Metric Embedded Discriminative Vocabulary Learning for High-Level Person Representation

Yang Yang, Zhen Lei, Shifeng Zhang, Hailin Shi, Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{yang.yang, zlei, shifeng.zhang, hailin.shi, szli}@nlpr.ia.ac.cn

Abstract

A variety of encoding methods for bag of word (BoW) model have been proposed to encode the local features in image classification. However, most of them are unsupervised and just employ k-means to form the visual vocabulary, thus reducing the discriminative power of the features. In this paper, we propose a metric embedded discriminative vocabulary learning for high-level person representation with application to person re-identification. A new and effective term is introduced which aims at making the same persons closer while different ones farther in the metric space. With the learned vocabulary, we utilize a linear coding method to encode the image-level features (or holistic image features) for extracting high-level person representation. Different from traditional unsupervised approaches, our method can explore the relationship (same or not) among the persons. Since there is an analytic solution to the linear coding, it is easy to obtain the final high-level features. The experimental results on person re-identification demonstrate the effectiveness of our proposed algorithm.

1 Introduction

Person re-identification (Re-ID) has recently attracted a lot of attentions due to its many critical applications such as long-term multicamera tracking (Bi Song and Roy-Chowdhury 2010), forensic search (Roberto Vezzani and Cucchiara 2013) and crowd movements analysis in public places (Martin Hirzer and Bischof 2012). The task of person Re-ID is to match persons from several disjoint cameras. To address it, a commonly used framework is (1) appearance based person representation and (2) metric learning for matching them. Owing to large viewpoint changes, illumination, different poses, background clutter and occlusions, there is often large intra-class appearance variations, which makes the descriptive representations of the same person unstable. To that end, metric learning methods are used to reduce the intra-class variations in feature space. Such a learned metric is able to describe the transitions among different cameras, and thus suitable for real world scenarios.

However, because most existing methods of feature extraction are based on descriptive methods such as features extracted in Kiss metric (KISSME) (Kostinger et al. 2012), symmetry-driven accumulation of local features (SDALF) (M. Farenzena et al. 2010), Comb (Kviatkovsky, Adam, and Rivlin 2013a), salient color names based color descriptor (SCNCD) (Yang et al. 2014b), mid-level filters (MLF) (Rui Zhao and Wang 2014) and fusion of color models (FCMs) (Yang et al. 2014a), the extracted features are of less discriminative power. It may influence the final matching rate despite of the application of metric learning methods in the stage of matching persons.

Recently, various coding methods, which are based on a vocabulary learned by k-means, are used to encode local features (e.g. SIFT descriptors) and then a pooling method (e.g., max pooling or average pooling) is utilized to obtain a holistic image feature with statistical meanings. Simple though they are, surprising good results have been reported in classification tasks. Inspired by them, we can employ a coding method to encode the image-level features to learn high-level features. To make the coding coefficient be more discriminative, we can learn a 'good' (e.g., with discriminative power) vocabulary based on the labeled training images instead of simply using k-means.

In this paper, we propose a novel method named metric embedded discriminative vocabulary learning (MED_VL) to learn a vocabulary with discriminative power. Because only pairwise relationships ('same' or 'different') of training images are obtained for person re-identification, we incorporate the equivalence constraints into the objective function of MED_VL to construct a discriminative vocabulary. It can be considered as a supervised learning approach which aims to make the same persons closer while different ones farther in the metric space. Owing to the merits of metric learning method which learns the transitions among cameras, we can measure the similarity of two persons in different cameras in a better manner. With the learned vocabulary, a linear coding method is employed to encode each image-level feature and then the final high-level person representation is then obtained. Based on the same metric learning approach in the stage of matching persons, the final high-level feature which owns semantic information performs better than the original image-level feature.

The main contributions of the paper are two-fold: (1) We employ a linear coding method for feature coding. To our best knowledge, this is the first work to exploit coding methods to learn high-level features from the image-level features

for person re-identification. (2) We propose a novel method - MED_VL to learn the vocabulary for linear coding. It is more discriminative than the one learned by k-means.

The remainder of the paper is organized as follows: Section 2 gives a brief review of related works on coding methods as well as the metric leaning method used in this paper; In section 3, we describe in details the proposed method including vocabulary learning and linear coding; An evaluation of our method on the publicly available person re-identification datasets in section 4, and finally, section 5 makes a conclusion of the paper.

2 Related Work

In this section, we first review the commonly utilized coding methods and then make a brief introduction to the used metric learning approach. Let $X = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_n] \in \mathcal{R}^{d\times n}$ be a set of d-dimensional image-level features, where $\vec{x}_i \in \mathcal{R}^d, i = 1, 2, ..., n$ denotes the feature of the i-th image. Given a vocabulary (or a set of basis vectors) $B = [\vec{b}_1, \vec{b}_2, ..., \vec{b}_k] \in \mathcal{R}^{d\times k}$, different coding methods can be applied to convert each d-dimensional original feature into a k-dimensional high-level one.

2.1 Different Coding Methods

Soft-assignment Coding (SAC) (Jan C. van Gemert and Smeulders 2008): For an image-level feature $\vec{x}_i$, there are k nonzero coding coefficients. The j-th coding coefficient is computed by

$$\vec{s}_{ij} = \frac{\exp(-\gamma\|\vec{x}_i - \vec{b}_j\|_2^2)}{\sum_{l=1}^{k}\exp(-\gamma\|\vec{x}_i - \vec{b}_l\|_2^2)}. \quad (1)$$

where γ is a smoothing factor which controls the softness of the assignment. Each coding coefficient denotes the degree of membership of $\vec{x}_i$ to the corresponding basis vector in B.

Locality-constrained Coding (LLC) (Wang et al. 2010): Other than soft-assignment coding which employs all k basis vectors to encode the features, LLC incorporates the locality constraint which leads to smaller coefficients for those basis vectors farther away from $\vec{x}_i$ in Euclidean space. The LLC code $\vec{s}_i$ is computed by the following criteria:

$$\vec{s}_i = \underset{1^T\vec{s}_i=1}{\operatorname{argmin}} \|\vec{x}_i - B\vec{s}_i\|_2^2 + \lambda\|\vec{d}_i \odot \vec{s}_i\|_2^2, \quad (2)$$

where $\vec{d}_i = exp(dist(\vec{x}_i, B)/\delta)$, $dist(\vec{x}_i, B) = [dist(\vec{x}_i, \vec{b}_1), dist(\vec{x}_i, \vec{b}_2), ..., dist(\vec{x}_i, \vec{b}_k)]^T$, $dist(\vec{x}_i, \vec{b}_j)$ means the Eulidean distance between $\vec{x}_i$ and $\vec{b}_j$ and $\odot$ denotes the element-wise multiplication. δ is employed for adjusting the weight decay speed for the locality adaptor. Additionally, an approximated LLC is also presented for fast encoding.

Salient Coding (SC) (Huang et al. 2011): Based on the 'saliency', which means that the nearest code is much closer to the input feature than other codes, SC defines a new saliency degree:

$$s_{ij} = 1 - \frac{\|\vec{x}_i - \vec{b}_j\|_2^2}{\frac{1}{k_0-1}\sum_{m\neq j}^{k_0}\|\vec{x}_i - \vec{b}_m\|_2^2} \quad (3)$$

where $k_0 < k$ denotes the number of basis vectors used for coding every time. It is efficient and easy to implement.

Although a coding coefficient learned by a coding method has been successfully used as an alternative to its original feature in classification tasks, most of them only focus on how to reflect the fundamental characteristic of coding. For example, in (Wang et al. 2010), it considers that locality is more essential than sparsity. Then, LLC incorporates the locality constraint into the objective function. Another classical example is that in (Huang et al. 2011), it believes saliency is the fundamental characteristic of coding and a novel salient coding algorithm is proposed to stably extract saliency representation. However, both of them simply employ k-means to construct the vocabulary which will reduce the discriminative power of the final features (or coding coefficients). Therefore, there is a need to learn a 'good' vocabulary based on the training images.

2.2 The Metric Learning Method - KISSME

In (Kostinger et al. 2012), a simple yet effective method - KISSME is proposed to learn a global Mahalanobis distance metric defined in Eq. 4 from equivalence constraints.

$$d_M(\vec{x}, \vec{y}) = (\vec{x} - \vec{y})^T M(\vec{x} - \vec{y}). \quad (4)$$

where $\vec{x}$ and $\vec{y}$ are features of a pair of images. In consideration of the fact that there is a bijection between the set of Mahalanobis metric and that of multivariate Gaussian distribution, KISSME directly computes M by Eq. 5 with the help of a log like ratio.

$$M = \Sigma_{\mathcal{D}}^{-1} - \Sigma_{\mathcal{S}}^{-1}. \quad (5)$$

where $\Sigma_{\mathcal{S}}$ and $\Sigma_{\mathcal{D}}$ denotes the covariance matrixes of similar pairs and dissimilar pairs, respectively. To make M be a positive semi-definite matrix, the authors of (Kostinger et al. 2012) further re-project it onto the cone of positive semi-definite matrixes, i.e., clipping the spectrum of M by eigen-analysis.

In this paper, we employ KISSME to learn a good metric. However, we find that adding the positive semi-definite matrix constraint on M does not bring better results but consumes additional computation time in experiments. Therefore, we do not take the constraint into consideration, i.e., we just employ Eq. 5 to compute M. Additionally, we normalize each feature with the l_2 norm to slightly improve the performance of KISSME.

3 The Proposed Method

In this section, we first present our metric embedded discriminative vocabulary learning (MED_VL) from equivalence constraints. By taking into consideration the relationship of each training pair of learned coding coefficient s_i, s_j, it aims at encouraging similar pairs (or pairs of same persons) to be closer than dissimilar pairs (or pairs of different persons) in the metric space. Then, with the learned vocabulary, we show how to efficiently learn the final high-level features from original image-level features based on a linear coding method.

3.1 Metric Embedded Discriminative Vocabulary Learning

Considering that person re-identification problem is lacking class labels, we incorporate the equivalence constraints into our objective function. The similarity between a pair is measured in the metric space. Thus, a metric embedded discriminative vocabulary is learned with pairwise constraints.

Suppose we have a coding coefficient matrix $S = [\vec{s}_1, \vec{s}_2, ..., \vec{s}_n] \in \mathcal{R}^{k\times n}$ corresponding to the original data matrix $X = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_n] \in \mathcal{R}^{d\times n}$. Each column of S denotes a new representation of each data (i.e., corresponding column of X) in the new space. With the training data, one may hope that similar pairs are closer than dissimilar pairs. Due to the fact that metric learning method is able to learn the transitions among cameras, we measure the distance between a pair of data from different cameras in the metric space. Then, we minimize the following term:

$$\frac{1}{2}\sum_{i,j=1}^{n}(\vec{s}_i - \vec{s}_j)^T M(\vec{s}_i - \vec{s}_j)W_{ij} = Tr(S^T MSL) \quad (6)$$

where $M \in \mathcal{R}^{k\times k}$ is the matrix parameterizing the metric space, $L = D - W$ is the Laplacian matrix, $D = diag(d_1, d_2, ..., d_n)$ is a diagonal matrix with $d_i = \sum_{j=1}^{n} W_{ij}$ and W is defined by Eq. 7.

$$W_{ij} = \begin{cases} 1/n_{\mathcal{S}} & , \quad if(\vec{x}_i, \vec{x}_j) \in \mathcal{S} \\ -1/n_{\mathcal{D}} & , \quad if(\vec{x}_i, \vec{x}_j) \in \mathcal{D} \end{cases} \quad (7)$$

where $n_{\mathcal{S}}$ and $n_{\mathcal{D}}$ denotes the number of similar and dissimilar pairs, respectively. With Eq. 6 being as a regularizer, our objective function is defined as

$$\begin{aligned} &\min_{B,S} \|X - BS\|_F^2 + \alpha Tr(S^T MSL) + \beta\|S\|_F^2 \\ &s.t. \ \ \|\vec{b_i}\|_2^2 \le C, i = 1, 2, ..., k. \end{aligned} \quad (8)$$

where the parameters α and β are used to control the contributions of corresponding terms and M is fixed and directly learned in Eq. 5 based on the initial coding coefficients of training data. In Eq. 8, the first term denotes the reconstruction error. The second term is employed to ensure that similar pairs are closer than dissimilar pairs in the metric space. The last term is to avoid overfitting.

Although the objective function in Eq. 8 is convex for B only or S only, it is not convex in both variables together. We then solve this problem by optimizing B and S iteratively in the following.

Initialization of B, S, M We should initialize B, S and M before learning the discriminative vocabulary B. To be specific, B is firstly initialized by k-means. Then, by solving the linear coding defined in Eq. 9, we can obtain the initial value of S via Eq. 10 where $I \in \mathcal{R}^{k\times k}$ is the identity matrix.

$$\min_{S} \|X - BS\|_F^2 + \beta\|S\|_F^2 \quad (9)$$

$$S = (B^T B + \beta I)^{-1}(B^T X) \quad (10)$$

If we initialize M directly using Eq. 5, there exists a singular value if k is larger than n. To address this problem, we first apply PCA to learn a projecting matrix $P \in \mathcal{R}^{k\times m}, m < n$. Then, we have $(P^T S)^T M_0 (P^T S) = S^T (P M_0 P^T) S$ where M_0 is learned based on the PCA-reduced training data via Eq. 5. Thus, we are able to initialize M by PM_0P^T. We then optimize the S and B iteratively with the fixed M.

Learning Linear Codes S When B is fixed, the problem Eq. 8 becomes

$$\min_{S} \|X - BS\|_F^2 + \alpha Tr(S^T MSL) + \beta\|S\|_F^2. \quad (11)$$

To solve Eq. 11, we optimize each vector $\vec{s}_i$ alternatively, while holding other vectors $\vec{s}_j (j \neq i)$ constant. Optimization of Eq. 11 is equivalent to

$$\min_{\vec{s}_i} \mathcal{F}(\vec{s}_i) + \beta\|\vec{s}_i\|_2^2. \quad (12)$$

where $\mathcal{F}(\vec{s}_i) = \|\vec{x}_i - B\vec{s}_i\|_2^2 + \alpha(2\vec{s}_i^T(MS\vec{L}_i) - \vec{s}_i^T M\vec{s}_i L_{ii}) + \beta\|\vec{s}_i\|_2^2$ with $\vec{L}_i = \sum_{l=1}^{n} L_{li}$.

Therefore, an analytic solution can be obtained when we have $\frac{\partial}{\partial \vec{s}_i}\mathcal{F}(\vec{s}_i) = \vec{0}$:

$$\vec{s}_i^{new} = (B^T B - \alpha L_{ii} M + \beta I)^{-1}(B^T \vec{x}_i - \alpha MS\vec{L}_i). \quad (13)$$

Learning Dictionary B When the coefficient matrix S is given, we employ the Lagrange dual in (Honglak Lee and Ng 2007) to optimize the following least squares problem with quadratic constraints in Eq. 14. There is an analytical solution of B: $B^{new} = XS^T(SS^T + \Lambda)^{-1}$ where Λ is a diagonal matrix of optimal dual variables.

$$\min_{B,S} \|X - BS\|_F^2 \ \ s.t. \ \ \|\vec{b_i}\|_2^2 \le C, i = 1, 2, ..., k \quad (14)$$

Algorithm 1 Metric Embedded Discriminative Vocabulary Learning

Input: Data matrix X, Laplacian matrix L, parameters α and β and iteration number T.
Output: B.

1: Initialize B with k-means, S via Eq. 10 and M by PM_0P^T.
2: **for** t=1,2,...,T **do**
3: Update the coding coefficients S using Eq. 11;
4: Update the vocabulary B using Eq. 14.
5: **end for**

3.2 Linear Coding

Once we have obtained the discriminative vocabulary B (described in Algorithm 1), we employ the linear coding defined by Eq. 9 to encode all the image-level features and then the final high-level features (high-level person representation) are obtained. In comparison with the original features, the final obtained features, which own latent semantic information, are more compact and discriminative. We will compare their performances in the following experiments.

Figure 1: Some examples from the datasets: VIPeR (left) and PRID 450S (right).

4 Experimental Results

In this section, we evaluate the proposed algorithm on two publicly available datasets: VIPeR dataset (Gray, Brennan, and Tao 2007), PRID 450S dataset (Roth et al. 2014) for person re-identification problems. Some examples from these two datasets are shown in Fig. 1. In both datasets, each person has two images obtained from two disjoint cameras. Both of them are challenging datasets. Specifically, VIPeR dataset mainly suffers from arbitrary viewpoints and illumination variations between two camera views. There are 632 image pairs captured in outdoor environments. Due to different viewpoint changes, background interference, partial occlusion and illumination variations, PRID 450S dataset is also challenging. It contains 450 image pairs from two disjoint camera views.

4.1 Settings

In all experiments, half image pairs are randomly selected for training and the remaining are employed for test. During test, images from camera A are considered as probe and those from camera B as gallery. Then, we switch them. We regard the average results as one run. The final average results over 100 runs are reported in form of Cumulated Matching Characteristic (CMC) curve (Wang et al. 2007).

Features In experiments, we employ the image-level features provided by the authors of (Yang et al. 2014b): Image-only representation and foreground representation based on color names distributions and color histograms are extracted and fused in four different color spaces including original RGB, *rgb*, $l_1l_2l_3$, and HSV. As is suggested in (Yang et al. 2014b), the dimensions of features of both datasets are reduced to 70 by PCA.

Parameter Settings In our evaluations, we set α and β in Eq 8 to 0.5 and 0.2, respectively. The number of basis vectors in B is set to 120 and the iteration number T to 4. Before using KISSME, we employ PCA to reduce the 120-dimensional high-level features to 70 for both datasets. When SAC is compared, we set γ to 0.05. When we compare the coding methods - LLC and SC, we report two kinds of results based on (1) 5 nearest basis vectors and (2) all 120 basis vectors. We name them LLC(5), LLC(120), SC(5) and SC(120).

4.2 Comparison with Different Coding Methods

To validate whether the high-level features are better than original features, we employ different coding methods including SAC, LLC and SC to encode the input features. For these methods, we use k-means to construct the vocabulary. For our method, which employs MED_VL to construct a discriminative vocabulary, we name it MED_VL for simplicity in experiments. Then, we report the results on VIPeR and PRID 450S datasets in Tables 1 and 2, respectively. 'Original' denotes the input original features. For the sake of a fair comparison, all of the reported results are based on the same metric learning method - KISSME.

Rank	**1**	**5**	**10**	**20**
Original	38.9%	69.3%	81.0%	90.1%
SAC	39.3%	69.5%	81.1%	90.1%
LLC(5)	12.8%	32.5%	45.6%	60.9%
LLC(120)	39.3%	69.6%	81.3%	90.2%
SC(5)	11.6%	31.3%	45.3%	61.5%
SC(120)	39.5%	70.0%	81.6%	90.4%
MED_VL	**41.1%**	**71.7%**	**83.2%**	**91.7%**

Table 1: Comparison between the high-level features and original features on VIPeR dataset. Different coding methods including SAC, LLC and SC are analyzed. k-means is employed to construct the vocabulary except MED_VL which learn a discriminative vocabulary. Best in bold.

Rank	**1**	**5**	**10**	**20**
Original	42.6%	70.0%	79.8%	88.0%
SAC	43.4%	70.4%	80.5%	89.0%
LLC(5)	9.4%	27.8%	40.6%	56.9%
LLC(120)	43.3%	70.6%	80.5%	89.1%
SC(5)	8.5%	26.5%	39.7%	56.0%
SC(120)	44.0%	71.3%	81.4%	89.8%
MED_VL	**45.9%**	**73.0%**	**82.9%**	**91.1%**

Table 2: Comparison between the high-level features and original features on PRID 450S dataset. Different coding methods including SAC, LLC and SC are analyzed. k-means is employed to construct the vocabulary except MED_VL which learn a discriminative vocabulary. Best in bold.

From Tables 1 and 2, we can observe that all unsupervised coding methods including SAC, LLC(120) and SC(120) perform (at least slightly) better than the original features. It demonstrates the feasibility of coding methods for leaning high-level features. When we compare the performances among different coding methods all of which employ k-means to obtain the vocabulary, we find that LLC(120) performs similar as SAC while SC(120) is slightly better. The best results are achieved by our method MED_VL. By comparing MED_VL with the original feature, we can see that there are 2.2% and 3.3% increases at Rank 1 on VIPeR and PRID 450S datasets, respectively. This observation demonstrates that our proposed approach is able to learn more discriminative features than original ones. Additionally, it should be noted that on both datasets, LLC(5) and SC(5) performs poorly, it reflects that when we encode the image-level features, locality constraint harms the results. This is caused by the fact that if there are only several nonzero val-

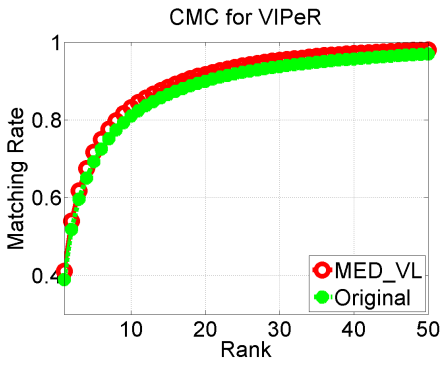

Figure 2: Comparison of MED_VL with original features on the VIPeR.

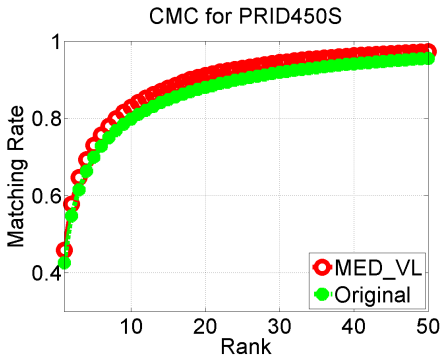

Figure 3: Comparison of MED_VL with original features on the PRID 450S.

ues in the coding coefficient, the obtained feature can not retain its original data information. In addition, Figs. 2 and 3 also compare them overally at Ranks 1-50 on VIPeR and PRID 450S datasets, respectively. It is obvious that high-level features obtained by MED_VL perform better than the original features at all Ranks.

4.3 Comparison with Euclidean Space Embedded

In Eq. 8, we learn the matrix M by KISSME to compute the training pairs in a metric space. However, we can also directly compute them in Euclidean space, i.e., $M = I$, where I is the identity matrix. We name them 'Euclidean' and 'Metric', respectively. Table 3 shows the results on VIPeR dataset while Table 4 shows the results on PRID 450S dataset. Ranks 1, 5, 10 and 20 are reported. On both datasets, we can find that the results based on Metric space embedded are better than those based on Euclidean space embedded.

Rank	**1**	**5**	**10**	**20**
Euclidean	39.5%	70.0%	81.6%	90.4%
Metric	**41.1%**	**71.7%**	**83.2%**	**91.7%**

Table 3: Comparison with Euclidean space embedded in Eq. 8 on VIPeR dataset. Best in bold.

Rank	**1**	**5**	**10**	**20**
Euclidean	45.1%	72.6%	82.7%	90.8%
Metric	**45.9%**	**73.0%**	**82.9%**	**91.1%**

Table 4: Comparison with Euclidean space embedded in Eq. 8 on PRID 450S dataset. Best in bold.

Rank	**1**	**10**	**20**	**Reference**
HCe	32.2%	66.9%	-	ICCV2015
CVPDL	34.0%	77.5%	88.6%	IJCAI2015
LOMO	40.0%	80.5%	91.1%	CVPR 2015
Final*	37.8%	81.2%	90.4%	ECCV 2014
MtMCML	28.8%	75.8%	88.5%	TIP 2014
SSCDL	25.6%	68.1%	83.6%	CVPR 2014
Mid-level	29.1%	67.1%	80.0%	CVPR 2014
SalMatch	30.2%	66.0%	79.2%	ICCV 2013
ColorInv	24.2%	57.1%	69.7%	TPAMI2013
LF	24.2%	67.1%	82.0%	CVPR 2013
KISSME	19.6%	62.2%	77.0%	CVPR 2012
MED_VL	**41.1%**	**83.2%**	**91.7%**	Proposed

Table 5: Comparison with the state-of-the-art methods on VIPeR dataset. Best in bold. *Copied directly from the corresponding paper.

4.4 Comparison with State-of-the-art Results

In this subsection, we compare the performance of the proposed method to the state-of-the-art results on VIPeR and PRID 450S datasets at Ranks 1, 10 and 20. On VIPeR datast, the compared methods include HS+CN+eSDC (HCe) (Liang Zhengy and Tian 2015), CVPDL (Sheng Li and Fu 2015), LOMO (Shengcai Liao and Li 2015), Final (Yang et al. 2014b), MtMCML (Lianyang Ma and Tao 2014), SSCDL (Xiao Liu and Bu 2014), Mid-level (Rui Zhao and Wang 2014), SalMatch (Rui Zhao and Wang 2013), ColorInv (Kviatkovsky, Adam, and Rivlin 2013b), LF (Sateesh Pedagadi and Boghossian 2013) and KISSME (Kostinger et al. 2012). Among the previous approaches, LOMO achieves the best results at all Ranks. Our method performs better than LOMO (1.1% higher at Rank 1) and achieve a new state-of-the-art result 41.1% at Rank 1. On PRID 450S dataset, the compared methods are Final (Yang et al. 2014b), KISSME (Kostinger et al. 2012) and EIML (Hirzer, Roth, and Bischof 2012). Our approach also performs the best (4.3% higher than Final(ImgF) at Rank 1) at all Ranks and achieves a new state-of-the-art result 45.9% at Rank 1.

4.5 Evaluation of Vocabulary Size

In this subsection, we evaluate the effect of the vocabulary size on the final results. The numbers are selected as 80, 120, 160, 200 and 300. For LLC and SC, we use all the basis vectors to encode the features. Tables 7 and 8 show the results on VIPeR and PRID 450S datasets, respectively. It can be seen that all of them remain relatively stable from 80 to 300. Finally, in Figs. 4 and 5, we also give some examples of querying results based on MED_VL on VIPeR and PIRD 450S dataset, respectively. Given a query, top 10

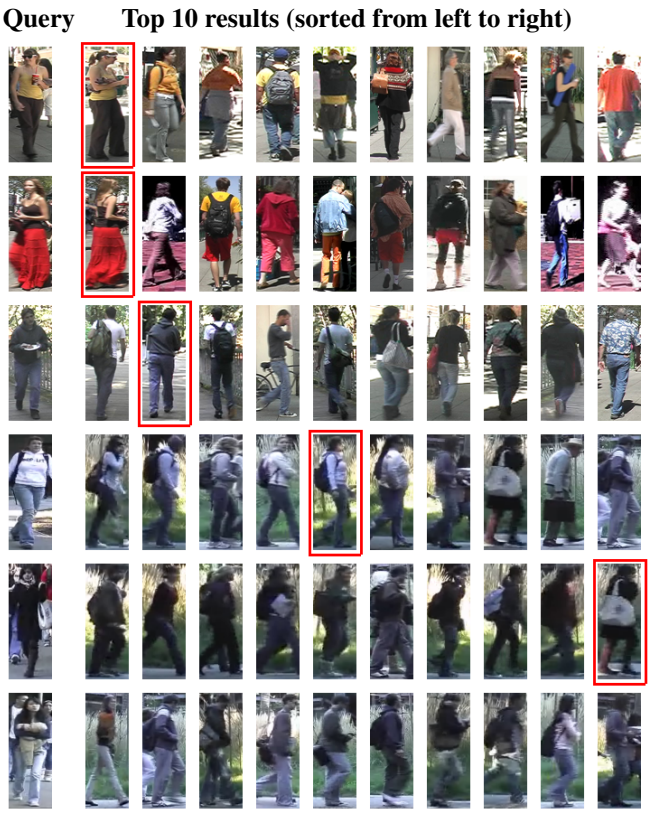

Figure 4: Examples of querying results based on MED_VL on VIPeR dataset.

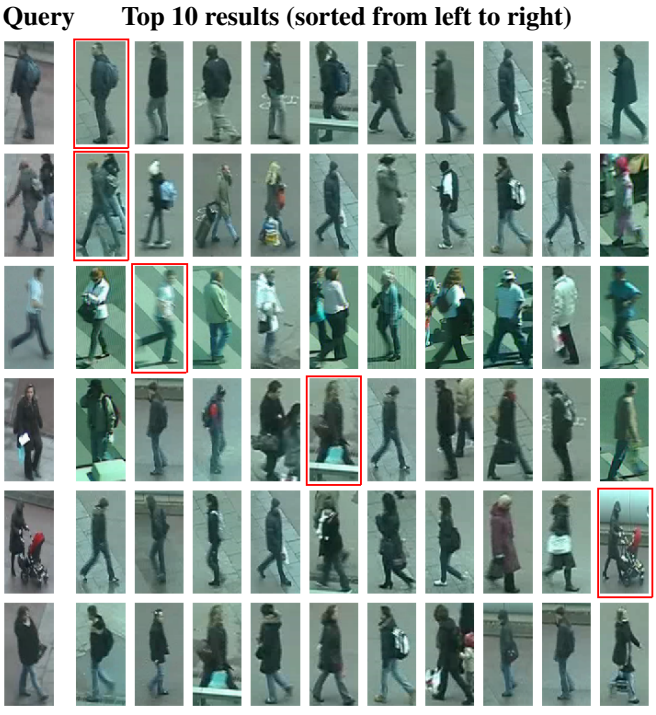

Figure 5: Examples of querying results based on MED_VL on PIRD 450S dataset.

Rank	1	10	20	Reference
Final*	41.6%	79.4%	87.8%	ECCV2014
KISSME	33.0%	71.0%	79.0%	CVPR2012
EIML	35%	68%	77%	AVSS2012
MED_VL	**45.9%**	**82.9%**	**91.1%**	Proposed

Table 6: Comparison with the state-of-the-art methods on PRID 450S dataset. Best in bold. *Copied directly from the corresponding paper.

No.	80	120	160	200	300
SAC	38.9%	39.3%	39.1%	38.8%	38.6%
LLC	38.7%	39.3%	39.0%	39.1%	38.6%
SC	39.0%	39.5%	39.2%	39.1%	39.3%
MED_VL	41.0%	41.1%	40.8%	41.0%	40.6%

Table 7: Evaluation of vocabulary size on VIPeR dataset. Rank 1 is shown.

No.	80	120	160	200	300
SAC	43.1%	43.4%	43.2%	42.9%	42.9%
LLC	43.4%	43.3%	43.3%	43.0%	43.1%
SC	43.9%	44.0%	40.1%	43.6%	43.8%
MED_VL	45.7%	45.9%	45.9%	45.5%	45.8%

Table 8: Evaluation of vocabulary size on PRID 450S dataset. Rank 1 is shown.

similar images (sorted from left to right) coming from the Gallery are shown. A red rectangular box is used to highlight the correct match. If there is no red rectangular box, it means that the correct match is not among the top 10 results. We can find that if the illumination is not severe or there are not serious background interference, it seems easier to find the 'right' person. However, in the condition that the appearance in one camera is different with that in the other camera, caused by illumination, different viewpoints or background interference, the appearance based method will fail. Actually, it is difficult even for human being to match them.

5 Conclusion

This paper presents a novel vocabulary learning method to construct a discriminative vocabulary based on which the high-level features are obtained by a linear coding method. In view of the fact that only pairwise relationship (similar/dissimilar) can be used for person re-identification problem, we incorporate equivalence constraints into our objective function which makes similar pairs closer than dissimilar pairs in the metric space. Experimental results on VIPeR and PRID 450S datasets show that our approach of learning high-level features can obtain better results than the direct application of original features and can also achieve superior performances than several classical coding methods. Additionally, we point out that locality constraint in coding methods can not represent the image-level features well and will harm the final re-identification rates.

Acknowledgments This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, #61572501, #61572536, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

Bi Song, Ahmed T. Kamal, C. S. C. D. J. A. F., and Roy-Chowdhury, A. K. 2010. Tracking and activity recognition through consensus in distributed camera networks. *Image Processing, IEEE Transactions on* 19(10):2564–2579.

Gray, D.; Brennan, S.; and Tao, H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*.

Hirzer, M.; Roth, P. M.; and Bischof, H. 2012. Person re-identification by efficient impostor-based metric learning. In *In: Proc. AVSS*.

Honglak Lee, Alexis Battle, R. R., and Ng, A. Y. 2007. Efficient sparse coding algorithms. In *NIPS*.

Huang, Y.; Huang, K.; Yu, Y.; and Tan, T. 2011. Salient coding for image classification. In *CVPR*.

Jan C. van Gemert, Jan-Mark Geusebroek, C. J. V., and Smeulders, A. W. 2008. Kernel codebooks for scene categorization. In *ECCV*.

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

Kviatkovsky, I.; Adam, A.; and Rivlin, E. 2013a. Color invariants for person reidentification. *IEEE Trans. on PAMI* 35(7):1622–1634.

Kviatkovsky, I.; Adam, A.; and Rivlin, E. 2013b. Color invariants for person reidentification. *IEEE Trans. on PAMI* 35(7):1622–1634.

Liang Zhengy, Liyue Sheny, L. T. S. W. J. W., and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Lianyang Ma, X. Y., and Tao, D. 2014. Person re-identification over camera networks using multi-task distance metric learning. *TIP* 23:3656C–3670.

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

Martin Hirzer, Peter M. Roth, M. K., and Bischof, H. 2012. Relaxed pairwise learned metric for person re-identification. In *ECCV*.

Roberto Vezzani, D. B., and Cucchiara, R. 2013. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys(CSUR)* 46(2).

Roth, P. M.; Hirzer, M.; Kostinger, M.; Beleznai, C.; and Bischof, H. 2014. Mahalanobis distance learning for person re-identification. In *Advances in Computer Vision and Pattern Recognition*.

Rui Zhao, W. O., and Wang, X. 2013. Person re-identification by salience matching. In *ICCV*.

Rui Zhao, W. O., and Wang, X. 2014. Learning mid-level filters for person re-identification. In *CVPR*.

Sateesh Pedagadi, S. V., and Boghossian, B. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*.

Sheng Li, M. S., and Fu, Y. 2015. Cross-view projective dictionary learning for person re-identification. In *IJCAI*.

Shengcai Liao, Yang Hu, X. Z., and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

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

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *CVPR*.

Xiao Liu, Mingli Song, D. T. X. Z. C. C., and Bu, J. 2014. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*.

Yang, Y.; Liao, S.; Lei, Z.; Yi, D.; and Li, S. Z. 2014a. Color models and weighted covariance estimation for person re-identification. In *ICPR*.

Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; and Li, S. Z. 2014b. Salient color names for person re-identification. In *ECCV*.