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

Person re-identification refers to the task of recognizing the same person under different non-overlapping camera views and across different time and places. We propose a novel algorithm coined as Kernelized View Adaptive Subspace Learning (KVASL), which tries to learn different projection matrices for each camera view to compensate for the specific distortion brought by different views. The kernel trick is adopted to catch more information such that nonlinear transformation is possible. We present the motivating example in Figure 1, which not only demonstrates the benefit of adopting different projection matrices, but also illustrates the necessity of the kernel trick.

Denote $|S|$,$|D|$ as the number of matched and mismatched image pairs in the training set. Our kernelized loss function is formulated as follows:

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

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, $\phi(A_{train}) = \left[\phi(x_{1A}^{train}), \cdots, \phi(x_{NA}^{train})\right]$, $\phi(B_{train}) = \left[\phi(y_{1B}^{train}), \cdots, \phi(y_{NB}^{train})\right] \in R^{m\times N}$ are the matrices formed by feature vectors in the kernel space of the corresponding camera views, $tr(\cdot)$ 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.

We adopt an alternately iterative gradient descent method to optimize our loss function. The gradients of Q_A,Q_B in Eq.(1) can be derived as follows:

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

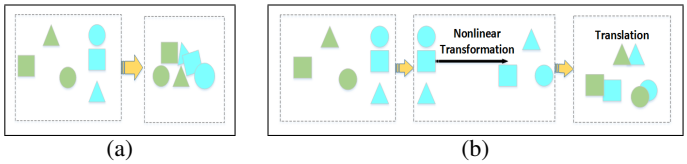
where X is the same matrix as in Eq.(1). In this case, the matrices to be learned can be updated in the following way:

$$Q_A(t+1) = Q_A(t) - \eta_{Q_A}\frac{\partial \ell}{\partial Q_A}, \quad Q_B(t+1) = Q_B(t) - \eta_{Q_B}\frac{\partial \ell}{\partial Q_B} \tag{3}$$

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, the distance between two test images can be derived as follows:

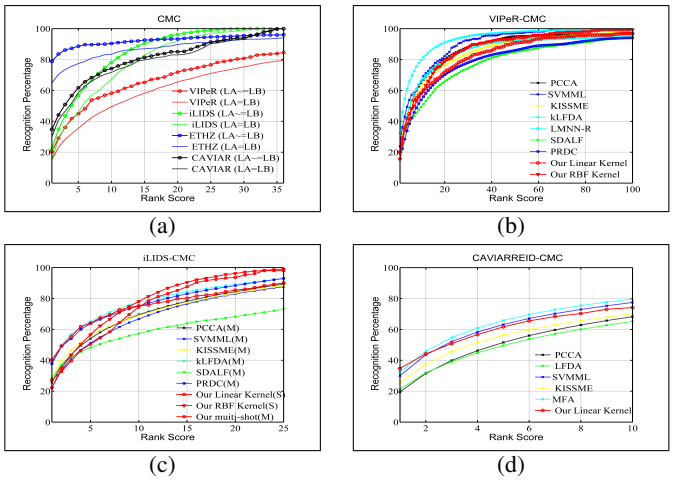
$$\begin{aligned}d_K^2(x_{iA}^{test}, y_{jB}^{test}) &= e_i^T K_A^{test} Q_A^T Q_A (K_A^{test})^T e_i \\ &- 2e_j^T K_B^{test} Q_B^T Q_A (K_A^{test})^T e_i + e_j^T K_B^{test} Q_B^T Q_B (K_B^{test})^T e_j\end{aligned} \tag{4}$$

We implement our algorithm on four publicly available datasets, the comparison results with the baseline method demonstrate the superiority of view-adaptive projection matrices over using the same projection matrices. We also compare our algorithm with some state-of-the-art algorithms (SVMML [2], KISSME [1], kLFDA [3] et al.), which demonstrates the effectiveness of our algorithm. Detailed comparison results are shown in Figure 2 and Table 1.



(a) (b)

Figure 1: A conceptual illustration of how different projection functions can be more flexible than the same ones. (a)The same transformation; (b) Different transformations.



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

Figure 2: (a) Comparison results with the baseline method.(b) Results on VIPeR dataset.(c) Results on iLIDS dataset.(d) Results on CAVIAR4REID dataset. Red lines correspond to results of our algorithm.

Method	r=1	r=5	r=10	r=15
KISSME [1]	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: Comparison of our algorithm with [1] on the ETHZ dataset.

[1] 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.

[2] 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.

[3] 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.