# Metric Learning in Codebook Generation of Bag-of-Words for Person Re-identification

Lu Tian, Ranran Huang and Yu Wang
*Department of Electronic Engineering, Tsinghua University, Beijing, China*

Keywords: Person Re-identification, Bag-of-Words, Metric Learning.

Abstract: Person re-identification is generally divided into two part: the first is how to represent a pedestrian by discriminative visual descriptors and the second is how to compare them by suitable distance metrics. Conventional methods isolate these into two parts, the first part usually unsupervised and the second part supervised. The Bag-of-Words (BoW) model is a widely used image representing descriptor in part one. Its codebook is simply generated by clustering visual features in Euclidean space, however, it is not optimal. In this paper, we propose to use a metric learning techniques of part two in the codebook generation phase of BoW. In particular, the proposed codebook is clustered under Mahalanobis distance which is learned supervised. Then local feature is compared with the codewords in the codebook by the trained Mahalanobis distance metric. Extensive experiments prove that our proposed method is effective. With several low level features extracted on superpixel and fused together, our method outperforms state-of-the-art on person re-identification benchmarks including VIPeR, PRID 450S, and Market-1501.

## 1 INTRODUCTION

Person re-identification (Gong et al., 2014) is an important task in video surveillance systems. The key challenge is the large intra-class appearance variations, usually caused by various human body poses, illuminations, and different camera views. Furthermore, the poor quality of video sequences makes it difficult to develop robust and efficient features.

Generally speaking, person re-identification can be divided into two parts: first how to represent a pedestrian by discriminative visual descriptors and second how to compare them by suitable distance metrics. Bag of words (BoW) model and its variants is one of the most widely used part one image descriptor technology in person re-id systems with significant performance (Lu and Shengjin, 2015). In the traditional BoW approaches, images are divided into patches and local features are first extracted to represent these patches. Then a codebook of visual words is generated by unsupervised clustering. After that, the image is represented by histogram vectors obtained by mapping and quantizing the local features into the visual words in the codebook.

However, it is not optimal to cluster visual words by k-means in Euclidian space, which implicitly assumes that local features of the same person usually have closer Euclidian distance, which does not always stand in practical.

Part two metric learning methods learn suitable distance metrics of image descriptors to distinguish correct and wrong matching pairs. However, conventional methods always isolate part one and part two, the first part usually unsupervised and the second part supervised.

To this end, this paper proposes to borrow some part two metric learning techniques to learn a suitable distance for local features in part one BoW model. In particular, a Mahalanobis distance is trained on local features extracted from pedestrian images. Then codebook of visual words is clustered under this Mahalanobis distance. We formulate the codebook generation task as a distance metric learning problem and propose to use KISSME (Köstinger et al., 2012) to solve it efficiently. When integrated with conventional part two metric learning methods, our proposed method also achieves good performance. The overall framework of our proposed method is shown in Fig 1. Finally, we outperform state-of-the-art result by applying KISSME (Köstinger et al., 2012) metric learning for local features in the BoW model and Null Space (Zhang et al., 2016a) metric learning for image descriptors after the BoW model.

In summary, our contributions are three-fold: 1),



Tian, L., Huang, R. and Wang, Y.
Metric Learning in Codebook Generation of Bag-of-Words for Person Re-identification.
DOI: 10.5220/0007251102980306
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 298-306
ISBN: 978-989-758-351-3

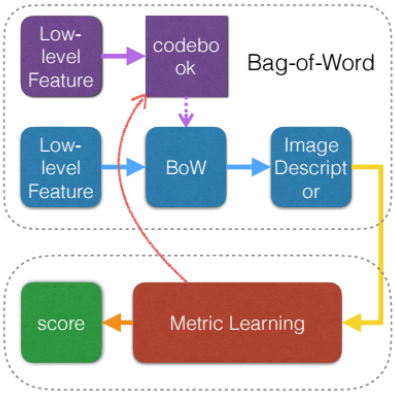


Figure 1: The framework of metric learning in codebook generation of Bag-of-Words.

to the best of our knowledge, we are the first to propose metric learning for BoW low level features; 2), we propose KISSME (Köstinger et al., 2012) to learn a suitable metric for low level features; 3) we integrate the proposed local feature level metric learning method with conventional part two image descriptor level metric learning methods and achieve state-of-the-art results.

The rest of this paper is organized as follows. In Section 2, a brief discussion of several related works on person re-identification is made. In Section 3 we introduce our method. The experimental results are shown and examined in Section 4. Finally, we draw our conclusions in Section 5.

## 2 RELATED WORK

Generally speaking, person re-id include two basic parts: how to represent a pedestrian and how to compare them, and most efforts on person re-id could be roughly divided into these two categories (Zheng et al., 2016b).

The first category focuses on discriminative visual descriptor extraction. Gray and Tao (Gray and Tao, 2008) use RGB, HS, and YCbCr color channels and 21 texture filters on luminance V channel, and partition pedestrian images into horizontal strips. Farenzena et al. (Farenzena et al., 2010) compute a symmetrical axis for each body part to handle viewpoint variations, based on which the weighted color histogram, the maximally stable color regions, and the recurrent high-structured patches are calculated. Zhao et al. (Zhao et al., 2013) propose to extract 32-dim LAB color histogram and 128-dim SIFT descriptor from each 10*10 patch. Das et al. (Das et al., 2014) use HSV histograms on the head, torso and legs. Li et al. (Li et al., 2013) aggregate local color features by hierarchical Gaussianization (Zhou et al., 2009; Chen et al., 2015) to capture spatial information. Pedagadi et al. (Pedagadi et al., 2013) extract color histograms from HSV and YUV spaces and then apply PCA dimension reduction. Liu et al. (Liu et al., 2014) extract HSV histogram, gradient histogram, and the LBP histogram from each patch. Yang et al. (Yang et al., 2014) propose the salient color names based color descriptor (SCNCD) and different color spaces are analyzed. In (Liao et al., 2015), LOMO is proposed to maximize the occurrence of each local pattern among all horizontal sub-windows to tackle viewpoint changes and the Retinex transform and a scale invariant texture operator are applied to handle illumination variations. In (Lu and Shengjin, 2015), Bag-of-Words (BoW) model is proposed to aggregate the 11-dim color names feature (Van de Weijer et al., 2007) from each local patch.

The second category learns suitable distance metrics to distinguish correct and wrong match pairs. Specifically, most metric learning methods focus on Mahalanobis form metrics, which generalizes Euclidean distance using linear scaling and rotation of the feature space, and the distance between two feature vectors $x_i$ and $x_j$ could be written as

$$s(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{M} (x_i - x_j)}, \quad (1)$$

where $\mathbf{M}$ is the positive semi-definite Mahalanobis matrix. Weinberger and Saul (Weinberger and Saul, 2009) propose the large margin nearest neighbor learning (LMNN) which sets up a perimeter for correct match pairs and punishes those wrong match pairs. In (Köstinger et al., 2012), KIEEME is proposed under the assumption that $x_i - x_j$ is a Gaussian distribution with zero mean. Hirzer et al. (Hirzer et al., 2012) obtained a simplified formulation and a promising performance by relaxing the positivity constraint required in Mahalanobis metric learning. Li et al. (Li et al., 2013) propose locally-adaptive decision functions (LADF) combining a global distance metric and a locally adapted threshold rule in person verification. Chen et al. (Chen et al., 2015) add a bilinear similarity in addition to the Mahalanobis distance to model cross-patch similarities. Liao and Li (Liao and Li, 2015) propose weighting the positive and negative samples differently. In (Liao et al., 2015), XQDA is proposed as an extension of Bayesian face and KISSME, in that a discriminant subspace is further learned together with a distance metric. It learns a projection $w$ to the low-dimensional subspace in a similar way as linear discriminant analysis (LDA) (Scholkopft and Mullert, 1999) with

$$\mathcal{J}(w) = \frac{w^T S_b w}{w^T S_w w} \quad (2)$$

maximized, where $S_b$ is the between-class scatter matrix and $S_w$ is the within-class scatter matrix. Zhang et al. (Zhang et al., 2016a) propose Null Space to further employ the null Foley-Sammon transform to learn a discriminative null space with the projection $w$ where the within-class scatter is zero and between-class scatter is positive, thus maximizing $\mathcal{J}(w)$ to positive infinite.

Recently some works based on deep learning are also used to tackle person re-id problem. Filter pairing neural network (FPNN) (Li et al., 2014) is proposed to jointly handle misalignment, photometric and geometric transforms, occlusions and background clutter with the ability of automatically learning features optimal for the re-identification task. Ahmed et al. (Ahmed et al., 2015) present a deep convolutional architecture and propose a method for simultaneously learning features and a corresponding similarity metric for person re-identification. Compared to handcrafted features and metric learning methods, Yi et al. (Yi et al., 2014) proposes a more general way that can learn a similarity metric from image pixels directly by using a "siamese" deep neural network. A scalable distance driven feature learning framework based on the deep neural network is presented in (Ding et al., 2015). Zheng et al. (Zheng et al., 2016c) propose a new siamese network that simultaneously computes identification loss and verification loss, which learns a discriminative embedding and a similarity measurement at the same time. Pose invariant embedding (PIE) is proposed as a pedestrian descriptor in (Zheng et al., 2017), which aims at aligning pedestrians to a standard pose to help re-id accuracy.

# 3 THE APPROACH

## 3.1 Review of Bog-of-Words in Person Re-identification

The BoW model represents an image as a collection of visual words. We briefly review the BoW model in person re-identification in previous approaches (Lu and Shengjin, 2015; Zheng et al., 2015). First, an pedestrian image $i$ is segmented as superpixels by SLIC method (Achanta et al., 2012). Superpixel algorithms cluster pixels into perceptually meaningful atomic regions according to the pixel similarity of color and texture, which capture image redundancy and provide a convenient primitive to compute robust image features. To enhance geometric constraints, the pedestrian image is usually partitioned into horizontal strips with equal width. Then in superpixel $k$ of strip $j$, the low level high-dimensional appearance features are extracted as $\mathbf{f}_{i,j,k} \in \mathcal{R}^d$ and $d$ is the feature vector length. These low level features may contain much noise and redundancy, and are often difficult to use directly. Hence, a codebook $\mathcal{C} = \{\mathbf{c}(l)\}$ of visual words is generated by clustering (usually standard k-means) on these features and each word $\mathbf{c}$ corresponds to a cluster center with $l$ in a finite index set. The mapping, termed as a quantizer, is denoted by: $\mathbf{f} \rightarrow \mathbf{c}(l(\mathbf{f}))$. The function $l(\cdot)$ is called an encoder, and function $\mathbf{c}(\cdot)$ is called a decoder (Gray, 1984). The encoder $l(\mathbf{f})$ maps any $\mathbf{f}$ to the index of its nearest codeword in the codebook $\mathcal{C}$. Here multiple assignment (MA) (Jegou et al., 2008) is employed, where the local feature $\mathbf{f}_{i,j,k}$ is assigned to some of the most similar visual words by measuring the distance between them. Thus the histogram of the visual words representing strip $j$ is obtained by encoding the local features into the codebook, which is denoted as $\mathbf{d}_{i,j} = histogram\{l(\mathbf{f}_{i,j,k}) | k \in strip_j\}$. Each visual word is generally weighted using the TF scheme [2], [3]. We also use pedestrian parsing and background extraction techniques (Luo et al., 2013) and only the superpixels which contain pedestrian parts are considered and counted in our BoW model. The BoW descriptor of image $i$ is the concatenation of $\mathbf{d}_i = [\mathbf{d}_{i,1}, \cdots, \mathbf{d}_{i,j}, \cdots, \mathbf{d}_{i,J}]$. Finally, the distance of two images $i1$ and $i2$ can be directly calculated as the Euclidean distance between $\mathbf{d}_{i1}$ and $\mathbf{d}_{i2}$, that is,

$$s(i1, i2) = \sqrt{(\mathbf{d}_{i1} - \mathbf{d}_{i2})^T \cdot (\mathbf{d}_{i1} - \mathbf{d}_{i2})}. \quad (3)$$

Or conventional part two metric learning methods can be applied to improve re-id performance by supervised labels. Most of them focus on Mahalanobis based metrics, which generalizes Euclidean distance using linear scalings and rotations of the feature space and can be written as

$$s(i1, i2) = \sqrt{(\mathbf{d}_{i1} - \mathbf{d}_{i2})^T \mathbf{M} (\mathbf{d}_{i1} - \mathbf{d}_{i2})}, \quad (4)$$

where $\mathbf{M}$ is the positive semi-definite Mahalanobis matrix.

Fusing different low level features together could provide more rich information. We consider four different appearance based features: color histograms (CH or namely HSV) (Lu and Shengjin, 2015), color names (CN) (Berlin and Kay, 1991; Van de Weijer et al., 2007), HOG (Dalal and Triggs, 2005), and SILTP (Liao et al., 2010) to cover both color and texture characteristics. They are all $l_1$ normalized followed by $\sqrt{(\cdot)}$ operator before BoW quantization, as the Euclidean distance on root feature space is equivalent to the Hellinger distance on original feature

space, and Hellinger kernel performs better considering histogram similarity (Arandjelović and Zisserman, 2012). The fusion is applied at image descriptor level, which has been demonstrated effective. Different codebooks $\mathcal{C}^{HSV}$, $\mathcal{C}^{CN}$, $\mathcal{C}^{HOG}$, and $\mathcal{C}^{SILTP}$ are generated for each low level feature separately, thus the BoW image descriptor of each feature is calculated respectively. Then the final descriptor of image $i$ is concatenated as $\mathbf{d}_i = [\mathbf{d}_i^{HSV}, \mathbf{d}_i^{CN}, \mathbf{d}_i^{HOG}, \mathbf{d}_i^{SILTP}]$.

#### 3.1.1 Color Histograms

HSV is typically used to describe color characteristics within one region. First, the image is transferred to the HSV color space. Then the statistical distribution of hue (H) and saturation (S) channels is calculated respectively with each channel quantized to 10 bins. Luminance (V) channel is excluded because of huge illumination changes in person re-identification tasks.

#### 3.1.2 Color Names

CN are semantic attributes obtained through assigning linguistic color labels to image pixels. Here, we use the descriptors learned from real-world images like Google Images to map RGB values of a pixel to 11 color terms (Van de Weijer et al., 2007). The CN descriptor assigns each pixel an 11-D vector, each dimension corresponding to one of the 11 basic colors. Afterward, the CN descriptor of a superpixel region is computed as the average value of each pixel.

#### 3.1.3 HOG

HOG is a classical texture descriptor which counts occurrences of gradient orientation in localized portions of an image. We separate gradient orientation into 9 bins and calculate on the gray image.

#### 3.1.4 Scale Invariant Local Ternary Pattern

SILTP (Liao et al., 2010) descriptor is an improved operator over the well-known Local Binary Pattern (LBP) (Ojala et al., 1996). LBP has a nice invariant property under monotonic gray-scale transforms, however, it is not robust to image noises. SILTP improves LBP by introducing a scale invariant local comparison tolerance, achieving invariance to intensity scale changes and robustness to image noises. Within each superpixel, we extract 2 scales of SILTP histograms ($SILTP^{0.3}_{4,3}$ and $SILTP^{0.3}_{4,5}$) as suggested in (Liao et al., 2015).

### 3.2 Bag-of-Words Framework and Codebook Generation

Codebook generation is a critical step of building the BoW model. Conventional approach simply clusters low level appearance features by unsupervised k-means in Euclidean space. In this paper, we suggest applying supervised metric learning methods and cluster features in Mahalanobis space with its trained distance metrics.

We denote the feature vector of superpixel $k$ in the strip $j$ of image $i$ as $\mathbf{f}_{i,j,k}$, whereas $\mathbf{f}_{i,j,k} \in \mathcal{R}^d$ and $d$ is the feature vector length. And ($\mathbf{f}_{i1,j,k1}$, $\mathbf{f}_{i2,j,k2}$) is a pairwise feature instance where they belong to two superpixels in the same horizontal strip $j$ of two different images. Here, only features belonging to the same horizontal strip are collected as pairwise instance, which is quite reasonable because of the geometric constraints of pedestrian images, thus dramatically reduce the amount of pairwise feature instances as well as the computational complexity. We further denote $\mathcal{P}$ as the positive set of pairwise feature instances where the first feature and the second feature belong to same person, i.e, $(\mathbf{f}_{i1,j,k1}, \mathbf{f}_{i2,j,k2}) \in \mathcal{P}, id(i1) = id(i2)$. And we denote $\mathcal{N}$ as the negative set of pairwise feature instances, i.e, $(\mathbf{f}_{i1,j,k1}, \mathbf{f}_{i2,j,k2}) \in \mathcal{N}, id(i1) \neq id(i2)$. The goal of our task is to learn a distance metric $\mathbf{M}'$ (to be distinguished with $\mathbf{M}$ in conventional part two metric learning methods) to effectively measure distance between any two visual features $\mathbf{f}_{i1,j,k1}$ and $\mathbf{f}_{i2,j,k2}$, which is often represented as

$$d(\mathbf{f}_{i1,j,k1}, \mathbf{f}_{i2,j,k2}) = \sqrt{(\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2})^T \mathbf{M}' (\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2})}, \quad (5)$$

where matrix $\mathbf{M}'$ is the $d \times d$ Mahalanobis matrix that must be positive and semi-definite.

Many metric learning methods are proposed to learn an optimized $\mathbf{M}'$. In this paper, we use KISSME (Köstinger et al., 2012) and apply it in our BoW codebook generation. KISSME is a bayesian method and only assumes $(\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2})$ is gaussian distribution, which is quite reasonable in our case. The computation is simple yet the algorithm is effective:

$$\Delta_{\mathbf{P}} = \sum_{(\mathbf{f}_{i1,j,k1}, \mathbf{f}_{i2,j,k2}) \in \mathbf{P}} (\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2}) \cdot (\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2})^T \quad (6)$$

$$\Delta_{\mathbf{N}} = \sum_{(\mathbf{f}_{i1,j,k1}, \mathbf{f}_{i2,j,k2}) \in \mathbf{N}} (\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2}) \cdot (\mathbf{f}_{i1,j,k1} - \mathbf{f}_{i2,j,k2})^T \quad (7)$$

$$\mathbf{M}' = \Delta_{\mathbf{P}}^{-1} - \Delta_{\mathbf{N}}^{-1}. \quad (8)$$

Our codebook can be generated by clustering low-level features under the learned distance metric as above. We collect all the features with background removed. Then k-means clustering is applied based on the optimized Mahalanobis distance metric $\mathbf{M}'$. Finally, we build our codebook on the clustering centers.

Applying our codebook in test phase is straightforward. We first extract low-level features from a novel test image. Then the feature is compared with visual words in the codebook by the trained Mahalanobis distance $\mathbf{M}'$. Finally, the visual word histogram of a pedestrian image strip is calculated and the image descriptor is the concatenation of all stripes in one image.

The image descriptor generated above can be compared directly under Euclidean distance or conventional part two metric learning methods. These part two metric learning methods operate on image descriptor level, while our proposed method operates on low level visual features in part one. We will demonstrate in section 4 that our proposed method can be directly integrated with these conventional methods with a significant performance boost.

# 4 EXPERIMENTS

To evaluate the effectiveness of our method, we conducted experiments on 3 public benchmark datasets: the VIPeR (Gray et al., 2007), the PRID 450S (Roth et al., 2014), and the Market-1501 (Zheng et al., 2015; Zheng et al., 2016a) datasets. The conventional evaluation protocol split the dataset into training and test part. For unsupervised methods evaluation, only test samples are used. The BoW codebook size is set to 350 for each feature. An average of 500 superpixels per image are generated by SLIC method and its compactness parameter is set to 20. Considering re-identification as a ranking problem, the performance is measured in Cumulative Matching Characteristics (CMC).

## 4.1 Datasets

### 4.1.1 VIPeR

The 1264 images which are normalized to 128x48 pixels in the VIPeR dataset are captured from 2 different cameras in outdoor environment, including 632 individuals and 2 images for each person. It is the large variances in viewpoint, pose, resolution, and illumination that makes VIPeR very challenging. In conventional evaluations, the dataset is randomly divided into 2 equal parts, one for training, and the other for testing. In one trial, images are taken as probe sequentially and matched against the opposite camera. 10 trials are repeated and the average result is calculated.

### 4.1.2 PRID 450S

450 single-shot image pairs depicting walking humans are captured from 2 disjoint surveillance cameras. Pedestrian bounding boxes are manually labeled with a vertical resolution of 100-150 pixels, while the resolution of original images is 720*576 pixels. Moreover, part-level segmentation is provided describing the following regions: head, torso, legs, carried object at torso level (if any) and carried object below torso (if any). Like VIPeR, we randomly partition the dataset into two equal parts, one for training, and the other for testing. 10 trials are repeated.

### 4.1.3 Market-1501

Market-1501 consists of 32668 detected person bounding boxes of 1501 individuals captured by 6 cameras (5 high-resolution and 1 low-resolution) with overlaps. Each identity is captured by 2 cameras at least, and may have multiple images in one camera. For each identity in test, one query image in each camera is selected, therefore multiple queries are used for each identity. Note that, the selected 3368 queries are hand-drawn, instead of DPM-detected as in the gallery. The provided fixed training and test set are used under both single-query and multi-query evaluation settings.

## 4.2 Exploration of Metric Learning in BoW Codebook Generation

We first compare the performance of our proposed method against conventional baseline BoW approaches on VIPeR dataset. The performance is evaluated on 3 different part two metric learning methods (KISSME (Köstinger et al., 2012), XQDA (Liao et al., 2015), Null Space (Zhang et al., 2016a)) on image descriptor level respectively as well as directly applying Euclidean distance on image descriptors without part two metric learning methods. The baseline method applies BoW descriptor simply on Euclidean space without any pedestrian labels, which is totally unsupervised. As shown in Figure 2, our proposed method performs better than baseline method with 1.7% rank 1 recognition rate gain. When part two metric learning methods are integrated, the performance gain on

Table 1: Comparison to the State-of-the-Art Results on VIPeR.

| method | r1 (%) | r5 (%) | r10 (%) | r20 (%) | r30 (%) |
|---|---|---|---|---|---|
| SCSP (Chen et al., 2016a) | 53.5 | 82.6 | 91.5 | 96.6 | - |
| Kernel X-CRC (Prates and Schwartz, 2016) | 51.6 | 80.8 | 89.4 | 95.3 | 97.4 |
| FFN (Wu et al., 2016b) | 51.1 | 81.0 | 91.4 | 96.9 | - |
| Triplet Loss (Cheng et al., 2016) | 47.8 | 74.7 | 84.8 | 91.1 | 94.3 |
| LSSL (Yang et al., 2016) | 47.8 | 77.9 | 87.6 | 94.2 | - |
| Metric Ensembles (Paisitkriangkrai et al., 2015) | 44.9 | 76.3 | 88.2 | 94.9 | - |
| LSSCDL (Zhang et al., 2016b) | 42.7 | - | 84.3 | 91.9 | - |
| LOMO + Null Space (Zhang et al., 2016a) | 42.3 | 71.5 | 82.9 | 92.1 | - |
| NLML (Huang et al., 2015) | 42.3 | 71.0 | 85.2 | 94.2 | - |
| Semantic Representation (Shi et al., 2015) | 41.6 | 71.9 | 86.2 | 95.1 | - |
| WARCA (Jose and Fleuret, 2016) | 40.2 | 68.2 | 80.7 | 91.1 | - |
| LOMO + XQDA (Liao et al., 2015) | 40.0 | 68.0 | 80.5 | 91.1 | 95.5 |
| Deep Ranking (Chen et al., 2016b) | 38.4 | 69.2 | 81.3 | 90.4 | 94.1 |
| SCNCD (Yang et al., 2014) | 37.8 | 68.5 | 81.2 | 90.4 | 94.2 |
| Correspondence Structure Learning (Shen et al., 2015) | 34.8 | 68.7 | 82.3 | 91.8 | 94.9 |
| Baseline BoW | 48.7 | 77.5 | 87.0 | 93.9 | - |
| **Proposed + Null Space** | 50.0 | 79.0 | 88.1 | 94.5 | 97.0 |

Table 2: Comparison to the State-of-the-Art Results on PRID 450S.

| method | r1 (%) | r5 (%) | r10 (%) | r20 (%) | r30 (%) |
|---|---|---|---|---|---|
| Kernel X-CRC (Prates and Schwartz, 2016) | 68.8 | 91.2 | 95.9 | 98.4 | 99.0 |
| FFN (Wu et al., 2016b) | 66.6 | 86.8 | 92.8 | 96.9 | - |
| LSSCDL (Zhang et al., 2016b) | 60.5 | - | 88.6 | 93.6 | - |
| Semantic Representation (Shi et al., 2015) | 44.9 | 71.7 | 77.5 | 86.7 | - |
| Correspondence Structure Learning (Shen et al., 2015) | 44.4 | 71.6 | 82.2 | 89.8 | 93.3 |
| SCNCD (Yang et al., 2014) | 41.6 | 68.9 | 79.4 | 87.8 | 95.4 |
| Baseline BoW | 68.0 | 88.0 | 93.8 | 97.2 | - |
| **Proposed + Null Space** | **70.7** | **90.7** | **94.8** | **97.8** | **99.2** |

rank 1 recognition rate reaches 1.8% with KISSME metric learning, 0.7% with XQDA metric learning, and 1.3% with Null Space metric learning.

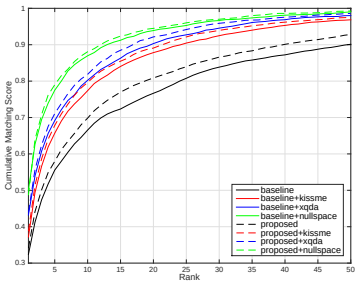


Figure 2: CMC curves on the VIPeR dataset, by comparing the proposed approach to conventional baseline methods. Euclidean distance, KISSME, XQDA, and Null Space are employed on image descriptor level respectively.

The improvement of our proposed method against baseline BoW method is most notable, because the baseline method is totally unsupervised, while the proposed method applies supervised label data on BoW low level feature level. The baseline method with KISSME metric learning outperforms our proposed method without any part two metric learning methods, which suggests that our proposed local feature level metric learning method is an improvement but not replacement of conventional image descriptor level metric learning methods.

## 4.3 Comparison to the State-of-the-Art Results

In this section, we compare our proposed method with the state-of-the-art approaches. Specifically, we adopt Null Space as the part two image descriptor level metric learning method.

We first compare our approach with the state-of-the-art results on VIPeR in Table 1. We obtain a rank 1 re-identification rate of 50.0% on VIPeR, which is comparable to the best result.

Table 2 compares our results to the state-of-the-art approaches on PRID 450S. We yields rank 1 re-

Table 3: Comparison to the State-of-the-Art Results on Market-1501.

| | methods | r1 (%) | mAP (%) |
|---|---|---|---|
| Metric learning | WARCA (Jose and Fleuret, 2016) | 45.16 | - |
| | TMA (Martinel et al., 2016) | 47.92 | 22.31 |
| | SCSP (Chen et al., 2016a) | 51.90 | 26.35 |
| | LOMO+Null Space (Zhang et al., 2016a) | 55.43 | 29.87 |
| | Baseline BoW | 63.87 | 36.04 |
| | **Proposed+Null Space** | **64.13** | **36.21** |
| Deep learning | PersonNet (Wu et al., 2016a) | 37.21 | 18.57 |
| | CAN (Liu et al., 2016a) | 48.24 | 24.43 |
| | SSDAL (Su et al., 2016) | 39.4 | 19.6 |
| | Triplet CNN (Liu et al., 2016b) | 45.1 | - |
| | Histogram Loss (Ustinova and Lempitsky, 2016) | 59.47 | - |
| | Gated Siamese CNN (Varior et al., 2016) | **65.88** | **39.55** |

identification rate of 70.7% with Null Space metric learning, which is superior to the best result (Prates and Schwartz, 2016) by 1.9%.

As for the large scale datasets like Market-1501, we roughly classify supervised learning methods into two categories, the first conventional metric learning based approaches, and the second deep learning based approaches. Our method yields rank 1 recognition of 64.13% and mAP of 36.21% under the single query mode with Null Space (Zhang et al., 2016a) metric learning, which outperforms the best metric learning approaches by 8.7% on rank 1 and 6.3% on mAP, as shown in Table 3. Our result even outperforms many other deep learning based approaches and is comparable to the recent state-of-the-art method Gated Siamese CNN (Varior et al., 2016), which is quite outstanding because Market-1501 is generally considered more suitable for deep learning based methods with its large image volume.

# 5 CONCLUSIONS

In this paper, we propose an improved BoW method that learns a suitable metric distance of low level features in codebook generation for person re-identification. The approach uses KISSME metric learning for local features, and can be effectively integrated with conventional image descriptor level metric learning algorithms. Experiments demonstrate the effectiveness and robustness of our method. The proposed method outperforms state-of-the-art results on VIPeR, PRID 450S, and Market-1501 integrated with part two Null Space metric learning method.

# ACKNOWLEDGEMENTS

The work was supported by the National Natural Science Foundation of China under Grant Nos. 61071135 and the National Science and Technology Support Program under Grant No. 2013BAK02B04.

# REFERENCES

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282.

Ahmed, E., Jones, M., and Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916.

Arandjelović, R. and Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2911–2918. IEEE.

Berlin, B. and Kay, P. (1991). *Basic color terms: Their universality and evolution*. Univ of California Press.

Chen, D., Yuan, Z., Chen, B., and Zheng, N. (2016a). Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1268–1277.

Chen, D., Yuan, Z., Hua, G., Zheng, N., and Wang, J. (2015). Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1565–1573.

Chen, S.-Z., Guo, C.-C., and Lai, J.-H. (2016b). Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367.

Cheng, D., Gong, Y., Zhou, S., Wang, J., and Zheng, N. (2016). Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Das, A., Chakraborty, A., and Roy-Chowdhury, A. K. (2014). Consistent re-identification in a camera network. In *European Conference on Computer Vision*, pages 330–345. Springer.

Ding, S., Lin, L., Wang, G., and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE.

Gong, S., Cristani, M., Yan, S., and Loy, C. C. (2014). *Person re-identification*, volume 1. Springer.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3. Citeseer.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer.

Gray, R. (1984). Vector quantization. *IEEE Assp Magazine*, 1(2):4–29.

Hirzer, M., Roth, P. M., Köstinger, M., and Bischof, H. (2012). Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, pages 780–793. Springer.

Huang, S., Lu, J., Zhou, J., and Jain, A. K. (2015). Nonlinear local metric learning for person re-identification. *arXiv preprint arXiv:1511.05169*.

Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer.

Jose, C. and Fleuret, F. (2016). Scalable metric learning via weighted approximate rank component analysis. *arXiv preprint arXiv:1603.00370*.

Köstinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., and Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3610–3617.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206.

Liao, S. and Li, S. Z. (2015). Efficient psd constrained asymmetric metric learning for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3685–3693.

Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., and Li, S. Z. (2010). Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1301–1306. IEEE.

Liu, H., Feng, J., Qi, M., Jiang, J., and Yan, S. (2016a). End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*.

Liu, J., Zha, Z.-J., Tian, Q., Liu, D., Yao, T., Ling, Q., and Mei, T. (2016b). Multi-scale triplet cnn for person re-identification. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 192–196. ACM.

Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., and Bu, J. (2014). Semi-supervised coupled dictionary learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3557.

Lu, T. and Shengjin, W. (2015). Person re-identification as image retrieval using bag of ensemble colors. *IEICE TRANSACTIONS on Information and Systems*, 98(1):180–188.

Luo, P., Wang, X., and Tang, X. (2013). Pedestrian parsing via deep decompositional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2648–2655.

Martinel, N., Das, A., Micheloni, C., and Roy-Chowdhury, A. K. (2016). Temporal model adaptation for person re-identification. In *European Conference on Computer Vision*, pages 858–877. Springer.

Ojala, T., Pietikäinen, M., and Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59.

Paisitkriangkrai, S., Shen, C., and van den Hengel, A. (2015). Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855.

Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. (2013). Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325.

Prates, R. and Schwartz, W. R. (2016). Kernel cross-view collaborative representation based classification for person re-identification. *arXiv preprint arXiv:1611.06969*.

Roth, P. M., Hirzer, M., Koestinger, M., Beleznai, C., and Bischof, H. (2014). Mahalanobis distance learning for person re-identification. In Gong, S., Cristani, M., Yan, S., and Loy, C. C., editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition, pages 247–267. Springer, London, United Kingdom.

Scholkopft, B. and Mullert, K.-R. (1999). Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1(1):1.

Shen, Y., Lin, W., Yan, J., Xu, M., Wu, J., and Wang, J. (2015). Person re-identification with correspondence structure learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3200–3208.

Shi, Z., Hospedales, T. M., and Xiang, T. (2015). Transferring a semantic representation for person re-identification and search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4184–4193.

Su, C., Zhang, S., Xing, J., Gao, W., and Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. *arXiv preprint arXiv:1605.03259*.

Ustinova, E. and Lempitsky, V. (2016). Learning deep embeddings with histogram loss. In *Advances In Neural Information Processing Systems*, pages 4170–4178.

Van de Weijer, J., Schmid, C., and Verbeek, J. (2007). Learning color names from real-world images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

Varior, R. R., Haloi, M., and Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244.

Wu, L., Shen, C., and Hengel, A. v. d. (2016a). Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*.

Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., and Zheng, W.-S. (2016b). An enhanced deep feature representation for person re-identification. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE.

Yang, Y., Liao, S., Lei, Z., and Li, S. Z. (2016). Large scale similarity learning using similar pairs for person verification. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., and Li, S. Z. (2014). Salient color names for person re-identification. In *European Conference on Computer Vision*, pages 536–551. Springer.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE.

Zhang, L., Xiang, T., and Gong, S. (2016a). Learning a discriminative null space for person re-identification. *arXiv preprint arXiv:1603.02139*.

Zhang, Y., Li, B., Lu, H., Irie, A., and Ruan, X. (2016b). Sample-specific svm learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhao, R., Ouyang, W., and Wang, X. (2013). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016a). Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer.

Zheng, L., Huang, Y., Lu, H., and Yang, Y. (2017). Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*.

Zheng, L., Yang, Y., and Hauptmann, A. G. (2016b). Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*.

Zheng, Z., Zheng, L., and Yang, Y. (2016c). A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*.

Zhou, X., Cui, N., Li, Z., Liang, F., and Huang, T. S. (2009). Hierarchical gaussianization for image classification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1971–1977. IEEE.