LOCALIZED MULTI-KERNEL DISCRIMINATIVE CANONICAL CORRELATION ANALYSIS FOR VIDEO-BASED PERSON RE-IDENTIFICATION

Guangyi Chen[1,2,3], *Jiwen Lu*[1,2,3,*], *Jianjiang Feng*[1,2,3], *and Jie Zhou*[1,2,3]

[1]Department of Automation, Tsinghua University, Beijing, China
[2]State Key Lab of Intelligent Technologies and Systems, Beijing, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, China
{chen-gy16,lujiwen,jfeng,jzhou}@tsinghua.edu.cn

ABSTRACT

This paper presents a localized multi-kernel discriminative canonical correlation analysis (LMKDCCA) approach for video-based person re-identification, which aims to match persons from pedestrian videos captured by non-overlapping cameras. Unlike conventional methods, our approach models each pedestrian video as a point on the Riemannian manifold and learns similarity over these points under the multiple kernel learning framework. For each given person video, we first represent it as a symmetric positive definite (SPD) matrix which lies on a Riemannian manifold and compute the similarity of multiple SPDs. Then, we develop an LMKDCCA algorithm to learn a nonlinear distance metric which effectively combines these SPDs to exploit complementary information for similarity measure. Experimental results on the iLIDS-VID and PRID 2011 datasets show that our approach achieves the state-of-the-arts.

***Index Terms*—** Person re-identification, canonical correlation analysis, multiple kernel learning

1. INTRODUCTION

Person re-identification, which refers to matching pedestrians across non-overlapping cameras, has numerous potential applications in visual surveillance and receives increasing interests in recent years [1]. It is a challenging problem since the image or video quality is intrinsically limited by the complex inter-camera variances like variations of camera viewpoints, poses, illumination changes and partial occlusions. Many existing works [2–4] focus on either robust appearance feature representation or discriminative metric learning on still images to reduce the influence of inter-camera variances.

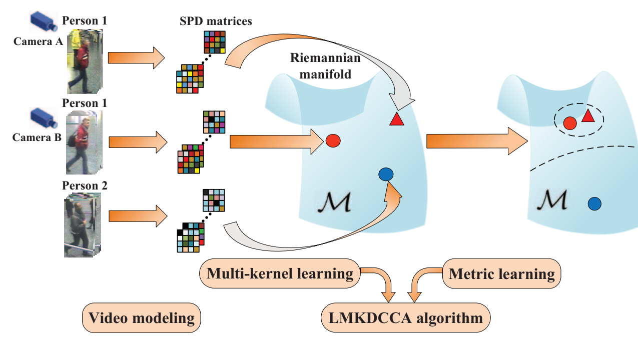

Fig. 1. The basic idea of our approach. For each video, we model it with multiple SPD matrices indicted by different kernels, which lie on a Riemannian manifold. Then, we develop an LMKDCCA algorithm to iteratively combine SPDs and learn a nonlinear distance metric.

Instead of still image, many video based person re-identification approaches have been developed in recent years [5–15]. The reason is that videos contain more abundant spatial-temporal information about the motion of pedestrians, and weaken the disturbance of pose variations and occlusion. Existing video based works can be divided into two categories: *video representation* and *metric learning*. Video representation methods focus on the utilization of temporal information. For example, methods in [5, 13] combine the inter-frame information with appearance information, and Liu *et al*. [8] develop a space-time body-action method. Metric learning methods concentrate on reducing the intra-class variance, such as ranking and selecting video segments [5], learning a dictionary to sparsely encode features [7], and learning the top-push distance [11]. Although conventional video-based methods have made great progress, there are several challenges still unsolved, such as the representation of video, the similarity of walking actions and the increase of intra-class variance.

* Corresponding author.
This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001004, the National Natural Science Foundation of China under Grants 61672306, 61572271, 61527808, 61373074 and 61373090, the National 1000 Young Talents Plan Program, the National Basic Research Program of China under Grant 2014CB349304, the Ministry of Education of China under Grant 20120002110033, and the Tsinghua University Initiative Scientific Research Program.

To address these limitations, we propose an LMKDCCA approach which represents videos as multiple SPD matrices and learns a metric under the multi-kernel learning framework. Figure 1 illustrates the basic idea of our approach. Unlike previous space-time features such as Optical Flow Energy [13] and HOG3D [16], we model videos by multiple SPD matrices on a Riemannian manifold. The manifold embodies the inherent structure and the combined representation overcomes the influence of various person poses and video lengths. Furthermore, our LMKDCCA algorithm projects the SPD matrices to a pair of discriminative geodesic subspaces and combines multiple SPDs locally to get a robust accurate video representation with complementary information. Performance evaluations on two available video datasets including PRID 2011 [17] and iLIDS-VID [5] show the effectiveness of our proposed approach.

2. APPROACH

In this section, we describe the LMKDCCA approach which models videos by multiple SPD matrices and learns a nonlinear distance metric in detail.

2.1. Video Modeling by SPD Matrix

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a pedestrian video containing n frames, where $x_i \in R^d$ denotes the feature of the ith frame in the video. We model the video as multiple $d \times d$ SPD matrices $S^m = \{S^m_{ij}\}$, which lie on a Riemannian manifold. One effective SPD representing method is induced by kernel function as [18]:

$$S_{kernel}(i,j) = \langle \phi(f_i), \phi(f_j) \rangle = \kappa(f_i, f_j), \tag{1}$$

where $\phi(\cdot)$ is an implicit nonlinear mapping function, $\langle \cdot, \cdot \rangle$ means inner product, $\kappa(\cdot, \cdot)$ is the corresponding kernel function, and $f_i, 1 \le i \le d$ is the ith row of X.

Different kennel functions will induce different SPD matrices. For example, we employ the Gaussian RBF kernel to get the SPD matrix as:

$$\kappa_{RBF}(f_i, f_j) = exp(-\gamma ||f_i - f_j||^2), \tag{2}$$

where γ is the radial scale parameter of Gaussian RBF kernel.

Furthermore, to fuse appearance feature with kernel matrices, we suppose that frames x_i obey a Gaussian distribution $N(\mu, \Sigma)$, where μ is the mean of frame features, and Σ is the covariance matrix. Then we get the SPD matrix by [19]:

$$S_{Gaussian} = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}. \tag{3}$$

Compared with previous pedestrian video representing methods, modeling videos with SPD matrices has two advantages. 1) The SPD matrix lies on a Riemannian manifold, which embodies the inherent structure of data and concentrates the discriminative parts of videos. 2) Our representation exploits correlation information of all frames instead of adjacent frames, which is not affected by the variation of poses, video lengths and occlusions. Moreover, by combining multiple SPD matrices calculated in different ways, we exploit complementary information to represent person videos robustly and validly.

2.2. Localized Multi-kernel Discriminative CCA

Having modeled pedestrian videos captured from two cameras as SPD matrices $X_i \in Sym^+_d$ and $Y_i \in Sym^+_d$ on a Riemannian manifold, we learn projection operations $f_x(\cdot), g_y(\cdot)$ to map points on the manifolds into the best pair of geodesic subspaces respectively. In the projection subspaces, the similarity between SPDs is measured more accurately. In order to learn discriminative projection operations, we maximize inter-class variations and minimize intra-class variations with an optimization problem as follows:

$$\begin{aligned} &\min \sum_{i,j} \omega_{ij} ||f_x(X_i) - g_y(Y_j)||^2_F \\ &s.t \sum_i ||f_x(X_i)||^2_F = 1, \sum_j ||g_y(Y_j)||^2_F = 1, \end{aligned} \tag{4}$$

where ω_{ij} is the pair-wise label which equals to 1 if X_i and Y_j belong to the same people. Otherwise, it equals to -1/n.

Generally, it is difficult to compute the distance between two points on the Riemannian manifold directly. Hence, we introduce the implicit function $\phi : X_i \to \phi(X_i)$ which maps SPD matrices to a Hibert space. Thus, the projection operations are written as linear matrices lying on the span of projected training data. And we employ the kernel trick on the Riemannian manifold to calculate the distance between projected points:

$$W_x^T \phi(X_i) = \sum_n \alpha_n \phi(X_n)^T \phi(X_i) = \sum_n \alpha_n K(X_n, X_i), \tag{5}$$

$$W_y^T \phi(Y_i) = \sum_n \beta_n \phi(Y_n)^T \phi(Y_i) = \sum_n \beta_n K(Y_n, Y_i), \tag{6}$$

where W_x, W_y are original linear projection matrices, and $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^T, \beta = [\beta_1, \beta_2, \ldots, \beta_n]$ are projection operations induced by kernel function. $K(\cdot, \cdot)$ means the kernel function and we apply the kernel proposed by Wang *et al.* in [20], which is effective and easy to be calculated:

$$K(X_i, X_j) = tr[log(X_i) \cdot \log(X_j)]. \tag{7}$$

Then, with this kernel function, we compute the distance of two SPD matrices on the Riemannian manifold and reformulate the optimization function (4) in the following:

$$\begin{aligned} &\min \sum_{i,j} \omega_{ij} ||\alpha^T K_x^{(i)} - \beta^T K_y^{(j)}||^2_F \\ &s.t \alpha^T \alpha = I, \beta^T \beta = I, \end{aligned} \tag{8}$$

where $K_x^{(i)}, K_y^{(i)}$ stand for the ith column of kernel matrices.

$$\begin{aligned} K_x^{(i)} &= [K(X_1, X_i), K(X_2, X_i), \ldots, K(X_n, X_i)]^T \\ K_y^{(i)} &= [K(Y_1, Y_i), K(Y_2, Y_i), \ldots, K(Y_n, Y_i)]^T, \end{aligned} \tag{9}$$

Furthermore, we weight different SPD matrices obtained in different ways to exploit complementary information and get more accurate video representing. Different from other methods which assume weights are same to all the samples, we argue that weights should be data-adaptive and introduce a localized multi-kernel learning [21] algorithm, which combines multiple SPD matrices locally. In this framework, we apply the gating function $\eta_m(\cdot)$, which is learned from data to combine SPD matrices as follows:

$$\begin{aligned} Kx_\eta(i,j) &= \sum_{m=1}^{p} \eta_m(K_x^{m(i)}) K_x^m(i,j) \eta_m(K_x^{m(j)}) \\ Ky_\eta(i,j) &= \sum_{m=1}^{p} \eta_m(K_y^{m(i)}) K_y^m(i,j) \eta_m(K_y^{m(j)}), \end{aligned} \tag{10}$$

where the K_x^m, K_y^m are kernel matrices computed by the mth pair of SPD matrix representations about videos X, Y. And $K^{m(i)}$ means the ith column of K^m. We select the softmax function [22] as gating function $\eta_m(\cdot)$ due to its nonnegativity and monotonicity:

$$\eta_m(K_x^{m(i)}) = \frac{exp(v_m^T K_x^{m(i)} + v_{m0})}{\sum_{k=1}^{p} exp(v_k^T K_x^{m(i)} + v_{k0})}, \tag{11}$$

where the $v_k \in R^{n \times 1}$ and $v_{k0} \in R^1$ are the parameters to be learned. To this end, the final objective function can be written as follows by applying Lagrange multiplier method:

$$\begin{aligned} \min \mathcal{L} = &\sum_{i,j} \omega_{ij} ||\alpha^T K x_\eta^{(i)} - \beta^T K y_\eta^{(j)}||_F^2 \\ &+ \mu ||\alpha||_F^2 + \mu ||\beta||_F^2. \end{aligned} \tag{12}$$

To solve the optimization problem in (12), we update projection matrices α, β and the parameters v_m, v_{m0} of gating function η_m by using an iterative algorithm. Firstly, we initialize the parameters v_m, v_{m0}, and compute the weighted SPD matrix and corresponding kernel Kx_η, Ky_η. The problem turns to be the classical kernel based CCA (8) and can be solved by the following generalized eigenvalue problem:

$$\begin{bmatrix} 0 & K_{xy} \\ K_{xy}^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \Lambda \begin{bmatrix} K_{xx} & 0 \\ 0 & K_{yy} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \tag{13}$$

where $K_{xy} = Kx_\eta Ky_\eta, K_{xx} = Kx_\eta Kx_\eta, K_{yy} = Ky_\eta Ky_\eta$. Then, by fixing α, β, we use gradient descent method to update v_m and v_{m0} as follows:

$$\begin{aligned} v_m^{t+1} &= v_m^t - l \frac{\partial \mathcal{L}}{\partial v_m} \\ v_{m0}^{t+1} &= v_{m0}^t - l \frac{\partial \mathcal{L}}{\partial v_{m0}}, \end{aligned} \tag{14}$$

Algorithm 1: LMKDCCA

Input: Training data with different representing K_x^m, K_y^m, learning rate l, iteration number N and convergence condition ϵ.

Output: projection matrices α, β and the parameters of gating function v_m, v_{m0}

```
Initialize the parameters v_m^0, v_m0^0;
for all t = 1, 2, ..., N do
    Get the weighted kernel Kx_η, Ky_η by (10);
    Solve the eigenvalue problem in (13)
    Obtain projection matrices α, β;
    Update v_m, v_m0 by using (14);
    if t > 1 and |L_t − L_{t−1}| < ϵ then go to return
    end if
end for
return α, β and v_m, v_m0
```

where l is the learning rate. Algorithm 1 summaries the detail procedure of our method LMKDCCA.

Moreover, we learn projection operators for the mean of appearance representation with the same learning framework as (4), which is formulated by classical CCA [24] about conventional vector feature $\bar{x}, \bar{y}$ and linear projection matrices W_x, W_y. Finally, we measure the similarity of person videos by combing the distance of SPD matrix representation and average appearance representation.

In the testing stage, given the gallery videos $G = \{G_k\}$ and probe video P, we first compute kernel matrices as (10) with learned parameters. Then, we respectively calculate the scores of SPD matrix representations and average appearance representation. Finally, we match the P with G_k by combining the two scores with a rate σ:

$$k = arg \min_k d_S(G_k, P) + \sigma d_M(G_k, P), \tag{15}$$

where σ, as a normalization coefficient, is equal to the ratio between norms of two features.

3. EXPERIMENTS

3.1. Datasets and Setting

We evaluated our method on two available pedestrian video datasets including iLIDS-VID [5] and PRID 2011 [17]. The iLIDS-VID dataset contains 600 pieces of videos for 300 randomly sampled people, which have variable lengths from 23 to 193 frames. We randomly selected half of pedestrians for training and the other are used to test, and took the average cumulative matching characteristic (CMC) cure in ten trials as the evaluating indicator, referring to the experiment settings in [5]. For the PRID 2011 dataset which includes 400 videos with 5 to 675 frames, we selected 178 persons with more than 27 frames in both cameras. The dataset was randomly divided into training set and testing set by half, which is same to [5].

Table 1. Comparison with state-of-the-art person re-identification methods on the iLIDS-VID and PRID 2011 datasets.

Method	iLIDSVID				PRID 2011			
	Rank=1	Rank=5	Rank=10	Rank=20	Rank=1	Rank=5	Rank=10	Rank=20
DynFV+LDFV [15]	28.8	55.0	70.6	82.0	43.6	69.0	79.4	92.7
DVDL [7]	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6
SDALF+DVR [5]	41.3	63.5	72.7	83.1	48.3	74.9	87.3	94.4
TDL [11]	56.7	80.0	87.6	93.6	56.3	87.6	95.6	98.3
McLaughlin [12]	58.0	84.0	91.0	96.0	70.0	90.0	95.0	97.0
AvgTAPR [14]	55.0	87.5	93.8	97.2	68.6	94.6	97.4	98.9
mvRMLLC+ST+Alignment [13]	69.1	89.9	**96.4**	**98.5**	66.8	91.3	96.2	98.8
STFV3D+KISSME [8]	44.3	71.7	83.7	91.7	64.1	87.3	89.9	92.0
LOMO+KISSME+SRID [23]	65.5	85.4	91.3	95.7	83.0	95.3	97.5	99.3
LOMO+SPD	29.4	56.8	69.7	81.9	51.2	83.5	92.2	97.0
KDCCA+SPD	48.7	80.5	89.4	96.1	70.6	93.6	98.4	99.8
KDCCA+appearance	60.3	80.6	87.3	90.9	76.7	92.8	95.9	98.0
GMKDCCA	70.6	90.1	93.8	97.3	83.0	96.1	99.4	99.8
LMKDCCA	**73.3**	**90.5**	94.7	98.1	**86.4**	**97.5**	**99.6**	**100**

In the experiments, we extracted the LOMO [2] feature as the original representation for each frame in the videos. Moreover, with the rate of positive samples and negative samples, $r = 1 : 2$, and the step size, $l = 1.5 \times 10^{-4}$, the loop of iterative optimization algorithm in Algorithm 1 stopped at the 12th iteration.

3.2. Results and Analysis

We report the performance of our method LMKDCCA and most existing video based person re-identification approaches in Table 1. The comparison with the state-of-the-art methods and some contrast experiments is as follows.

Compared with state-of-the-art methods: The first group of Table 1 tabulates the matching rate of the state-of-the-art methods on both two datasets. The results show that our proposed approach outperforms than most other state-of-the-art methods. For instance, with the same feature, the Rank-1 matching rate of our method is 4.1% higher than the KISSME-SRID method on the PRID 2011 dataset and 11% on the iLIDS-VID dataset. Compared with methods which model videos with the inter-frame spatial-temporal information like DVR and STFV3D+KISSME, our approach improves 65% and 70% respectively on the iLIDS-VID dataset. The reason is that our method takes the connection of all frames of video into account on the Riemannian manifold, while most compared method just consider the information of adjacent frames. However, similarity of samples is limited with the discrimination increasing, the Rank-10 and Rank-20 matching rate of our performance are less than mvRMLLC method. Moreover, TDL [11] and mvRMLLC+ST+Alignment [13] have higher performance on iLIDS-VID dataset, KISSME-SRID is particularly effective on the other, while our LMKDCCA algorithm performs outstanding on both two datasets because we combine different SPD matrix representations to exploit complementary information.

Evaluations with different adjustment: In the second group of Table 1, we analysed our approach LMKDCCA by comparing with different adjustments. We modeled videos by Gaussian SPD matrix representation with LOMO feature as the baseline. And we evaluated our basic kernel based discriminative CCA (KDCCA) method which learns the metric on the Riemannian manifold. It gets a great improvement that rises the Rank-1 matching rate almost 20% on the both two datasets and achieves the comparable performance with the method applying the average appearance representation as feature. In addition, we combined three SPD matrices and the mean of appearance feature to exploit complementary information for discriminative similarity measure, where two SPD matrices are induced by RBF kernel with $\gamma = 1, 10$ in (2) and one is Gaussian SPD matrix in (3). Compared with GMKDCCA method [25] which combines the SPD matrices for all the samples in a same weight, our proposed LMKDCCA method weights different SPD matrix representations locally and gets a better performance due to the considering characteristics of each sample.

4. CONCLUSIONS

In this work, we have proposed a SPD matrix model on the Riemannian manifold to represent person videos for re-identification. The model contains the global inter-frame connection and overcomes the variations of pose, video length and occlusion. In addition, we have developed a localized multi-kernel based discriminative canonical correlation analysis algorithm (LMKDCCA) method, to weight different SPD matrices for a more valid representing and learn a distance metric. We evaluated our method on two public video person re-identification datasets, and demonstrated the superiority of our proposed approach over state-of-the-art methods. For future work, we are studying a more effective metric over the Riemannian manifold to compute the geodesic distance between two person videos.

5. REFERENCES

[1] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016, pp. 1335–1344.

[2] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[3] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372.

[4] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVPR*, June 2015, pp. 1846–1855.

[5] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by video ranking," in *ECCV*, 2014, pp. 688–703.

[6] Srikrishna Karanam, Yang Li, and Richard J Radke, "Sparse re-id: Block sparsity for person re-identification," in *CVPR Workshops*, 2015, pp. 33–40.

[7] Srikrishna Karanam, Yang Li, and Richard J Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015, pp. 4516–4524.

[8] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *ICCV*, 2015, pp. 3810–3818.

[9] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*. Springer, 2016, pp. 868–884.

[10] Xiaoke Zhu, Xiao-Yuan Jing, Fei Wu, and Hui Feng, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," in *IJCAI*, 2016, pp. 3552–3559.

[11] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng, "Top-push video-based person re-identification," in *CVPR*, June 2016, pp. 1345–1353.

[12] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, June 2016, pp. 1325–1334.

[13] Jiaxin Chen, Yunhong Wang, and Yuan Y Tang, "Person re-identification by exploiting spatio-temporal cues and multi-view metric learning," *IEEE SPL*, vol. 23, no. 9, pp. 998–1002, 2015.

[14] Changxin Gao, Jin Wang, Leyuan Liu, Jin-Gang Yu, and Nong Sang, "Temporally aligned pooling representation for video-based person re-identification," in *ICIP*, 2016, pp. 4284–4288.

[15] Mengran Gou, Xikang Zhang, Angels Rates-Borras, Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps, "Person re-identification in appearance impaired scenarios," in *BMVC*, pp. 1–10.

[16] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008, pp. 1–10.

[17] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *SCIA*, 2011, pp. 91–102.

[18] Lei Wang, Jianjia Zhang, Luping Zhou, Chang Tang, and Wanqing Li, "Beyond covariance: Feature representation with nonlinear kernel matrices," in *ICCV*, 2015, pp. 4570–4578.

[19] Peihua Li, Qilong Wang, and Lei Zhang, "A novel earth mover's distance methodology for image matching with gaussian mixture models," in *ICCV*, 2013, pp. 1689–1696.

[20] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *CVPR*, 2012, pp. 2496–2503.

[21] Mehmet Gönen and Ethem Alpaydin, "Localized multiple kernel learning," in *ICML*, 2008, pp. 352–359.

[22] Jiwen Lu, Gang Wang, and Pierre Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *ICCV*, 2013, pp. 329–336.

[23] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," *arXiv preprint arXiv:1605.09653*, 2016.

[24] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[25] Yen-Yu Lin, Tyng-Luh Liu, and Chiou-Shann Fuh, "Multiple kernel learning for dimensionality reduction," *TPAMI*, vol. 33, no. 6, pp. 1147–1160, 2011.