LPCV: LEARNING PROJECTIONS FROM CORRESPONDING VIEWS FOR PERSON RE-IDENTIFICATION

Hong Liu, Qiao Guan

Key Laboratory of Machine Perception (Ministry of Education)
Shenzhen Graduate School, Peking University
{hongliu, guanqiao}@pku.edu.cn

ABSTRACT

Person re-identification is an important topic in visual surveillance, which aims at recognizing an individual over disjoint camera views. As a major aspect of person re-identification, distance metric learning has been widely studied to seek a discriminative matching metric. However, most existing distance metric learning methods learn an identical projection matrix for all camera views, while ignoring the own characteristic of each view. To address this issue, we propose a novel method to learn projections from corresponding views (LPCV) for person re-identification. First, we use the labeled features to learn different projections for different views. Then, these projections are used to transform tested features into a new feature space. Finally, we use this new feature space to identify a person from one camera to another with a standard nearest-neighbor voting method. Experimental results on three challenging datasets VIPeR, PRID 450S and CUHK01 demonstrate that our method significantly performs favorably against the state-of-the-art methods, especially on the rank-1 matching rate.

Index Terms— Person Re-identification, Distance Metric Learning, Projection Matrix

1. INTRODUCTION

Person re-identification is a task of matching persons observed from non-overlapping camera views. It has many important applications in video surveillance [1] including threat detection, human retrieval, human tracking and activity analysis. As a hot topic, person re-identification has gained much attention among researchers in recent years and many methods [2] have been proposed to advance this field. However, it still remains a challenging problem since a person observed under different camera views often undergoes significant variations on viewpoints, poses, appearances and illuminations, which make intra-personal variations even larger than inter-personal variations.

The existing person re-identification methods mainly focus on either developing discriminative feature representations [3–6] or learning a distance metric [7–11]. In this paper, we mainly focus on distance metric learning. Given any feature representation and a set of training data consisting of matching image pairs across camera views, the objective is to learn a projection matrix that maximizes the inter-person divergence and minimizes the intra-person divergence.

However, traditional distance metric learning methods [7–11] learn an identical projection matrix for all camera views, while ignoring the nature that images under different camera views have different characteristics, e.g. individuals under the same camera views have more similar illumination and viewpoint than those under different camera views. Since there are a lot of differences between different camera views, it is reasonable to assume that learning a unique projection matrix for all views would be difficult to address this issue. Therefore, these methods do not make full use of the characteristic of each view and would probably lose some discriminative information. This issue thus limits the application of the existing methods and may result in a performance degradation.

In this paper, we propose a distance metric learning method, which learns projections from corresponding views. Specifically, we first learn different projection matrices for different camera views. Then, each image under different views is projected into a new feature space. Finally, we implement this method with a standard re-identification pipeline and show the improvements to the re-identification performance by thorough experiments on VIPeR, PRID 450S and CUHK01 datasets.

Relation to prior work: Mignon *et al*. [7] proposed PCCA to learn a projection with sparse pairwise similarity/dissimilarity constraints. Kostinger *et al*. [8] proposed KISSME to derive a Mahalanobis metric by computing the difference between the intra-class and inter-class covariance matrix. Zheng *et al*. [10] proposed PRDC that maximizes the probability of a pair of correctly matched images to have a smaller distance than that of an incorrectly matched pair.

This work is supported by National High Level Talent Special Support ProgramNational Natural Science Foundation of China (NSFC, No.61340046,61673030U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011), Natural Science Foundation of Guangdong Province (No.2015A030311034), Scientific Research Project of Guangdong Province (No.2015B010919004).

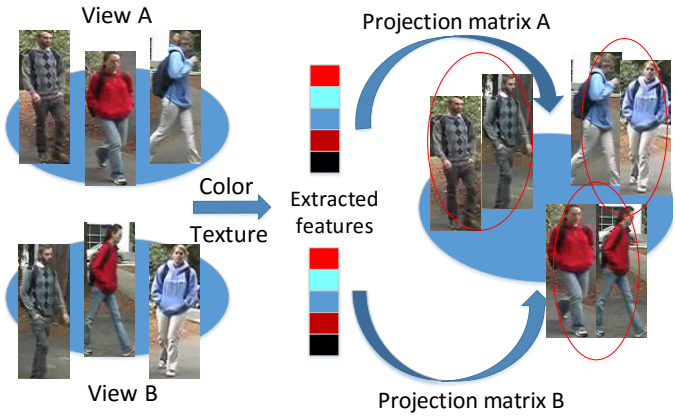


Fig. 1: **Flowchart of the proposed method**. After feature extraction, features extracted from different camera views are projected into a new feature space, the same person from different camera views are more likely to match.

Although having achieved inspiring re-identification results, these methods did not give sufficient consideration to the individual of each camera view when learning an identical projection matrix for all views, which may not be optimal when applying to person re-identification. This observation motivates us to explore a new way to make full use of the characteristics of different camera views.

2. ALGORITHM DESCRIPTION

In this section, we elaborate on how we perform our method to learn different projections for different camera views. The flowchart of our method is given in Fig.1. Existing methods learn an identical projection matrix for all camera views [7–11], i.e., most of them are based on the following distance function to measure the distance between the two cross-view samples $\mathbf{x}$ and $\mathbf{z}$.

$$D^2_{\mathbf{M}}(\mathbf{x},\mathbf{z}) = (\mathbf{x}-\mathbf{z})^T\mathbf{M}(\mathbf{x}-\mathbf{z}) = ||\mathbf{L}^T\mathbf{x}-\mathbf{L}^T\mathbf{z}||_2^2 \quad (1)$$

where $\mathbf{M}$ is a positive semidefinite matrix and can be factorized into $\mathbf{M} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is the projection matrix for all of the cross-view samples. This method doesn't make full use of characteristics of different camera views and thus may lose some discriminative information. To solve this problem, the learning projections from corresponding views (LPCV) method is introduced. For cross-view samples $\mathbf{x}$ (from view A) and $\mathbf{z}$ (from view B), after projection learning, we learn two different projection matrices $\mathbf{L}_A$ and $\mathbf{L}_B$. The details of our method are given below.

2.1. Learning Projections From Corresponding Views

Suppose we have a cross-view training set $\{\mathbf{X},\mathbf{Z}\}$, where $\mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m) \in \mathbf{R}^{d\times m}$ contains m samples in a d-dimensional space from one view and $\mathbf{Z} = (\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_n) \in \mathbf{R}^{d\times n}$ contains n samples in the same d-dimensional space but from the other view. Here $\mathbf{L}_A,\mathbf{L}_B \in \mathbf{R}^{d\times t}$ are the projection matrices of different camera views A and B. Note that $\mathbf{Z}$ is the same as $\mathbf{X}$ in the single-view matching scenario. The goal is to learn a distance function

$$D^2_{A,B}(\mathbf{x},\mathbf{z}) = ||\mathbf{L}_A^T\mathbf{x}-\mathbf{L}_B^T\mathbf{z}||_2^2 \quad (2)$$

where $\mathbf{x}$ and $\mathbf{z}$ are the features of each image under camera view A and B. To get a closed-form solution, we use loss function as follows rather than the log-logistic loss function in [10, 12]

$$F(\mathbf{L}_A,\mathbf{L}_B) = \sum_{i=1}^{m}\sum_{j=1}^{m} w^{i,j}_{A,A}D^2_{A,A}(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} w^{i,j}_{B,B}D^2_{B,B}(\mathbf{z}_i,\mathbf{z}_j) + \sum_{i=1}^{m}\sum_{j=1}^{n} w^{i,j}_{A,B}D^2_{A,B}(\mathbf{x}_i,\mathbf{z}_j) \quad (3)$$

where $w^{i,j}_{A,B}$ is the weight parameter between the i-th image of view A and j-th image of view B. We call (x_i,z_j) a positive sample pair if they are from the same class, and a negative pair otherwise. We define $w^{i,j}_{A,B}$ as:

$$w^{i,j}_{A,B} = \begin{cases} 1 & \text{if } (x_i,z_j) \in N_{pos} \\ \lambda \dfrac{\sum_{i,j\in N_{pos}} dist(\mathbf{x}_i,\mathbf{z}_j)}{\sum_{i,j\in N_{neg}} dist(\mathbf{x}_i,\mathbf{z}_j)} & \text{otherwise} \end{cases} \quad (4)$$

where $\sum_{i,j\in N_{pos}}(\cdot)$, $\sum_{i,j\in N_{neg}}(\cdot)$ denotes the total Euclidean distance of all positive and negative sample pairs, $dist(\mathbf{x}_i,\mathbf{z}_j)$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{z}_j$. λ is a negative value which not only controls the weight of inter-person distance but also contributes to the loss. This asymmetric weighting is important because positive and negative pairs are heavily unbalanced. The Eq.(3) has three parts in total, the first part represents the total distance under the view A. In a similar way, the second part represents the total distance under the view B and the third part represents the total distance between cross-view A and B. In this way, minimizing the loss Function F will reduce the intra-class distance and meanwhile enlarge the inter-class distance.

In order to avoid singularity of the covariance matrix, we add some constraints to the loss function. As a result, the problem is formulated as:

$$\begin{aligned} &min\ \ F(\mathbf{L}_A,\mathbf{L}_B) \\ &s.t.\ \ \mathbf{L}_A^T(\mathbf{X}\mathbf{X}^T+\gamma\mathbf{I}_1)\mathbf{L}_A = \mathbf{I} \\ &\qquad \mathbf{L}_B^T(\mathbf{Z}\mathbf{Z}^T+\gamma\mathbf{I}_1)\mathbf{L}_B = \mathbf{I} \end{aligned} \quad (5)$$

where $\mathbf{I}_1$ and $\mathbf{I}$ denote the corresponding identity matrix, $\mathbf{I}_1$ is the added constraint to avoid singularity and γ is the weight parameter. These constraints ensure the projected features of each view have unit amplitude and thus they are not shrunken to zero.

2.2. A Closed-Form Solution

Since each part is quite similar, we do not present the solution to all parts in detail. Take the first part as an example:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} w^{i,j}_{A,A}D^2_{A,A}(\mathbf{x}_i,\mathbf{x}_j) = tr\Big(2\mathbf{L}_A^T\mathbf{X}(\mathbf{W}_1-\mathbf{W}_2)\mathbf{X}^T\mathbf{L}_A\Big) \quad (6)$$

Here X=($\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_m$), $\mathbf{W}_1$ is a diagonal matrix whose i-th row and i-th column diagonal entries are defined as $\sum_{j=1}^{m} w_{A,A}^{i,j}$, $\mathbf{W}_2$ is a $m \times m$ matrix whose i-th row and j-th column element is $w_{A,A}^{i,j}$. In a similar way,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{B,B}^{i,j} D_{B,B}^2(\mathbf{z}_i, \mathbf{z}_j) = tr\Big(2\mathbf{L}_B^T\mathbf{Z}(\mathbf{W}_3 - \mathbf{W}_4)\mathbf{Z}^T\mathbf{L}_B\Big) \tag{7}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} w_{A,B}^{i,j} D_{A,B}^2(\mathbf{x}_i, \mathbf{z}_j) = tr\Big(\mathbf{L}_A^T\mathbf{X}\mathbf{W}_5\mathbf{X}^T\mathbf{L}_A + \mathbf{L}_B^T\mathbf{Z}\mathbf{W}_6\mathbf{Z}^T\mathbf{L}_B - \mathbf{L}_A^T\mathbf{X}\mathbf{W}_7\mathbf{Z}^T\mathbf{L}_B - \mathbf{L}_B^T\mathbf{Z}\mathbf{W}_8\mathbf{X}^T\mathbf{L}_A\Big) \tag{8}$$

where $\mathbf{W}_i$(i=1,2,$\cdots$ 8) only corresponds to w. The optimization problem of Eq.(3) can be written as below:

$$tr\left(\begin{pmatrix} \mathbf{L}_A^T & \mathbf{L}_B^T \end{pmatrix} \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{L}_A \\ \mathbf{L}_B \end{pmatrix} \right) \tag{9}$$

where $\mathbf{W}_{11} = \mathbf{X}(2\mathbf{W}_1 - 2\mathbf{W}_2 + \mathbf{W}_5)\mathbf{X}^T$, $\mathbf{W}_{12} = \mathbf{X}(-\mathbf{W}_7)\mathbf{Z}^T$,$\mathbf{W}_{21} = \mathbf{Z}(-\mathbf{W}_8)\mathbf{X}^T$, $\mathbf{W}_{22} = \mathbf{Z}(2\mathbf{W}_3 - 2\mathbf{W}_4 + \mathbf{W}_6)\mathbf{Z}^T$, so the optimization problem can be modified as :

$$\begin{aligned} min \quad & tr(\mathbf{L}^T\mathbf{W}\mathbf{L}) \\ s.t. \quad & \mathbf{L}^T\mathbf{M}\mathbf{L} = \mathbf{I} \end{aligned} \tag{10}$$

where $\mathbf{L}^T = (\mathbf{L}_A^T, \mathbf{L}_B^T)$, $\mathbf{M}$ is a block diagonal matrix defined as $\mathbf{M}$=diag($\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I}_1$, $\mathbf{Z}\mathbf{Z}^T + \gamma\mathbf{I}_1$) and $\mathbf{I}$ is the corresponding identity matrix.

The optimization problem of Eq.(10) can be solved by computing k eigenvectors corresponding to the smallest eigenvalues of the following eigen-decomposition problem:

$$\mathbf{W}\mathbf{L} = \mathbf{M}\mathbf{L}\mathbf{D} \tag{11}$$

where $\mathbf{L}$, $\mathbf{D}$ are eigenvectors and eigenvalues respectively, k is the dimension of sample features after dimensionality reduction, then $\mathbf{L}$, $\mathbf{L}_A$ and $\mathbf{L}_B$ can be calculated and the distances between any two pictures can be obtained. The complete algorithm is summarized in Algorithm 1.

Algorithm 1: The Solution of Proposed Method

Input: Cross-view training set $\mathbf{X}$, $\mathbf{Z}$, weight parameters λ and γ.
Output: Projection matrices $\mathbf{L}_A$ and $\mathbf{L}_B$

```
λ = −0.1        → Initialization of λ
γ = 0.001       → Initialization of γ
compute all w^{i,j}_{A,B} by (4)
formulate loss function by (2) and (3)
simplified loss function by (6), (7), (8) and (9)
compute L, L_A,L_B by (10) and (11)
return L_A, L_B
```

3. EXPERIMENTS AND DISCUSSIONS

3.1. Feature Representation

We utilize the Local Maximal Occurrence(LOMO) feature proposed in [13] for feature representation. The LOMO feature has shown impressive robustness against viewpoint changes and illumination variations by concatenating the maximal pattern of joint HSV histogram and SILTP descriptor. Considering that the dimensionality of LOMO is extremely high, PCA is used for dimensionality reduction. The dimension of LOMO feature is reduced to $n - 1$, where n is the number of training sample pairs for each dataset.

3.2. Datasets and Evaluation Protocol

We evaluate the proposed method on three challenging person re-identification datasets, VIPeR [14], PRID 450S [11] and CUHK01 [5]. All datasets are randomly divided into two parts, half for training and the remaining for testing. Experimental results are reported in the form of the average Cumulated Matching scores for 10 trials. Empirically we find that, $\lambda = -0.1, \gamma = 0.001$ as regularizers can be commonly applied to improve the results for the three datasets.

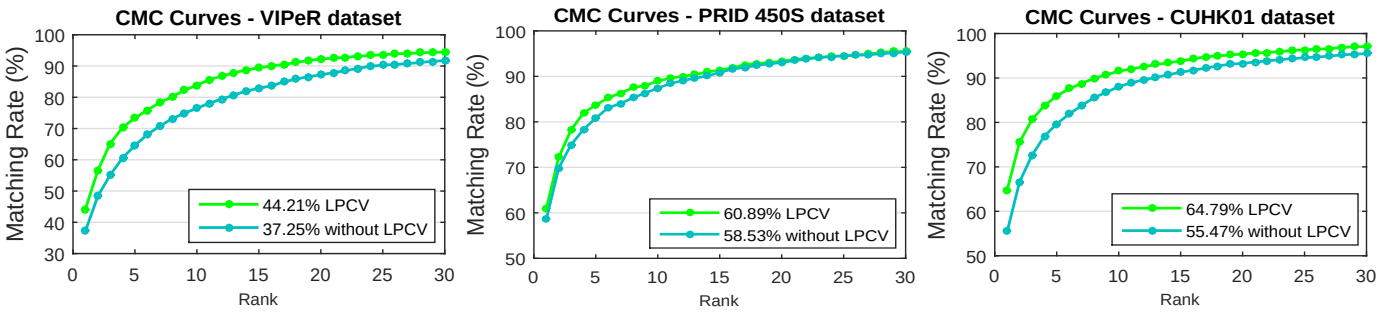
VIPeR dataset [14] contains 632 pairs of person images, captured by a pair of cameras in an outdoor environment. Images in VIPeR contain large variations in backgrounds, illuminations and viewpoints. All images are normalized to 128×48 pixels.

PRID 450S dataset [11] includes 450 single-shot pedestrian image pairs captured from two disjoint camera views. It is also a challenging person re-identification dataset due to the background interference, partial occlusion and viewpoint changes. In our experiments, all images are normalized to 128×64 pixels.

CUHK01 dataset [5] is captured in a campus environment with two camera views. It contains 971 individuals and each of them has two images in every camera view. Taking the evaluation method in [5], we normalize all images to 160×60 pixels.

3.3. Evaluation of the proposed LPCV

To validate the effectiveness of the proposed LPCV, we compare the matching results of our method with or without LPCV on VIPeR, PRID 450S and CUHK01. All parameters are the same besides the number of the learned projection matrices, the results are demonstrated in Fig.2. As illustrated in the figures, the LPCV significantly boosts the performance on the three datasets, for pictures under different views undergo significant variations on viewpoints, poses and illuminations. Learning an identical projection matrix for all views may lose some information. It also validates the assumption that images under different views have different characteristics, by learning projections from corresponding views, we can make better use of this discriminative information.



(a) Evaluate on VIPeR (b) Evaluate on PRID 450S (c) Evaluate on CUHK 01

Fig. 2: Evaluation of the proposed LPCV. Rank-1 matching rate is marked before the name of each method.

Table 1: Comparisons with the state-of-the-arts on VIPeP (P=316).

Method	rank=1	rank=10	rank=20
Ours	**44.21**	83.89	92.28
LSSCDL [15]	42.66	84.27	91.93
DNS [16]	42.28	82.94	92.06
Semantic [17]	41.60	**86.20**	**95.10**
IDLA [18]	34.81	75.63	84.49
PolyMap [19]	36.80	83.70	91.70
MLAPG [20]	40.73	82.34	92.37
XQDA [13]	40.00	80.51	91.08
LADF [9]	30.22	78.92	90.44
LF [6]	24.18	67.12	82.00
KISSME [8]	19.60	62.20	77.00

Table 2: Comparisons with the state-of-the-arts on PRID 450S (P=225).

Method	rank=1	rank=10	rank=20
Ours	**60.89**	88.98	93.33
LSSCDL [15]	58.98	89.24	93.64
Semantic [17]	44.90	77.50	86.70
LOMO+XQDA [13]	59.60	**89.60**	**93.91**
SCNCD [21]	26.90	64.20	74.90
LF [6]	24.18	67.12	82.00
KISSME [8]	33.00	71.00	79.00

3.4. Comparison with State-of-the-arts

Several state-of-the-art metric learning methods with the same feature representation are compared, and the state-of-the-art published results on the three datasets are also compared, detailed comparison results at rank 1, 10, 20 are shown in Table 1, Table 2 and Table 3. The variable P in the tables represents the number of the probe set. As shown in these three tables, in the rank 1 matching rate, our method outperforms all the listed state-of-the-art methods on VIPeR dataset, PRID 450S dataset and achieves competitive performance compared with LSSCDL [15] on CUHK01 dataset. As we all know, rank 1 is more important than the latter ranks in person re-identification problem. The achieved results in rank 1 validate the superiotity of our proposed method. Meanwhile, we can see that our matching rate in rank 10 and rank 20 are a little inferior to the state-of-the-arts. The reasonable explanation is that the process of the projection learning may be affected by the loss function.

Table 3: Comparisons with the state-of-the-arts on CUHK01 (P=486).

Method	rank=1	rank=10	rank=20
Ours	64.79	91.54	95.41
LSSCDL [15]	**65.97**	**92.12**	**95.64**
LOMO+XQDA [13]	63.21	90.04	95.16
DNS [16]	64.98	89.92	94.36
KFLDA [22]	54.63	86.87	92.02
Mid-level Filter [5]	34.30	64.96	74.96
IDLA [18]	47.53	80.25	87.45

3.5. Running cost

We conduct the proposed method with Matlab implementation on a desktop PC with 3.2GHz CPU and 16G RAM, and report the running time of each stage averaged over 10 random trials on the VIPeR dataset. The total running time (including training and test) is 37.8 seconds for each trail. Therefore, it indicates that our model is efficient.

4. CONCLUSIONS

In this paper, we propose a method to learn projections from corresponding views (LPCV) and apply it to person re-identification problem. LPCV makes better use of discriminative information of each camera because it learns different projection matrices for different camera views. Experiments on three challenging person re-identification datasets VIPeR, PRID 450S and CUHK01 demonstrate the superiority of the proposed method over the state-of-the-art methods. In addition, our method has low time consumption, which is applicable for the real-world surveillance systems. Future work focus on exploring a better loss function and applying it to more real-world applications.

5. REFERENCES

[1] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara, "People reidentification in surveillance and forensics: A survey," *in Acm Computing Surveys*, pp. 1–37, 2013.

[2] Shaogang Gong, Marco Cristani, Change Loy Chen, and Timothy M Hospedales, "The re-identification challenge," *in Person Re-Identification*, pp. 1–20, 2014.

[3] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," *in Proceedings of CVPR*, pp. 3586–3593, 2013.

[4] Hong Liu, Liqian Ma, and Can Wang, "Body-structure based feature representation for person re-identification," *in Proceedings of ICASSP*, pp. 1389–1393, 2015.

[5] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," *in Proceedings of CVPR*, pp. 144–151, 2014.

[6] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," *in Proceedings of CVPR*, pp. 3318–3325, 2013.

[7] Alexis Mignon and Frdric Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," *in Proceedings of CVPR*, pp. 2666–2672, 2012.

[8] Martin Köstinger, Martin Hirzer, Paul Wohlhart, and Peter M. Roth, "Large scale metric learning from equivalence constraints," *in Proceedings of CVPR*, pp. 2288–2295, 2012.

[9] Zhen Li, Shiyu Chang, Feng Liang, Thomas S. Huang, Liangliang Cao, and John R. Smith, "Learning locally-adaptive decision functions for person verification," *in Proceedings of CVPR*, pp. 3610–3617, 2013.

[10] Wei Shi Zheng, Shaogang Gong, and Tao Xiang, "Person re-identification by probabilistic relative distance comparison," *in Proceedings of CVPR*, pp. 649–656, 2011.

[11] Peter M Roth, Martin Hirzer, Martin Köstinger, Csaba Beleznai, and Horst Bischof, *Mahalanobis distance learning for person re-identification*, Springer, 2014.

[12] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," *in Proceedings of ICCV*, pp. 498–505, 2009.

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

[14] Doug Gray, Shane Brennan, and Hai Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," *in Proceedings of PETS*, 2007.

[15] Ying Zhang, Baohua Li, Huchuan Lu, Atshushi Irie, and Xiang Ruan, "Sample-specific svm learning for person re-identification," *in Proceedings of CVPR*, pp. 1278–1287, 2016.

[16] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a discriminative null space for person re-identification," *in Proceedings of CVPR*, pp. 1239–1248, 2016.

[17] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang, "Transferring a semantic representation for person re-identification and search," *in Proceedings of CVPR*, pp. 4184–4193, 2015.

[18] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," *in Proceedings of CVPR*, pp. 3908–3916, 2015.

[19] Dapeng Chen, Zejian Yuan, Gang Hua, Nanning Zheng, and Jingdong Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," *in Proceedings of CVPR*, pp. 1565–1573, 2015.

[20] Shengcai Liao and Stan Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," *in Proceedings of ICCV*, pp. 3685–3693, 2015.

[21] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li, "Salient color names for person re-identification," *in Proceedings of ECCV*, pp. 536–551, 2014.

[22] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier, "Person re-identification using kernel-based metric learning methods," *in Proceedings of ECCV*, pp. 1–16, 2014.