown (e.g. the dimension of RBF kernel space is infinite). In this case, we rewrite the formulation as:

$$d_K^2(x_{iA}^{test}, y_{jB}^{test}) = e_i^T K_A^{test} Q_A^T Q_A (K_A^{test})^T e_i - 2 e_j^T K_B^{test} Q_B^T Q_A (K_A^{test})^T e_i + e_j^T K_B^{test} Q_B^T Q_B (K_B^{test})^T e_j \quad (9)$$

where $K_A^{test} = \phi(A_{test})^T \phi(A_{train}), K_B^{test} = \phi(B_{test})^T \phi(B_{train}) \in R^{N_{test} \times N_{train}}$, Q_A, Q_B are the learned projection matrices, and $\phi(A_{test})$, $\phi(B_{test})$ are constructed in the same way as $\phi(A_{train})$,$\phi(B_{train})$ in Section 2.2. Finally, the gallery images in the test set are ranked based on their distances between the probe images.

3 Experimental Results

Feature Representation and Experimental Setting. We adopt the simple feature representation as in [10] in our experiments. The images are divided into overlapping blocks of size 16×8 and stride 8×8. On each block, we extract HSV and LAB histograms, each with 24 bins per channel. Afterwards, texture information is embedded into the representation by extracting LBP [14] on each block. Finally, the local features are concatenated and projected into a 34-dimensional subspace by PCA as in [10].

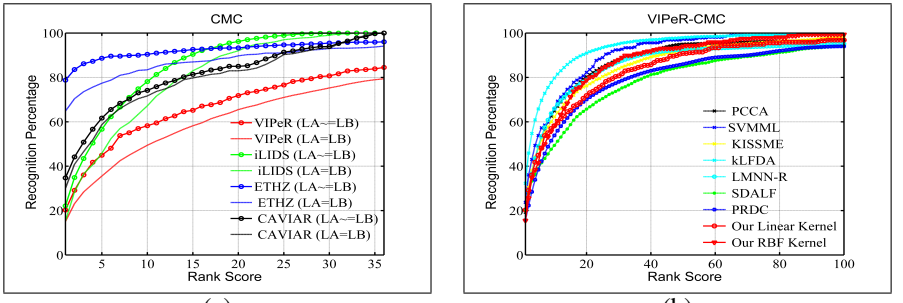
To perform single-shot person re-identification on the three multi-shot benchmark datasets (iLIDS [17], ETHZ [5] and CAVIAR4REID [3]), we first randomly select p people to form the training set, then two images of each person are randomly selected to construct the view-related sets. We also evaluate our algorithm in the multi shot case on the iLIDS dataset. The performance of our algorithm is evaluated by the Cumulative Matching Characteristic (CMC) curve, which represents the expectation of finding the correct match in the top n matches. In each experiment, both the training-testing splits and the random selection of two images for each person (on the iLIDS [17], ETHZ [5] and CAVIAR4REID [3] datasets) are performed 10 *times* and average performance is recorded. Detailed comparison results are presented in the following part.

In the training stage, the projection matrices can not be directly learned in the proposed kernelized method, we instead learn the transform matrices Q_A, Q_B according to Equation (6)(7)(8). In the testing phase, the distance in the kernel space is calculated according to Equation (9), detailed derivation is illustrated in the supplementary material. In our experiments, all the other parameters are automatically tuned by cross validation. Figure 4 presents some re-identification results of our algorithm on the VIPeR dataset.

Comparison with the Baseline Method. To show the benefits of exploiting different projection functions, we compare our algorithm with the baseline algorithm. The baseline algorithm here means setting the projection matrices to be the same in our algorithm while the experimental and parameter settings remain the same. The detailed experimental results are shown in Figure 2(a). As shown in Figure 2(a), our algorithm outperforms the baseline method on all the four challenging datasets. This validates that our algorithm can flexibly adapt to different views and reduce the influence of inter-camera variations than using the same projection matrix.

VIPeR Dataset. VIPeR is the largest and most challenging person re-identification dataset consisting of 632 people with two images from two cameras for each person. It bears great variations in pose and illumination, most of the examples contain a viewpoint change of more than 90 degrees.

We extract simple features as illustrated in the feature representation part and then we randomly select 316 people to form the training set, the left images then form the test set.



(a) (b)

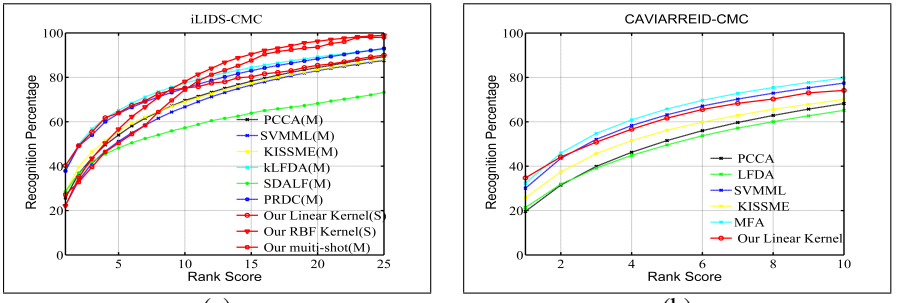
Figure 2: (a) Comparison results with the baseline method. (b) Comparison results on the VIPeR dataset.

The detailed parameter setting on this dataset is as follows: $\mu_A = \mu_B = 0.1$, $\lambda = 0.07$, $\eta_A = \eta_B = 0.001$, which is tuned by cross validation. The comparison results with other state-of-the-art algorithms (PCCA [13], SVMML [11], KISSME [10], kLFDA [16], LMNN-R [4], SDALF [6], PRDC [18]) on VIPeR are shown in Figure 2(b). As we can see, our algorithm (both linear kernel and the RBF kernel) achieves competitive results with existing algorithms. Besides, due to the great variation posed by this challenging dataset, we find that the RBF kernel performs relatively better than the linear kernel. This denotes that the kernel trick catches more information of the features and can lead to a better subspace (features of the same person come closely and faraway otherwise) when the dataset is large with great variation.

iLIDS Dataset. The iLIDS dataset is another publicly available dataset captured at an airport arrival hall. It contains 479 images of 119 pedestrians, with each image subjected to great illumination changes and occlusions.

To better exploit our algorithm (view adaptive), since images in the original iLIDS dataset are mixed with no view information, we manually seperate the images into two sets according to their camera view information like the VIPeR dataset. Then we perform both single-shot and multi-shot person re-dentification with the proposed algorithm. In the single-shot case, we randomly select one image of each pedestrian from the manually separated sets. As for the multi-shot case, all the images of the training set are used to learn the projection matrices. The detailed parameter setting on this dataset is as follows: $\mu_A = \mu_B = 0.03$, $\lambda = 0.1$, $\eta_A = \eta_B = 0.1$, which is tuned by cross validation. The comparison results with some state-of-the-art algorithms (PCCA [13], SVMML [11], KISSME [10], kLFDA [16], SDALF [6], PRDC [18]) are demonstrated in Figure 3(a). As shown in the figure, our algorithm with linear kernel performs very well in the low rank case and achieves the best rank one performance with 41.6 percent recognition rate. The RBF kernel achieves similar low rank recognition rate as some other existing algorithms (PCCA [13], KISSME[10], SVMML[11]), but it converges faster to 100 percent than all the other algorithms, which is a valuable property for the person re-identification algorithm in that it is likely to have high confidence of the returned matches in a relatively low rank. As to the performance of our algorithm in the multi-shot case, we find that it achieves similar performance with RBF kernel in the single-shot case although converges relatively slower, which indicates that more training data may bring benefits to improving the convergence rate when the dataset is small with great variations.

CAVIAR4REID Dataset. The CAVIAR4REID dataset is another widely used dataset



(a) (b)

Figure 3: (a) Comparison results on the iLIDS dataset. (b) Comparison results on the CAVIAR4REID dataset.

for person re-identification. It consists of 72 pedestrians, each containing multiple images captured from two different cameras in an indoor shopping mall. These images cover a wide range of poses and resolutions.

The detailed parameter setting on this dataset is as follows: $\mu_A = \mu_B = 0.03$, $\lambda = 0.1$, $\eta_A = \eta_B = 0.01$, which is tuned by cross validation. As this dataset is very small, with only features of 36 people (total 72 images) for training, we only evaluate the performance of linear kernel in the single-shot setting to avoid over-fitting and compare the result with some state-of-the-art metric learning algorithms. As illustrated in Figure 3(b), our algorithm achieves similar results with LFDA [15], KISSME[10] and SVMML[11] in the low rank case. The PCCA [13] and MFA [16] perform slightly better than our algorithm, but it should be noted that our algorithm only needs two images of each pedestrian, while other algorithms need features of all the images (10 to 20 images for each person) as input.

ETHZ Dataset. We also evaluate our algorithm on the ETHZ dataset, which is captured from moving cameras. The most challenging aspects of this dataset are illumination changes and occlusions. The dataset is structured as follows: *SEQ*.#1 contains 83 pedestrians with total 4857 images, *SEQ*.#2 contains 35 pedestrians with total 1936 images, and *SEQ*.#3 contains 28 pedestrians with total 1762 images.

We combine the three sequences as a whole to perform person re-identification such that it's not likely to overfit. We also experiment in the single-shot setting for the sake of efficiency, randomly selecting two images of each pedestrian to form the training and testing sets. The detailed parameter setting on this dataset is as follows: $\mu_A = \mu_B = 0.03$, $\lambda = 0.1$, $\eta_A = \eta_B = 0.1$, which is tuned by cross validation. On this dataset, we compare our algorithm with [10], which adopts the same feature and has the same single-shot setting as ours. The results are averaged over 10 runs. Detailed Comparison results are shown in Table 1. We can see from the table that our algorithm with linear kernel achieves better

Method	r=1	r=5	r=10	r=15
KISSME [10]	65.51	83.67	87.34	89.59
Our Linear Kernel	**78.77**	**88.63**	**90.41**	**92.60**
Our RBF Kernel	70.55	85.75	89.45	91.23

Table 1: The comparison results of our algorithm with [10] on the ETHZ dataset. In the evaluation, the same features and single-shot setting are adopted for fair comparison.

performance than both the RBF kernel and the algorithm in [10]. This demonstrates the effectiveness of our algorithm and also indicates that linear kernel is a better choice when the dataset is small with relatively small variations in appearance.

4 Conclusions

In this paper, we present a new subspace learning algorithm, which tries to explicitly learn the projection matrices for each camera view, such that the inter-camera variations are minimized in the projected subspace. We also introduce the kernel trick to better tackle the linearly non-separable situation. The linear and RBF kernel are evaluated and the reported performance on four publicly available datasets are compared with some other state-of-the-art algorithms. The extensive experiments demonstrate the effectiveness of our algorithm. We find that the kernel trick helps to improve the recognition rate when the dataset is large (with more people) and bears great variations. Collecting more images for each person (corresponding to multi-shot cases on the iLIDS dataset) also leads to improved performance when the dataset is relatively small with great variation. Besides, we also find that linear kernel is a better choice when the dataset is small with relatively small variations in appearance.

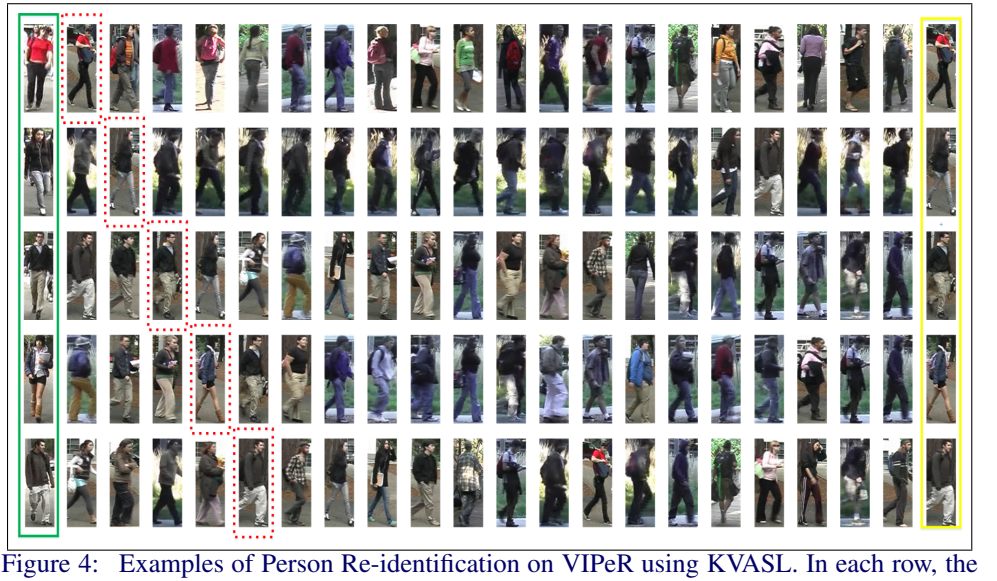


Figure 4: Examples of Person Re-identification on VIPeR using KVASL. In each row, the left-most image in green box is the probe, images in the middle are the top 20 matched gallery images with a highlighted red box for the correctly matched, and the right-most image in the yellow box shows a true match.

5 Acknowledgements

This research is partially supported by the NSFC (No. 61221001 and No. 61171172).

References

[1] Le An, Mehran Kafai, Songfan Yang, and Bir Bhanu. Reference-based person re-identification. In *Proc. AVSS*, pages 244–249, 2013.

[2] Amir Beck. Quadratic matrix programming. *SIAM Journal on Optimization*, 17(4): 1224–1238, 2007.

[3] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Proc. BMVC*, pages 1–11, 2011.

[4] Mert Dikmen, Emre Akbas, Thomas S. Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Proc. ACCV*, pages 501–512, 2010.

[5] Andreas Ess, Bastian Leibe, and Luc J. Van Gool. Depth and appearance for mobile scene analysis. In *Proc. ICCV*, pages 1–8, 2007.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proc. CVPR*, pages 2360 –2367, June 2010.

[7] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proc. ECCV*, pages 262–275, 2008.

[8] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Proc. SCIA*, pages 91–102, 2011.

[9] Martin Hirzer, Peter M. Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *Proc. ECCV*, pages 780–793, 2012.

[10] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Proc. CVPR*, pages 2288–2295, 2012.

[11] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith. Learning locally-adaptive decision functions for person verification. In *Proc. CVPR*, pages 3610–3617, 2013.

[12] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Proc. ECCV*, pages 413–422, 2012.

[13] Alexis Mignon and Frédéric Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *Proc. CVPR*, pages 2666–2672, 2012.

[14] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[15] Sateesh Pedagadi, James Orwell, Sergio A. Velastin, and Boghos A. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proc. CVPR*, pages 3318–3325, 2013.

[16] Fei Xiong, Mengran Gou, Octavia I. Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Proc. ECCV*, pages 1–16, 2014.

[17] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *Proc. BMVC*, pages 1–11, 2009.

[18] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *Proc. CVPR*, pages 649–656, 2011.