Unsupervised Learning of Multi-Level Descriptors for Person Re-Identification

Yang Yang,[1] **Longyin Wen,**[2] **Siwei Lyu,**[2] **Stan Z. Li**[1]
[1]Center for Biometrics and Security Research & National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190
[2]University at Albany, State University of New York, Albany, NY, 12222
yyangcv@gmail.com, lywen.cv.workbox@gmail.com, slyu@albany.edu, szli@nlpr.ia.ac.cn

Abstract

In this paper, we propose a novel coding method named weighted linear coding (WLC) to learn multi-level (e.g., pixel-level, patch-level and image-level) descriptors from raw pixel data in an unsupervised manner. It guarantees the property of saliency with a similarity constraint. The resulting multi-level descriptors have a good balance between the robustness and distinctiveness. Based on WLC, all data from the same region can be jointly encoded. Consequently, when we extract the holistic image features, it is able to preserve the spatial consistency. Furthermore, we apply PCA to these features and compact person representations are then achieved. During the stage of matching persons, we exploit the complementary information resided in multi-level descriptors via a score-level fusion strategy. Experiments on the challenging person re-identification datasets - VIPeR and CUHK 01, demonstrate the effectiveness of our method.

1 Introduction

Recent years have seen a major progress in person re-identification. Its objective is to automatically associate images of the same person captured by cameras at different locations and time. This is a particular issue in a network of cameras collaborating with each other, in which the appearances of the same person may exhibit drastic variations in different camera views due to difference in illumination, view angles and poses.

A popular pipeline in existing person re-identification methods includes person representation and similarity learning. In the former, one seeks features of detected human image that are both robust to intra-personal variations in illumination, view angles and poses and distinctive in inter-personal description (Yang et al. 2014b; Zhao, Ouyang, and Wang 2014; Yang et al. 2014a; Liao et al. 2015). In the latter, one aims to find a proper metric or metrics to measure the similarity between any two images from disjoint cameras (Köstinger et al. 2012; Yang et al. 2016b; Li et al. 2015). Considering that image representation is arguably the most fundamental task, we focus on how to extract features which have both intra-personal invariance and inter-personal discrimination.

In previous works (Yang et al. 2014b; Zhao, Ouyang, and Wang 2014; Yang et al. 2014a; Shen et al. 2015; Köstinger et al. 2012; Farenzena et al. 2010; Kviatkovsky, Adam, and Rivlin 2013; Zhao, Ouyang, and Wang 2013; Karanam, Li, and Radke 2015; Li and Wang 2013; Yang, Jin, and Tao 2012), color and texture are often used to characterize human appearance in person re-identification. Handcrafted textural descriptors (e.g., SIFT, LBP) achieve good performances in image classification (Wang et al. 2010; Huang et al. 2011) and face recognition (Ahonen, Hadid, and Pietikäinen 2006). Whereas the performance is pretty poor when they are directly applied and used alone in person re-identification (Yang, Jin, and Tao 2012). This is because person images from cameras are often low in resolution with many image noises, thereby leading to the difficulty in accurately describing the image textures.

Among the color features, color histograms (e.g., RGB histogram) are most commonly used. However, it is unstable when there are large variations in illumination (Yang et al. 2014a). To improve the robustness of color features, Yang et.al. (Yang et al. 2014b) propose a novel salient color name based color descriptor to describe colors. Although it shows good performance in addressing the issue of person re-identification, it may not sufficiently distinguish different persons with similar clothing color. In view of this, a common strategy is to combine the color features with textural features to compensate each other (Köstinger et al. 2012; Farenzena et al. 2010; Karanam, Li, and Radke 2015; Yang, Jin, and Tao 2012).

In this paper, we propose a coding method to learn descriptors from raw pixel data. This is motivated by (Yang et al. 2014b), which utilizes a learned 16-dimensional color descriptor to replace raw 3-dimensional pixel. Different with (Yang et al. 2014b), our method learns three types of descriptors that contains multi-levels: pixel-level, patch-level and image-level, respectively. Each level has a corresponding relationship with a particular granularity of context. Specifically, pixel-level represents color information, patch-level corresponds to textural information or local shape patterns, and image-level provides entity (i.e., tracked person) information. As such, the integration of features based on these descriptors can exploit the complementary information resided in them, thus leading to more reliable, robust and distinct features for person re-identification.

Specifically, we address the problem of person re-identification based on the following model: 1) learning multi-level descriptors from raw pixel data, 2) extracting holistic image features and 3) fusing in the score-level. Among them, the speed and effectiveness of the first step largely determine the performances of our method. Therefore, our main contribution is to provide an efficient and effective solution to it. Motivated by the success of coding methods – locality-constrained linear coding (LLC) (Wang et al. 2010) and salient coding (SC) (Huang et al. 2011) for bag of visual words models in object recognition and scene classification tasks, we present a simple and effective coding approach that we refer to as ***weighted linear coding*** (WLC) to learn descriptors from unlabeled raw pixel data. WLC is a reconstruction based coding method that finds coefficients to minimize the coding error. In WLC, we use a similarity constraint to weigh the importance of each coding coefficient by means of lowering the corresponding coding value on the bases (obtained by k-means) which is dissimilar with the input data. It guarantees the property of saliency. WLC also includes an F-norm regularization to the coefficient matrix to avoid coding coefficient becoming unbounded or overfit to the training data. Compared to LLC and SC, WLC does not require the process of finding local bases and can encode all input data simultaneously. Hence, based on WLC, multi-level descriptors can be learned quickly and efficiently. With the learned descriptors from raw pixel data, we can extract different holistic image features under a horizontal stripes partition (Yang et al. 2014b). Finally, a score-level fusion is utilized to compute the similarity between any two images based on these different holistic image features. Though it is simple, our model obtains good performance on both the VIPeR (Gray, Brennan, and Tao 2007) and the CUHK 01 (Li, Zhao, and Wang 2012) benchmarks. We have released the MATLAB code [1] for future research on person re-identification.

2 Related Work

In this section, we simply review several existing person representation methods for person re-identification.

Color and hand-crafted feature. Recent works (Farenzena et al. 2010; Kviatkovsky, Adam, and Rivlin 2013; Zhao, Ouyang, and Wang 2013; Yang et al. 2014b; Zhao, Ouyang, and Wang 2014; Yang et al. 2014a; Liao et al. 2015; Yang, Jin, and Tao 2012; Li et al. 2012) have shown the importance of features in person re-identification. Reliable features should come from all cues that can provide identification of individual tracked humans, which include color, shape and context. In (Farenzena et al. 2010), an approach named Symmetry-Driven Accumulation of Local Features is proposed, which extracts features modeling three complementary aspects of human appearance: HSV histogram, Maximally Stable Color Regions and the Recurrent Highly Structured Patches. The extracted features are weighted according to the distance from the vertical axis, thus minimizing the effects of pose variations. Although color distribution changes under different imaging condition, the intradistribution structure is invariant while being discriminative (Kviatkovsky, Adam, and Rivlin 2013). With color information being the only cue, good recognition performances are achieved. In (Yang et al. 2014a), the illumination invariance and distinctiveness of different color models are evaluated for person re-identification. Then, features in different color models are fused and good results are obtained. To increase the robustness of colors, salient color names based color description (SCNCD) is proposed in (Yang et al. 2014b), which also shows complementary information with traditional color histograms. In (Liao et al. 2015), an effective feature representation named local maximal occurrence (LOMO) is presented, which maximizes the horizontal occurrence of local features to make a stable representation against viewpoint changes. Matsukawa et.al. (Matsukawa et al. 2016) further present a novel region descriptor based on hierarchical Gaussian distribution of pixel features, which represents the region as a set of multiple Gaussian distributions.

Attribute feature. Attribute features are also introduced for solving person re-identification task in (Layne, Hospedales, and Gong 2012). The extracted attribute features are often more powerful than original features and can be interpreted by humans. Due to the difficulty of obtaining sufficient domain-specific annotations, Shi et.al. (Shi, Hospedales, and Xiang 2015) put forward a new method of transferring a learned semantic attribute model from existing fashion datasets to person re-identification dataset.

Learning based feature. To make the features discriminative, a novel method named metric embedded discriminative vocabulary learning is proposed in (Yang et al. 2016a). The obtained image-level features of the same persons are closer while different ones farther in the metric space. Additionally, dictionary learning methods are also used in person re-identification. In (Liu et al. 2014), two coupled dictionaries, which relate to different cameras, are learned to make the extracted features robustness with different views. Different with (Liu et al. 2014), only one dictionary is discriminatively trained that is viewpoint invariant in (Karanam, Li, and Radke 2015).

In addition, convolutional neural network (CNN) has also been adopted in person re-identification (GuangrunWang et al. 2016; Zhu et al. 2014; Ahmed, Jones, and Marks 2015). It explores the hierarchical structure and extract discriminative features based on training samples. However, due to the lack of sufficient labeled training samples, the performance of CNN based methods is not superior to traditional ones on small datasets, e.g., VIPeR.

3 Method

In this section, we describe our method in detail. We first introduce our coding method, WLC, to learn multi-level (pixel-level, patch-level and image-level) descriptors from raw pixel data in Sec. 3.1. Then, three types of holistic image features are extracted under the horizontal stripes partition (Sec. 3.2). The last step is to combine the extracted features by fusing their individual similarity scores in matching two images for person re-identification (Sec. 3.3).

[1] http://www.cbsr.ia.ac.cn/users/yyang/main.htm.

3.1 Weighted Linear Coding

To learn multi-level descriptors, we propose a simple yet efficient coding method. It aims to decompose the input data over a basis (or base) set while satisfying certain requirements (e.g., ”saliency-aware”). Unlike traditional handcrafted descriptors, our descriptors are learned respective to specifical training samples in an unsupervised manner and thus contain some kind of statistical property in the dataset.

Previous work (Coates and Ng 2011) has shown that for the recognition task, the basis set design is less critical than coding. Therefore, we focus our efforts on how to design the coding scheme while the basis set (or a codebook) is generated by k-means. Let $X = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_n] \in \mathcal{R}^{d\times n}$ be n d-dimensional input data (raw pixel data) from the same region (e.g., in the same stripe). Given a set of k basis vectors $B = [\vec{b}_1, \vec{b}_2, ..., \vec{b}_k] \in \mathcal{R}^{d\times k}$, we introduce the following coding scheme to learn multi-level descriptors:

$$\min_{S} \|X - BS\|_F^2 + \lambda_1 \|W^T S\|_F^2 + \lambda_2 \|S\|_F^2, \qquad (1)$$

where λ_1 and λ_2 are positive constants. $\|S\|_F = \sqrt{tr(S^T S)}$ is the Frobenius norm of matrix S. $S \in \mathcal{R}^{k\times n}$ is the matrix containing all the coding coefficients. W_{ij} of $W \in \mathcal{R}^{k\times n}$ is defined as

$$W_{i,j} = \frac{1}{Z} \frac{\|\vec{x}_j - \vec{b}_i\|_2^2}{\frac{1}{k-1}\sum_{l\neq i} \|\vec{x}_j - \vec{b}_l\|_2^2} \qquad (2)$$

with Z is a normalization factor (chosen so that the sum of j-th column of W equals to 1), $i = 1, ..., k$ and $j = 1, ..., n$. As (1) shows, WLC is a reconstruction based coding method. For each input data, its corresponding coding coefficient is used as the learned descriptors.

The first term of (1) is the reconstruction error. It chooses all basis vectors to reconstruct the input. The second term of (1) is a similarity constraint which corresponds to the requirement of the coding coefficient to be consistent with the similarity criterion, i.e., the coefficients over bases, which are different from the input data, should be small. As such, it can be regarded as a saliency-aware term (Huang et al. 2011), which is a relaxation of the locality constraint used in LLC (Wang et al. 2010). The third term is an F-norm regularization which penalizes large coefficients, and serves as a regularizer to the optimization problem.

With the bases fixed, the objective function of (1) is a convex quadratic function of S. Then, we can solve it by taking the derivative of (1) with regards to S to zero:

$$\frac{\partial}{\partial S}\mathcal{G}(S) = 2(-B^T X + B^T BS + \lambda_1 WW^T S + \lambda_2 S) = 0, \qquad (3)$$

where $\mathcal{G}(S) = \|X - BS\|_F^2 + \lambda_1 \|W^T S\|_F^2 + \lambda_2 \|S\|_F^2$.

$$S = (B^T B + \lambda_1 WW^T + \lambda_2 I)^{-1}(B^T X), \qquad (4)$$

where $I \in \mathcal{R}^{k\times k}$.

In comparison with closely related methods – LLC and SC, WLC has better running efficiency. This is due to two factors: (1) Both LLC and SC require the assignment of the input data to several local bases. This procedure will increase their computation complexity and lead to longer running time. By contrast, there is no need to select bases for each data point and running time can be saved on that step. (2) Because different bases may be selected for different input data, both LLC and SC can only encode one input data (e.g., $\vec{x}_1$) at a time. However, our WLC is able to encode all the input data simultaneously.

	b_1	b_2	b_3	b_4	b_5
LLC	0	0	0.291	0.334	0.375
SC	0	0	-0.698	0.874	-0.858
WLC	-0.070	-0.280	0.402	0.955	0.042

Table 1: A comparison of the response codes based on LLC, SC and WLC.

Table 1 gives an example of responses in LLC, SC and WLC. Assume that there are 5 bases and both LLC and SC use 3 nearest bases to encode the input. From Table 1, we can find that both LLC and SC have high responses over b_3, b_4 and b_5, which are 3 closest bases to the input. Additionally, the response codes are determined by reconstruction error and salient degree for LLC and SC, respectively. Unlike them, WLC has high responses over b_2, b_3 and b_4 while low responses over others. This phenomenon reflects that under our definition of similarity constraint in (1), b_1 and b_5 are different from the input and thus the corresponding coding coefficient are small. The response codes in WLC are determined by the reconstruction error, the similarity constraint and the F-norm regularization.

We should also note that in (4), when we use WLC to encode the input data X, all of $\vec{x}_i, i = 1, 2, ..., n$ share the same WW^T. That is to say, all elements in X will be transformed using the same matrix $(B^T B + \lambda_1 WW^T + \lambda_2 I)^{-1} B^T$. This is useful in the feature extraction stage because we can jointly encode all data from one stripe, and their relationships (similarity) with the bases are shared. Therefore, WLC can preserve spatial consistency, i.e., data from the same spatial locations (in the same stripe) are considered simultaneously while those from different spatial locations will be processed independently.

With (4), we can obtain multi-level descriptors from raw pixel data. An example is shown in Fig. 1. For patch-level and image-level, we simply vectorize the corresponding matrix as the input data. For example, the dimensions of raw pixel data (a vector) from a $u\times v$ image are 3×1, $7\times 7\times 3$ (for a 7×7 patch, see Sec. 4.2) and $u \times v \times 3$ in pixel-level, patch-level and image-level, respectively. After WLC, multi-level descriptors are obtained. The dimensions of these descriptors depend on the number of bases in the codebook B.

From the (3), we know that the computation complexity of WLC is $\mathcal{O}(k^2 d + 2k^3 + kdn + k^2 n)$. Since the dimensions of the input data are 3, 147 and 60 for pixel-level, patch-level and image-level (see Sec. 4.2), respectively, its speed mainly depends on the value of k and n.

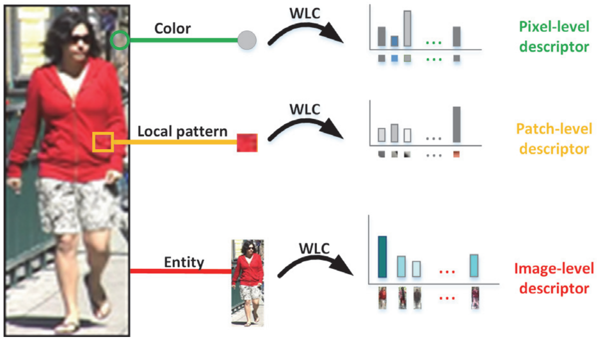


Figure 1: An example of multi-level descriptors learning based on WLC.

Method	pixel (s)	patch (s)	image (10^{-5}s)
LLC	0.51	0.43	6.88
SC	0.12	0.10	2.06
WLC	**0.03**	**0.07**	**1.26**

Table 2: Computing time based on WLC, SC and LLC for extracting features of an image (128×48). This is evaluated on a PC with the 3.40 GHz Core I7 CPU with 8 cores.

3.2 Extraction of Holistic Image Features

Due to camera-view changes or different poses, there exist many uncontrolled spatial misalignment problems for person re-identification. To overcome this problem, we adopt the horizontal stripes partition which divides an image into several horizontal stripes of equal size. Corresponding horizontal stripes have similar statistical information albeit being misaligned spatially. Furthermore, to reduce the effects of misalignment, we pool together the codes (pixel-level or patch-level descriptors) in each stripe by *average pooling* and then create a normalized histogram. Finally, the holistic image feature based on the pixel-level (or patch-level) descriptors is extracted by concatenating the histograms of all stripes while the image-level descriptor is directly used as the final holistic image feature.

Because of the running efficiency of WLC, we can extract different level features very fast. Table 2 shows the computing time for extracting multi-level features of an image (128×48).

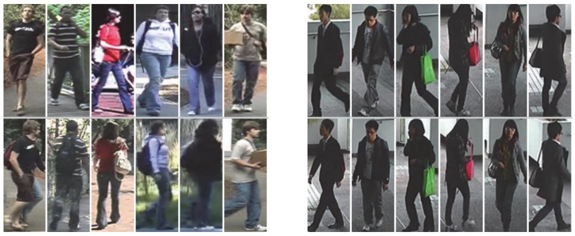
3.3 Score-level Fusion

With different holistic image features of all levels, we combine their effects in person re-identification at the step of computing similarity scores. The similarity between two images is computed as an unweighted[2] sum of the similarity scores of their corresponding features.

Specifically, we first use a similarity learning method (see Sec. 4.2) to compute the similarity score $SS(i), i = 1, 2, ..., m$ of the i-th holistic image feature for a given pair. If we use the mask, m is 16 (three levels in four color spaces without using mask and one level in four color spaces using mask). Then, the final similarity score FS is calculated based on:

$$FS = \sum_{i=1}^{m} SS(i). \qquad (5)$$

Finally, we can easily obtain the matching results based on their ranking and promising results are achieved.



(a) (b)

Figure 2: Some examples from (a) VIPeR and (b) CUHK 01. Each column is an image pair from one person.

4 Experiments

In this section, we evaluate our model on two publicly available datasets (VIPeR dataset and CUHK 01 dataset).

4.1 Datasets

VIPeR Dataset. The viewpoint invariant pedestrian recognition (VIPeR) dataset contains 632 image pairs, corresponding to 632 pedestrians. It was captured by two cameras in outdoor academic environments. This dataset is the most widely used in person re-identification. There are arbitrary viewpoints, pose changes and illumination variations between two disjoint camera views. Images from Camera A are mostly captured from 0 degree to 90 degree while those from Camera B mostly from 90 degree to 180 degree. All images are normalized to 128×48 pixels. In Fig. 2(a), we show some examples from VIPeR dataset.

CUHK 01 Dataset. CUHK 01 dataset has 971 persons and each person has two images in each of two camera views. It was collected in a campus environment. Camera A captures the frontal view or back view of pedestrians, while the side views are captured in Camera B. Compared with the VIPeR dataset, images in CUHK 01 dataset are of higher resolution and contains more persons and images. Fig. 2(b) shows some examples from CUHK 01 dataset. Because of significant viewpoints changes, it is also a challenging person re-identification dataset. Images are resized to 160×60 pixels for evaluation.

4.2 Evaluation Details

Training/testing samples. In our experiments, the final average results are reported in form of Cumulated Matching Characteristic (CMC) curve (Wang et al. 2007). The training set is formed from 50% of randomly chosen image pairs and the remaining 50% image pairs are used for testing. In testing, images from one camera are treated as probe and those

[2] A weighted one may improve the final results (Schapire and Singer 1999). We do not test it in this paper.

from the other camera as gallery. Then, we switch the probe and gallery. The average result of all probe-gallery combinations is regarded as one trial to form the CMC curve.

To compare the state-of-the art results, we report the results on VIPeR and CUHK 01 datasets based on the same 10 trials of training / testing samples. Note that for CUHK 01 dataset (485 persons for training while the remaining 486 persons for testing), there are two images for each camera view. During both training and testing, we randomly select an image and extract its corresponding feature.

Mask. In experiments, when we need to separate the foreground from the background, we use the masks on VIPeR dataset provided by (Yang et al. 2014b). On CUHK 01 dataset, we use the method in (Luo, Wang, and Tang 2013) to automatically generate the masks.

Color space. As suggested in (Yang et al. 2014b), we also employ 4 color spaces including RGB, rgb, $l_1 l_2 l_3$ and HSV to make our features robustness to illumination.

Similarity learning. We utilize a fast yet effective metric learning method - KISSME (Köstinger et al. 2012) to compute the similarity score of (5). It is defined in (6):

$$d_M(\vec{x}, \vec{y}) = (\vec{x} - \vec{y})^T M (\vec{x} - \vec{y}). \tag{6}$$

where $\vec{x}$ and $\vec{y}$ are features of a pair of images and M is computed by (7).

$$M = \Sigma_{\mathcal{D}}^{-1} - \Sigma_{\mathcal{S}}^{-1}. \tag{7}$$

where $\Sigma_{\mathcal{S}}$ and $\Sigma_{\mathcal{D}}$ are the covariance matrices (see (Yang et al. 2014b) for details) of similar pairs and dissimilar pairs, respectively.

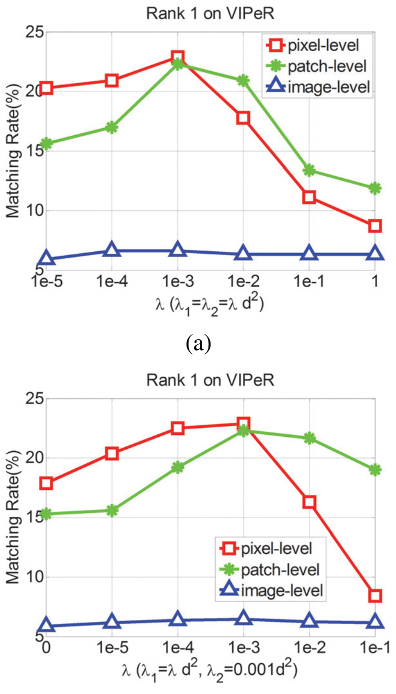
Parameter settings. Unless otherwise specified, we empirically set the parameters in our model as follows: 1) In (1), λ_1 = $0.001d^2$ and λ_2 = $0.001d^2$ with d is the dimension of raw pixel data. 2) The numbers of bases are 350, 70, 60 for learning pixel-level, patch-level and image-level descriptors, respectively. 3) In the patch-level, we adopt 7 × 7 with a stride of 1. 4) When holistic image features are extracted, we use 10 horizontal stripes for images. 5) For all datasets, the dimension of each holistic image feature is reduced to 60 by PCA. Note that the influence of the PCA dimensionality (60-100) is not too critical.

4.3 Performance Analysis

In this subsection, we choose the most widely used VIPeR dataset to evaluate our proposed coding method. We extract the features in RGB color space without using mask.

λ_1 **and** λ_2 **in** (1)**.** To simplify the meta-parameters in WLC, we set λ_1 = λ_2 (=λ_0) in (1), which suggests that the contributions between the second and third terms are identical. Additionally, to avoid tuning λ_0 for different dimensions of raw pixel data, we set λ_0 = λd^2 (d denotes the dimension of raw pixel data). As such, the constraints have stronger effects for higher dimensional data.

In Fig. 3(a), we show the matching rates of WLC with different values of λ. It shows that 1) when λ is set to 0.001, WLC achieves the best result across all levels; 2) pixel-level descriptor performs better than others and 3) the image-level descriptor using the whole image without partitioning into horizontal stripes consistently, showing the importance of partitioning.



(a)

(b)

Figure 3: Parameter analysis (on VIPeR dataset): (a) effect of λ and (b) effect of similarity constraint when λ_2 is set to $0.001d^2$.

We also show the effect of the similarity constraint (controlled by λ_1) in Fig. 3(b). To analyze its performance efficiently, we fix the parameter λ_2 unchanged (set to $0.001d^2$). We can find that when λ_1 is increased from 0 to $0.001d^2$, the performances of WLC are continually increasing for all levels. This phenomenon demonstrates the effectiveness of the similarity constraint.

Comparison with LLC and SC under different number of bases. In this subsection, we compare the performance of WLC with that of LLC and SC in learning multi-level descriptors. For a fair comparison, we use the same input features and bases for all three methods. Different numbers of bases are also evaluated. Specifically, in the pixel-level, we choose 300, 350, 400, 450 and 500. In the patch-level, we use 60, 70, 80, 90 and 100. In the image-level, we employ 60, 80, 100, 150 and 200. As is suggested in (Wang et al. 2010) and (Huang et al. 2011), for both LLC and SC, we set the number of nearest neighbors to 5.

In Tables 3, 4 and 5, we show the performance comparisons of different coding methods in the pixel-level, patch-level and image-level, respectively. As these results show, WLC performs better than both LLC and SC in extracting all level descriptors. During learning the pixel-level and patch-level descriptors, LLC outperforms SC while both LLC and SC fail when learning the image-level descriptors. We know that in SC, the coding coefficient is only based on the rela-

Method	**300**	**350**	**400**	**450**	**500**
LLC (%)	12.4	12.5	12.6	12.4	11.6
SC (%)	5.6	5.7	5.9	5.7	5.6
WLC (%)	**22.1**	**22.9**	**22.8**	**22.1**	**22.2**

Table 3: Comparison with SC and LLC (at Rank 1, on VIPeR dataset) in learning the pixel-level descriptors based on different numbers of bases.

Method	**60**	**70**	**80**	**90**	**100**
LLC(%)	17.4	18.8	19.0	19.1	19.2
SC(%)	14.6	16.6	16.4	16.4	16.5
WLC(%)	**21.6**	**22.3**	**21.7**	**20.7**	**19.0**

Table 4: Comparison with SC and LLC (at Rank 1, on VIPeR dataset) in learning the patch-level descriptors based on different numbers of bases.

Method	**60**	**80**	**100**	**150**	**200**
LLC(%)	0.3	0.3	0.4	0.4	0.3
SC(%)	0.4	0.2	0.4	0.3	0.2
WLC(%)	**6.6**	**6.5**	**6.6**	**6.4**	**6.6**

Table 5: Comparison with SC and LLC (at Rank 1, on VIPeR dataset) in learning the image-level descriptors based on different numbers of bases.

tionships between the input and several nearest bases while in LLC, the code is computed by minimizing the reconstruction error on several nearest bases. Therefore, these observations reflect that in learning the descriptors, minimizing the reconstruction error is still important.

4.4 Comparison with the State-of-the-art Methods on VIPeR Dataset

In this subsection, we compare our method with the state-of-the-art approaches on VIPeR dataset, including GOG (Matsukawa et al. 2016), LSSL (Yang et al. 2016b), MED_VL (Yang et al. 2016a), CSL (Shen et al. 2015), MetricEn (Paisitkriangkrai, Shen, and van den Hengel 2015), LOMO (Liao et al. 2015) and SCNCD (Yang et al. 2014b). Results on Ranks 1, 5 and 10 are shown in Table 6.

SCNCD represents a person based on 16 color names while based on SCNCD, MED_VL learns higher features for each person. Both of them employ KISSME as the similarity measure method, which is the same as our method. From Table 6, we can find that our method (denoted by 'Ours') performs significantly better than them. This phenomenon demonstrates the superiority of our multi-level person representation method over theirs.

In addition, GOG learns a novel hierarchical Gaussian descriptors to represent persons. When it is combined with XQDA (Liao et al. 2015), GOG achieves 49.7% at Rank 1. Our method performs better than GOG (1.7% higher) and obtains a new state-of-the-art result 51.4% at Rank 1 (combined with KISSME).

Rank	**1**	**5**	**10**	Reference
GOG	49.7%	**79.7%**	88.7%	CVPR16
LSSL	47.8%	77.9%	87.6%	AAAI16
MED_VL	41.1%	71.7%	83.2%	AAAI16
CSL	34.8%	68.7%	82.3%	ICCV15
MetricEn	45.9%	77.5%	**88.9%**	CVPR15
LOMO	40.0%	N/A	80.5%	CVPR15
SCNCD	37.8%	68.5%	81.2%	ECCV14
Ours	**51.4%**	76.4%	84.8%	Proposed

Table 6: Comparison with the state-of-the-art methods on VIPeR dataset.

Rank	**1**	**5**	**10**	Reference
GOG	57.8%	79.1%	**88.7%**	CVPR16
MetricEn	53.4%	76.4%	84.4%	CVPR15
LOMO	49.2%	75.7%	84.2%	CVPR15
Mid-level	34.3%	55.1%	65.0%	CVPR14
Ours	**65.8%**	**81.1%**	85.9%	Proposed

Table 7: Comparison with the state-of-the-art methods on CUHK 01 dataset.

In the experiments, we simply use KISSME as the baseline for a similarity measure. GOG adopt XQDA as the baseline while LSSL is used in (Yang et al. 2016b). Since both LSSL and XQDA show better performances than KISSME (Yang et al. 2016b; Liao et al. 2015), we can achieve better results by employing them to compute the similarity score (we do not test it in this paper).

4.5 Comparison with the State-of-the-art Methods on CUHK 01 Dataset.

In this subsection, we compare our method with the state-of-the-art approaches on CUHK 01 dataset, including GOG (Matsukawa et al. 2016), MetricEn (Paisitkriangkrai, Shen, and van den Hengel 2015), LOMO (Liao et al. 2015) and Mid-level (Zheng et al. 2015). Results on Ranks 1, 5 and 10 are shown in Table 7. Among the previous approaches, GOG achieves the best results at Rank 1. The Rank-1 identification of our method (combined with KISSME) is 65.8% (8.0% higher than GOG).

5 Conclusion

In this paper, we propose a new model to address person re-identification task. The main contribution is a simple and effective coding method (WLC) that we use to construct descriptors from raw pixel data at the pixel, patch and image levels. With these learned descriptors, it is easy to extract different holistic image features. During the stage of matching persons, we fuse these holistic image features at the similarity score level. The experimental results on the publicly available datasets - VIPeR and CUHK 01, demonstrate the effectiveness of our model. In the future work, how to combine the similarity score, instead of an unweighted sum strategy in (5), is still worth studying.

Acknowledgments This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801003), the Chinese National Natural Science Foundation Projects #61473291, #61572501, #61502491, #61572536, #61672521, AuthenMetric R&D Funds.

References

Ahmed, E.; Jones, M.; and Marks, T. K. 2015. An improved deep learning architecture for person re-identification. In *CVPR*, 3908–3916.

Ahonen, T.; Hadid, A.; and Pietikäinen, M. 2006. Face description with local binary patterns: Application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28(12):2037–2041.

Coates, A., and Ng, A. Y. 2011. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*.

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

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*.

GuangrunWang; Lin., L.; Ding, S.; Li, Y.; and QingWang. 2016. Dari: Distance metric and representation integration for person verification. In *AAAI*.

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

Karanam, S.; Li, Y.; and Radke, R. J. 2015. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*.

Köstinger, 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. 2013. Color invariants for person reidentification. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35(7):1622–1634.

Layne, R.; Hospedales, T.; and Gong, S. 2012. Person re-identification by attributes. In *BMVC*.

Li, W., and Wang, X. 2013. Locally aligned feature transforms across views. In *CVPR*.

Li, C.; Liu, Q.; Liu, J.; and Lu, H. 2012. Learning ordinal discriminative features for age estimation. In *CVPR*, 2570–2577.

Li, C.; Liu, Q.; Liu, J.; and Lu, H. 2015. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Neural Networks and leanring systems* 26(7):1551–1559.

Li, W.; Zhao, R.; and Wang, X. 2012. Human reidentification with transferred metric learning. In *ACCV*.

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

Liu, X.; Song, M.; Tao, D.; Zhou, X.; Chen, C.; and Bu, J. 2014. Semi-supervised coupled dictionary learning for person re-identification. In *CVPR*.

Luo, P.; Wang, X.; and Tang, X. 2013. Pedestrian parsing via deep decompositional network. In *ICCV*.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*.

Paisitkriangkrai, S.; Shen, C.; and van den Hengel, A. 2015. Learning to rank in person re-identification with metric ensembles. In *CVPR*.

Schapire, R. E., and Singer, Y. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning* 37:297–336.

Shen, Y.; Lin, W.; Yan, J.; Xu, M.; Wu, J.; and Wang, J. 2015. Person re-identification with correspondence structure learning. In *ICCV*.

Shi, Z.; Hospedales, T. M.; and Xiang, T. 2015. Transferring a semantic representation for person re-identification and search. 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*.

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*.

Yang, Y.; Lei, Z.; Zhang, S.; Shi, H.; and Li, S. Z. 2016a. Metric embedded discriminative vocabulary learning for high-level person representation. In *AAAI*.

Yang, Y.; Liao, S.; Lei, Z.; and Li, S. Z. 2016b. Large scale similarity learning using similar pairs for person verification. In *AAAI*.

Yang, Z.; Jin, L.; and Tao, D. 2012. A comparative study of several feature extraction methods for person re-identification. *Biometric Recognition* 7701:286–277.

Zhao, R.; Ouyang, W.; and Wang, X. 2013. Unsupervised salience learning for person re-identification. In *CVPR*.

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

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zhu, J.; Liao, S.; Yi, D.; Lei, Z.; and Li, S. Z. 2014. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *ICB*, 535–540.