Hindawi
Advances in Multimedia
Volume 2017, Article ID 9874345, 12 pages
https://doi.org/10.1155/2017/9874345

Research Article

Deep Learning for Person Reidentification Using Support Vector Machines

Mengyu Xu, Zhenmin Tang, Yazhou Yao, Lingxiang Yao, Huafeng Liu, and Jingsong Xu

Nanjing University of Science and Technology, Nanjing 210094, China

Correspondence should be addressed to Mengyu Xu; mengyxu@hotmail.com

Received 25 May 2017; Revised 20 July 2017; Accepted 23 August 2017; Published 10 October 2017

Academic Editor: Chong Wah Ngo

Due to the variations of viewpoint, pose, and illumination, a given individual may appear considerably different across different camera views. Tracking individuals across camera networks with no overlapping fields is still a challenging problem. Previous works mainly focus on feature representation and metric learning individually which tend to have a suboptimal solution. To address this issue, in this work, we propose a novel framework to do the feature representation learning and metric learning jointly. Different from previous works, we represent the pairs of pedestrian images as new resized input and use linear Support Vector Machine to replace softmax activation function for similarity learning. Particularly, dropout and data augmentation techniques are also employed in this model to prevent the network from overfitting. Extensive experiments on two publically available datasets VIPeR and CUHK01 demonstrate the effectiveness of our proposed approach.

1. Introduction

With the advances in computer version [1–4], machine learning [5–8], and deep neural networks [9, 10], we enter into an era that it is possible to build a real world identification system. Person reidentification (Re-ID) problem aims to recognize individuals across cameras at different locations and time from a distributed multicamera surveillance system in large public spaces [11]. Given a probe image captured from one camera, a person reidentification surveillance system attempts to identify the person from a gallery of candidate images taken from a different camera. The same person can be observed differently in cross-view cameras (see Figure 1). So it is quite difficult to find a kind of feature which is reliable and distinct and directly adapt to changes and misalignment in cross-view condition. Because of these challenge issues, researches in person reidentification still mainly focus on people appearance features, with the acceptable assumption that people will not change their clothing during the whole monitoring period.

Existing methods on this research topic have primarily focused on two aspects. The first aspect is to extract robust and discriminative feature descriptors to identify persons. It has been indicated that three important cues for person reidentification are color information, texture descriptors, and interest points; some of these features are learned from datasets and others are designed by hand. Low-level features such as biologically inspired features (BIF) [12], color histograms and variants [13–17], local binary patterns (LBP) [13, 14, 17, 18], Gabor features [14], and interest points (color SIFT [19, 20] and SURF [21]) were proposed to represent appearance features of different people from nonoverlapping cameras. Some other works have also investigated combinations of multiple visual features, including [13, 14, 16]. The second aspect is to develop metric learning methods to learn discriminative models. The idea of metric learning is to design classifiers to enforce features from the same person to be closer than those from different individuals. Usually used metric learning methods such as Large Margin Nearest Neighbour (LMNN) [16], Logistic Discriminant Metric Learning (LDML) [22], KISSME [18], and Marginal Fisher Analysis (MFA) [16] performed well in solving challenging issues. These approaches typically extract handcrafted features and subsequently learn the metrics. However, these methods optimize feature extraction and metric learning separately or sequentially which leads to suboptimal solutions easily.

In recent years, with the wide use of convolutional neural networks (CNN) in the tasks of object recognition,

(a) CUHK01 (b) VIPeR

Figure 1: Samples of pedestrian observed from CUHK01 and VIPeR datasets. The same person's appearance change in different camera views.

tracking [23], classification [24], and face recognition [25], it has been proved to have a strong automatic learning ability. However, CNN has little progress in person reidentification. In this paper, inspired by the outstanding performance on person reidentification and facial expression recognition in [26, 27], we introduce a deep learning architecture with joint representation learning and linear SVM top layer of CNN to measure the similarity of the comparing image pairs. We randomly select two pedestrian images and horizontally join them as a new resized input image. Joint representation learning method which refers to [26] reduces the complexity of the network rather than two input branches used in Siamese network. We replace the standard softmax layer with L2-SVM to measure the distance of pedestrians in different cameras and estimate whether the inputs of the two pedestrians are the same or not. Compared with softmax function for predicting class labels, we use linear SVM to measure the distance to the decision boundary that is more suitable for the person reidentification which is solved as ranking-like comparison issue. Since L1-SVM is not differentiable, we introduce L2-SVM which is differentiable during function optimization and more stable in numerical computation. Pretrained and dropout techniques are also used in the model to prevent the overfitting problem and boost the performance of person reidentification. The major contributions of this paper are twofold:

(i) We present a deep learning network combined joint representation learning with linear SVM to increase discriminative power of CNN network.

(ii) Extensive experiments are conducted on two benchmark datasets to validate the effectiveness of our architecture and achieve the best results.

2. Related Work

The typical workflow of existing person reidentification system is shown in Figure 2. It indicates that most of them focus on two main components: feature representation and metric learning. The aim of feature representation is to develop discriminate and robust appearance of the same pedestrian across different camera views.

Global features are divided into two categories: color based and texture based features. HSV [28] and LAB [29]

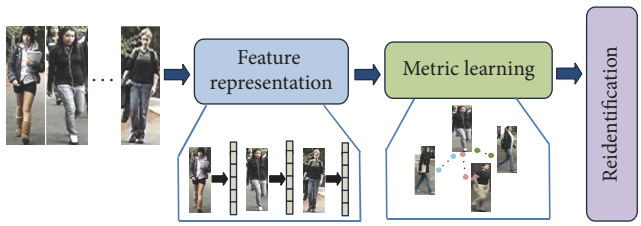

Figure 2: The general procedure of person reidentification.

color histograms are normal color based features. LBP histogram [30] and Gabor filter [14] are used to describe the textures of images. Recently, based on these traditional color and texture features, some more distinct and reliable feature representations for pedestrians have been proposed. Bazzani et al. [31] proposed to use a global mean color histogram and recurrent local patterns through local epitomic analysis to represent a person which is called the histogram plus epitome (HPE). Farenzena et al. [28] proposed to combine weighted HSV histogram of two separated human bodies with salient texture and stable color region as famous symmetry-driven method (SDALF) approach. Yang et al. [32] developed the semantic Salient Color Names based Color Descriptor (SCNCD) employing color naming. Local maximal occurrence (LOMO) features [33] and Scale Invariant Local Ternary Pattern (SILTP) histograms are used to analyse the horizontal occurrence of local features and maximize the occurrence to describe the mean information of pixel features. However, handcrafted features are difficult to achieve the balance between discriminative power and robustness which are highly susceptible to cross-view variations caused by illumination, occlusions, background clutter, and view orientation variations.

Besides feature representation, metric learning is also widely applied for person reidentification. Metric learning is formulated to learn the optimal similarity from features of training images which have strong interclass differences and intraclass similarities. Xiong et al. [34] proposed regularized PCCA (rPCCA), kernel LFDA (kLFDA), and Marginal Fisher Analysis (MFA) when the data space is undersampled. Chopra et al. proposed an algorithm to learn a similarity metric from data [35]. Zheng et al. [36] introduced the

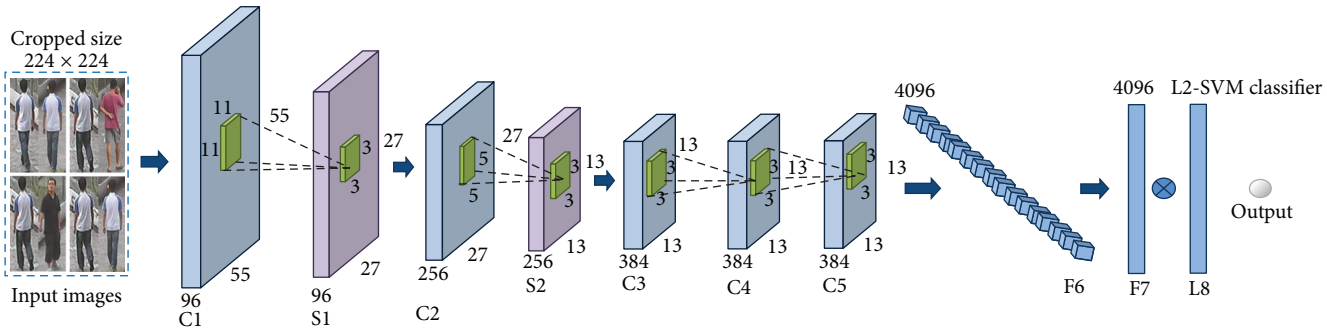

FIGURE 3: The framework of our proposed model. Both of positive and negative pairs are randomly selected as input images. The first to fifth layers are convolution layers and subsampling layers with Relu activation. Sixth and seventh layers are fully connected layers with 4096 neural units. The top layer is linear L2-SVM layer instead of traditional softmax layer to measure the similarity of input images.

Probabilistic Relative Distance Comparison (PRDC) model which aims to maximize the probability of a pair of right match having a smaller distance than that of a wrong match pair and optimizes the relative distance comparison. Prosser et al. [37] reformulated the person reidentification problem as a ranking problem and proposed the Ensemble RankSVM model learning a subspace where the potential true match is given highest ranking rather than any direct distance measure.

Recently, deep learning has become one of the state-of-the-art recognition algorithms, especially that CNN has shown great potential in computer vision tasks. Li et al. [38] propose a new filter pairing neural network (FPNN) that jointly optimizes feature learning, misalignment, occlusions, classification, photometric transforms, and geometric transforms to learn filter pairs encoding photometric transforms. Different from FPNN learning the joint representation of two images, Yi et al. [39] proposed Deep Metric Learning (DML) model inspired by a Siamese neural network that combines the separate modules together, learning the color feature, texture feature, and metric in a unified framework. Matsukawa and Suzuki [40] conducted a fine-tuning of CNN features on a pedestrian attribute dataset to bridge the gap of ImageNet classification and person reidentification and proposed a loss function for classifying combination attributes to increase discriminative power of CNN features. Ahmed et al. [41] presented a deep convolutional architecture with cross-input neighbourhood differences layer and subsequent layer that capture local relationships between the two input images based on mid-level features from each input image and summarized these differences.

3. Algorithm

In the person reidentification task, it usually needs to measure the similarity between gallery set and probe set. CNN is exactly proved to outperform on classification problems rather than comparison problems. Directly using CNN in person reidentification is not suitable and it is hard to leverage its power. In this section, we describe the proposed architecture of CNN specifically. Details of layers and the strategies we used in network training are introduced in the following subsections.

3.1. Joint Representation Learning. The standard pipeline of person reidentification includes feature extracting from input images and metric learning for those features across images. As mentioned above, optimizing feature representation and metric learning separately or sequentially easily leads to sub-optimal solutions. Different from this ordinary framework of learning metric over handcrafted features, we develop to use joint representation learning on input images in our network which is similar to deep rank CNN proposed by Chen et al. [26].

Motived by human assessment, it is used to assess two images whether they belong to the same person by comparing their depicted appearance separately. For instance, pictures A, B, and C are three quite similar but different pedestrian images. Setting picture C as probe image, the discriminative region between A and C is a handbag that appeared in C. Compared with B, pedestrian A wears dress, while B wears pants. As we compare different pedestrian images separately, some value information will be ignored or hidden when appearance features are extracted independently. In our proposed model, jointly representing two input pedestrian images and generating discriminative information will instead separately input images with two branches.

3.2. Architecture. Our deep learning network (see Figure 3) is composed of five convolutional layers (C1, C2, C3, C4, and C5) to extract features, three subsampling layers (S1, S2, and S5), and two fully connected layers (F6, F7). One branch is used as the input of network instead of two branches used in [27]. Different from the architecture of network proposed in [26], the top layer of our network ($L8$) is linear SVM instead of ranking layer which is more discriminative for different pedestrians, and we also optimize the gradient backpropagating learning problem in linear SVM. Randomly given two pedestrian images $\mathbf{I}$ and $\mathbf{J}$ observed from two cross-view cameras with three color channels (RGB) and sized $H_i \times W_i$ ($H_i = 2W_i$), then we join them horizontally. Since pedestrian images are not square-shaped and all of them are quite small, both of the images are resized to 12×256 in the experiment, and the new joint image is square with size of 256×256; then a 224×224 random crop is presented as the input to the whole network in order to get center areas

Table 1: The layer parameters of our network. The output dimension in the table is given by height × width × width. All convolution and fully connected layers use Relu activation function.

Name	Output dim	Filter size	Stride
C1	55 × 55 × 96	11 × 11	4
S1	27 × 27 × 96	3 × 3	2
C2	27 × 27 × 256	5 × 5	1
S2	13 × 13 × 256	3 × 3	2
C3	13 × 13 × 384	3 × 3	1
C4	13 × 13 × 384	3 × 3	1
C5	13 × 13 × 256	3 × 3	1
S5	6 × 6 × 256	3 × 3	2
F6	—	4096	—
F7	—	4096	—
L8	2	—	—

of images we focus on. Processed by this method, the aspect of images remains nearly unchanged and it avoids a large number of parameters contained in Siamese network. The processed images are represented as $x_i, i = 1, 2, \ldots, N$.

The first convolutional layer (C1) is convolved with 96 different filters (see Table 1) of size 11 × 11 with a stride of 4 in each horizontal and vertical directions. Then the 96 various 55 × 55 feature maps are passed through the Relu layer and subsampling layer (S1) with size of 3 × 3 to reduce the maps into 27 × 27 size. The Batch Normalization (BN) layer is employed before each of the Relu layers which allows the network to use much higher learning rates and less focus on initialization of weights and biases. The feature maps are more robust to illumination and variations. If we use K filters and each filter is in size of $m \times m \times C$, the output consists of C' channels of height H_i' and width W_i'. The convolution operation is expressed as function x_i^l:

$$x_i^l = \sigma\left(b_i^{(l)} + \sum_j k_{ij}^{(l)} * x_j^{(l-1)}\right), \tag{1}$$

where $x_i^{(l)}$ and $x_j^{(l-1)}$ represent the ith output channel at the lth layer and the jth input channel at the $(l-1)$th layer; $k_{ij}^{(l)}$ denotes convolutional kernel between the ith and jth feature map. The function $\sigma(\cdot)$ is the Relu neuron activation function of the network and represented as $\sigma(x) = \max(x, 0)$. The max-pooling operation is formulated as

$$x_{(i,j)}^{(l)} = \max_{\forall(p,q)\in\Omega_{(i,j)}} x_{(p,q)}^{(l)}, \tag{2}$$

where $\Omega_{(i,j)}$ represents the pooling region with index (i, j).

The second convolutional layer (C2) takes the outputs of S1 as input with filters of size 5 × 5 and gives 256 different 27 × 27 feature maps. The third and fourth convolutional layers (C3 and C4) are both with filters of size 3 × 3 and give 384 different 13×13 feature maps. With the same size of filters in C3 and C4, the fifth convolutional layer (C5) provides 256 different 13 × 13 feature maps. The two subsampling layers (S2 and S5) repeat the same pooling options as S1. The sixth and seventh fully connected layers (F6 and F7) connect with 6 × 6 × 256 neurons from S5 layer and reduce to 4096 nodes and form compact and robust features. The fully connected layers are expressed as

$$x^{(l)} = w^{(l)} \cdot x^{(l-1)} + b^{(l)}. \tag{3}$$

Instead of traditional softmax layer used in multiple classifications, we use L2-SVM objective for learning the lower level parameters in the top layer (L8) of the whole network to find the max margin of true match (+1) and false match (−1) over training sample pairs.

3.3. Linear SVM versus Softmax

3.3.1. Softmax. Softmax is usually used in deep learning technique at top layer of the network. It is a generalization of logistic regression to the case in multiclass classification. The class labels are formulated as $y^{(i)} \in \{1, \ldots, K\}$, where K is the number of classes. Let h_k be the activation in penultimate layer and let W_k be the weight connecting between penultimate layer and softmax layer. The input to softmax is represented as

$$a_i = \sum_k h_k W_{ki}. \tag{4}$$

The probability is defined as

$$p_i = \frac{\exp(a_i)}{\sum_j^{10} \exp(a_j)}. \tag{5}$$

So the predicted class label $\hat{i}$ would be

$$\hat{i} = \arg\max_i p_i. \tag{6}$$

3.3.2. Linear Support Vector. Softmax is usually used as activation function which is focused on classification and less suitable for ranking-like comparison issue of person reidentification. So in this paper, we proposed to use L2-SVM objection training CNN instead of softmax layer. In linear Support Vector Machines (SVM), corresponding data and labels are represented as (x_n, y_n), $x_n \in \mathbb{R}^D$, $t_n \in \{-1, +1\}$, $n = 1, \ldots, N$, and the linear SVM is defined as the following constrained optimization:

$$\min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max\left(1 - w^T x_n t_n, 0\right). \tag{7}$$

Equation (7) is known as typical L1-SVM, and a differentiable representation is known as L2-SVM, given as follows:

$$l(w) = \min_w \frac{1}{2} w^T w + C \sum_{n=1}^{N} \max\left(1 - w^T x_n t_n, 0\right)^2. \tag{8}$$

L2-SVM is differentiable during optimization and imposes a bigger loss for points which violate the margin. Equation (9) shows the predicted class labels of probe sets

$$\arg\max_t \left(w^T x\right) t. \tag{9}$$

We use the L2-SVM as objective function in our deep network and backpropagate the gradients from linear SVM layer to learn parameters of network. Therefore, the partial derivative of weight w is formulated as

$$\frac{\partial l(w)}{\partial w} = w - 2Cx_n t_n \left(\max\left(1 - w^T x_n t_n, 0\right)\right). \quad (10)$$

The penultimate activation h is given as

$$\frac{\partial l(w)}{\partial h} = -2Ct_n w \left(\max\left(1 - w^T h_n t_n, 0\right)\right). \quad (11)$$

In this way, a joint representation based L2-SVM neural network is obtained and the following section will show its performance on two public datasets.

3.4. Training Strategies Used in CNN

Dropout. During the training, random dropping units which are along with their connection from the neural network is an efficient technique to prevent overfitting and approximately combine exponentially different network architectures efficiently. The dropout technique is usually performed during supervised training and the network is likely forced to learn an averaging model. In this paper, we use dropout in the two fully connected layers (F6, F7) and randomly drop out 50% neurons of these two layers.

Data Augmentation and Data Balancing. Data augmentation is a widely used trick in deep learning. Since neural networks need to be trained on a huge number of training images to achieve satisfactory performance, the public datasets used in person reidentification usually contain limited images. In the training set, the positive pairs (the matched sample pairs) are generally fewer than negative pairs (nonmatched sample pairs). So in the experiment, doing data augmentation is better to boost the performance when training the deep network. In the training set, we randomly crop the input images into 224×224 patches and horizontally flip them around the y-axis. These augmented data will be used as new input of our network. To achieve data balancing, we online sample the same number of positive pairs and negative pairs with a 1 : 1 positive-negative ratio in each minibatch size of 32 images at the very beginning of the training process. As the whole network achieves a reasonably good configuration after the initial training, the positive-negative ratio will gradually reach 1 : 5 to alleviate overfitting.

Stochastic Gradient Descent. Our model is trained using minibatch stochastic gradient descent (SGD) for faster backpropagation and smoother convergence. In each iteration of the training phase, 32 images of a minibatch are the input of the network. We use the SGD with a momentum of 0.9, the learning rate of $\gamma = 10^{-4}$, and weight decay of 0.0005. Note that for every 10000 iterations the learning rate will decrease by $\gamma_{\text{new}} = 0.1 * \gamma$.

Pretraining and Fine-Tuning. The network proposed in this paper is a great depth network, so a great number of labeled images are needed to train the whole network. Before training on VIPeR and CUHK01 datasets, we use CUHK02 datasets to learn a pretrained model. When we test on different datasets, we fine-tune a few top layers of pretrained model with a small learning rate.

Table 2: Comparison of state-of-the-art results of feature representation reported with VIPeR database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	VIPeR ($p = 316$)			
	Top 1	Top 5	Top 10	Top 20
ELF6	8.73	18.76	23.75	31.75
gBiCov	9.87	27.64	36.75	48.96
HSV_Lab_LBP	12.47	26.95	33.37	44.16
Ours	*34.15*	*67.86*	*80.95*	*90.63*

Table 3: Comparison of state-of-the-art results of feature representation reported with CUHK01 database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	CUHK01 ($p = 485$)			
	Top 1	Top 5	Top 10	Top 20
ELF18	5.37	13.45	17.28	23.45
gBiCov	7.25	13.75	18.64	24.26
LOMO	10.80	23.20	27.35	36.12
Ours	*50.01*	*64.75*	*73.85*	*84.96*

Table 4: Comparison of state-of-the-art results of metric learning reported with VIPeR database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	VIPeR ($p = 316$)			
	Top 1	Top 5	Top 10	Top 20
LMNN	6.23	19.65	32.63	52.25
ITML	12.4	27.5	39.7	55.2
Euclidean	14.46	28.75	39.14	50.10
RDC	15.7	32.5	53.9	70.1
KISSME	25.78	56.24	70.14	82.92
Ours	*34.15*	*67.86*	*80.95*	*90.63*

4. Experiments

Our proposed network is implemented by Theano deep learning framework. The network is trained in NVIDIA TITAN X. We evaluate the proposed method on several famous person reidentification datasets carried out to compare with state-of-the-art approaches. The results are shown in Cumulative Matching Characteristic (CMC) curve. The cumulative matching scores are also shown in Tables 2–9.

4.1. Datasets and Evaluation Protocol

Datasets. We evaluate our method on two public datasets: VIPeR dataset and CUHK01 dataset. The deep learning model is pretrained on CUHK02 dataset. VIPeR dataset is a relatively small and quite challenging dataset in person

Table 5: Comparison of state-of-the-art results of metric learning reported with CUHK01 database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	CUHK01 ($p = 485$)			
	Top 1	Top 5	Top 10	Top 20
Euclidean	10.52	28.07	39.94	55.07
LMNN	13.45	31.33	42.52	54.11
ITML	16.0	28.5	45.3	60.1
KISSME	29.40	57.67	72.43	86.07
Ours	*50.01*	*64.75*	*73.85*	*84.96*

Table 6: Comparison of some other state-of-the-art results reported with VIPeR database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	VIPeR ($p = 316$)			
	Top 1	Top 5	Top 10	Top 20
L2-norm	10.89	22.37	32.34	45.19
L1-norm	12.15	26.01	32.09	34.72
aPRDC	16.14	37.72	50.98	65.95
RankSVM	14.00	37.00	51.00	67.00
SSCDL	25.60	54.15	68.10	83.60
eSCD	26.31	46.61	58.86	72.77
PCCA	19.62	51.55	68.23	82.92
rPCCA	21.96	54.78	70.95	85.29
SVMML	30.07	63.17	77.44	88.08
MFA	32.24	65.99	79.66	90.64
KLFDA	32.33	65.78	79.72	90.95
Ours	*34.15*	*67.86*	*80.95*	*90.63*

Table 7: Comparison of some other state-of-the-art results reported with CUHK01 database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	CUHK01 ($p = 485$)			
	Top 1	Top 5	Top 10	Top 20
L2-norm	5.6	16.0	22.9	30.6
SDALF	9.90	22.57	30.33	41.03
L1-norm	10.8	15.5	37.6	35.6
SVMML	30.23	55.58	67.49	78.92
KLFDA	32.76	59.01	69.63	79.18
MFA	38.09	56.34	64.59	72.62
Ours	*50.01*	*64.75*	*73.85*	*84.96*

Table 8: Comparison of CNN-based algorithms results reported with VIPeR database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	VIPeR ($p = 316$)			
	Top 1	Top 5	Top 10	Top 20
Deep_CNN	12.5	21.2	26.3	39.7
ImageNet + XQDA	19.7	44.5	58.1	72.9
DML	28.23	59.27	73.45	86.39
Ours	*34.15*	*67.86*	*80.95*	*90.63*

Table 9: Comparison of CNN-based algorithms results reported with CUHK01 database. The cumulative matching scores (%) at ranks 1, 5, 10, and 20 are listed.

Method	CUHK01 ($p = 485$)			
	Top 1	Top 5	Top 10	Top 20
FPNN	27.87	58.20	73.46	86.31
ImageNet + XQDA	28.5	52.3	63.6	74.9
FFN + XQDA	32.4	55.9	66.5	76.6
Ours	*50.01*	*64.75*	*73.85*	*84.96*

reidentification. It has 632 pedestrian pairs captured by two camera views in outdoor environment. Each pair contains two images of the same person seen from different viewpoints, including Cam A and Cam B. Images in Cam A are mainly from 0 to 90 degrees while images in Cam B are from 90 to 180 degrees. All images are normalized to 128×48.

The CUHK01 dataset is a larger dataset than VIPeR which contains 972 persons captured from two cross-views with 3884 images in a campus environment. Camera view A and camera view B include two images for the same person and view A captures the frontal or back view of the individuals while view B captures the profile view. All images are scaled to 160×60 pixels. The CUHK02 dataset contains five pairs of views (P1-P2). Images from P2-P2 were used to learn a pretrained model.

Evaluation Protocol. In each experiment on different datasets, we randomly divide each dataset into gallery set and probe set. The gallery set is composed of two kinds of image pairs: positive pairs and negative pairs. The positive pairs are created by the same people from different camera views, and the negative pairs are created by two separate people. Specifically, for VIPeR dataset, we set the number of individuals in the gallery/probe sets split to 316/316. For CUHK01 dataset, we use 485 pedestrians for training and 486 for testing. We compare our method with some state-of-the-art methods on VIPeR and CUHK01 datasets. The whole procedure is repeated ten times, and the average of Cumulative Matching Characteristic (CMC) curves are used to evaluate the performance of different approaches.

4.2. Comparison with Feature Representation

4.2.1. Experiments on VIPeR Dataset. In this experiment, we pretrained the network model with CUHK02 dataset and randomly divide the 632 pairs of images into half for training and half for testing. We compare our proposed approach with the following three available and typical person reidentification features: Ensemble of Local Features (ELF) [42], gBiCov [12], and HSV with Lab and LBP feature proposed in [18]. In the experiment, we used ELF6 implemented in [42].

We compared our proposed method with these three different kinds of features, results of CMC curves, and top-ranked matching rates shown in Figure 4(a) and Table 2. From Figure 4(a), it can be observed that our approach gives the best result. Comparing to the three baseline methods, the performance of our approach gains is over 20% at rank-1.

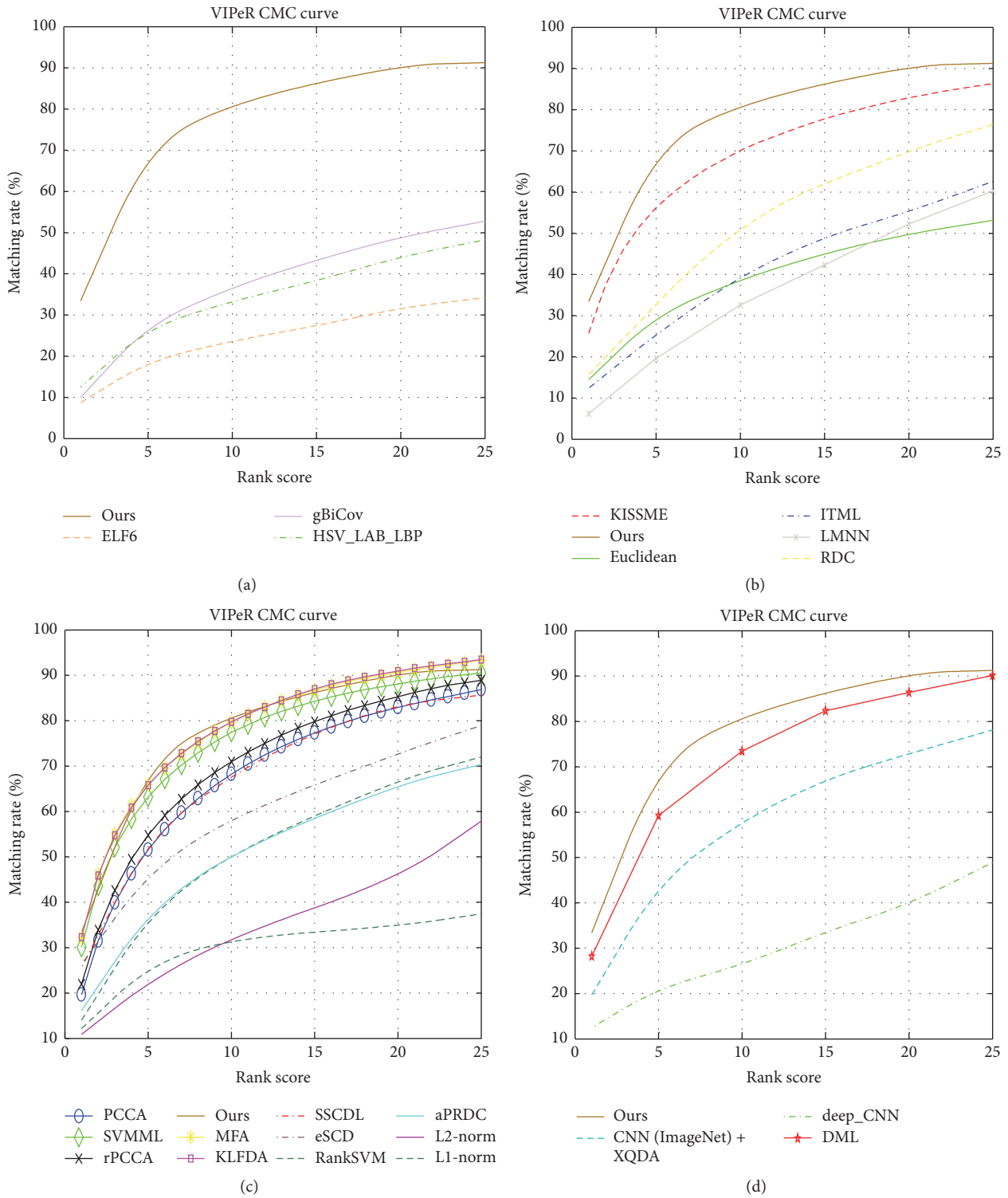

Figure 4: CMC curves on VIPeR data set. (a) Performance comparison with feature representation algorithms; (b) performance comparison with metric learning algorithms; (c) performance comparison with other state-of-the-art algorithms; (d) performance comparison with other CNN-based algorithms.

Such trend grows as the rank number increases. As shown in Table 2, our proposed method achieves a 34.15% rank-1 matching rate outperforming the ELF6 with 8.73%, gBiCov with 9.87%, and HSV_Lab_LBP with 12.47%. In our method, the feature learning is directly performed on the input images avoiding missing the critical information during the feature extracting by using handcrafted features. It confirms that utilizing deep convolutional neural network for learning

feature representation and similarity measurement is an effective solution for solving people reidentification tasks.

4.2.2. Experiments on CUHK01 Dataset. Same as the pretrained strategy for CUHK02 dataset used on VIPeR dataset, we chose the following approaches as baselines: ELF18 [42], gBiCov [12], and Local Maximal Occurrence (LOMO) representation [33]. The ELF18 feature is the same as ELF6 which is computed from eighteen equally divided horizontal stripes histograms rather than six.

The comparison results are shown in Figure 5(a) and Table 3. It is observed that our method outperforms the three feature representation methods by a large margin which is over 40% at all ranks and again validates its effectiveness. It is notable that our method achieves 50.01% rank-1 matching rate, outperforming the gBiCov which achieved a 7.25% rank-1 matching rate, by a more significant sizeable margin than VIPeR. The main reason for its superior performance on CUHK01 is that there are less positive pairs in VIPeR dataset even though we have used data augmentation strategy. It still lacks training data to train a robust network. Compared with VIPeR, CUHK01 is larger in scale and has more training data to feed into the deep network to learn a data-driven optimal framework.

4.3. Comparison with Metric Learning Algorithms

4.3.1. Experiments on VIPeR Dataset. We evaluated the proposed algorithm and several metric learning algorithms, including ITML [43], Euclidean [38], LMNN [16], KISSME [18], and RDC [44]. The results of Cumulative Matching Characteristic (CMC) curves are shown in Figure 4(b). It can be seen that our proposed method is better than the compared metric learning algorithms. To present the quantized comparison results more clearly, we summarize the performance comparison at several top ranks in Table 4. Note that our approach achieves a 34.15% rank-1 matching rate, outperforming the performance of KISSME nearly 10% at all ranks. The main reason for its superior performance is that our proposed framework is capable of joint representation learning and SVM rather than requiring two-step separate optimization.

4.3.2. Experiments on CUHK01 Dataset. We compare our proposed method with the same methods which have been validated on the VIPeR dataset. Figure 5(b) plots the CMC curves and Table 5 shows the ranking results of all methods on the CUHK01. It can be seen that our method outperforms state-of-the-art methods with a rank-1 recognition rate of 50.01% (versus 29.40% by the next best method). Notice that the second best method on this dataset is KISSME. Our method performs best over 1, 5, and 10, whereas KISSME is better at rank-20 and rank-25. Even though KISSME got better performance on rank-20 and rank-25, our proposed method still performs well.

4.4. Comparison with Other State-of-the-Art Algorithms

4.4.1. Experiments on VIPeR Dataset. We compare the performance of our algorithm with the following approaches: KLFDA [34], PCCA [45], rPCCA [34], SVMML [46], MFA [16], SSCDL [47], eSCD [29], RankSVM [37], aPRDC [48], L1-norm [49], and L2-norm. Figure 4(c) and Table 6 show the CMC curves and the matching rate comparing our method with state-of-the-art methods. It is obvious that our method gives the best result among these algorithms which achieves 34.15% rank-1 matching rate, outperforming the result of KLFDA with 32.33%. The other better performing method on the VIPeR dataset is MFA which achieved 32.24% rank-1 matching rate. Our method performs best over ranks 1, 5, and 10, whereas KLFDA and MFA perform better over ranks 15, 20, and 25. The experiment results suggest that even though our model suffers from a severe lack of training data, it still achieves state-of-the-art performance on the highly challenging VIPeR dataset.

4.4.2. Experiments on CUHK01 Dataset. We compare our method with several state-of-the-art approaches on CUHK01 dataset, such as KLFDA [34], SVMML [46], MFA [16], SDALF [29], L1-norm [49], and L2-norm. As shown in Figure 5(c) and Table 7, our method achieves more significant outperformance than KLFDA and MFA in all ranks on the CUHK01 dataset rather than VIPeR. It suggests that the large train dataset will improve the learning ability of CNN network.

Experiment results on both VIPeR and CUHK01 datasets clearly indicate that our proposed CNN method outperforms these feature representation and metric learning algorithms, particularly when sufficient training data are provided. In our proposed method, feature learning is directly performed on the input images. Joint input branch of the lower level layers designed in the framework transforms the input images gradually into the higher-level representation with more refined features without dramatic feature reduction. The linear SVM classifier layer effectively measures the similarity of representations among the people appearances.

4.5. Comparison with CNN-Based Algorithms.

In this section, we compare our method with five types of deep learning based person reidentification algorithms: FPNN [38], ImageNet + XQDA [40], FFN + XQDA [40], Deep_CNN [50], and DML [39]. ImageNet + XQDA algorithm is the combination of ImageNet feature and XQDA metric learning. We compare our method with it on both of VIPeR and CUHK01 datasets. FPNN and FFN + XQDA network model were trained on large-scale CUHK dataset because the other existing datasets are too small to train deep networks. Therefore, we compare our method with these two networks on CUHK01 and with DML on VIPeR dataset. It is notable that the train setting on CUHK01 of FPNN conducted in a different setting, with 871 pedestrians chosen for training and 100 for testing. Figures 4(d) and 5(d) and Tables 8 and 9 show the result of our experiments, and our method still achieves the best performance among these CNN-based approaches. The matching rate of our method on rank-1 outperforms ImageNet + XQDA more than 10% on both of VIPeR and CUHK01 datasets, far surpassing that of FPNN and Deep_CNN, which were only 27.87% and 12.5% separately.

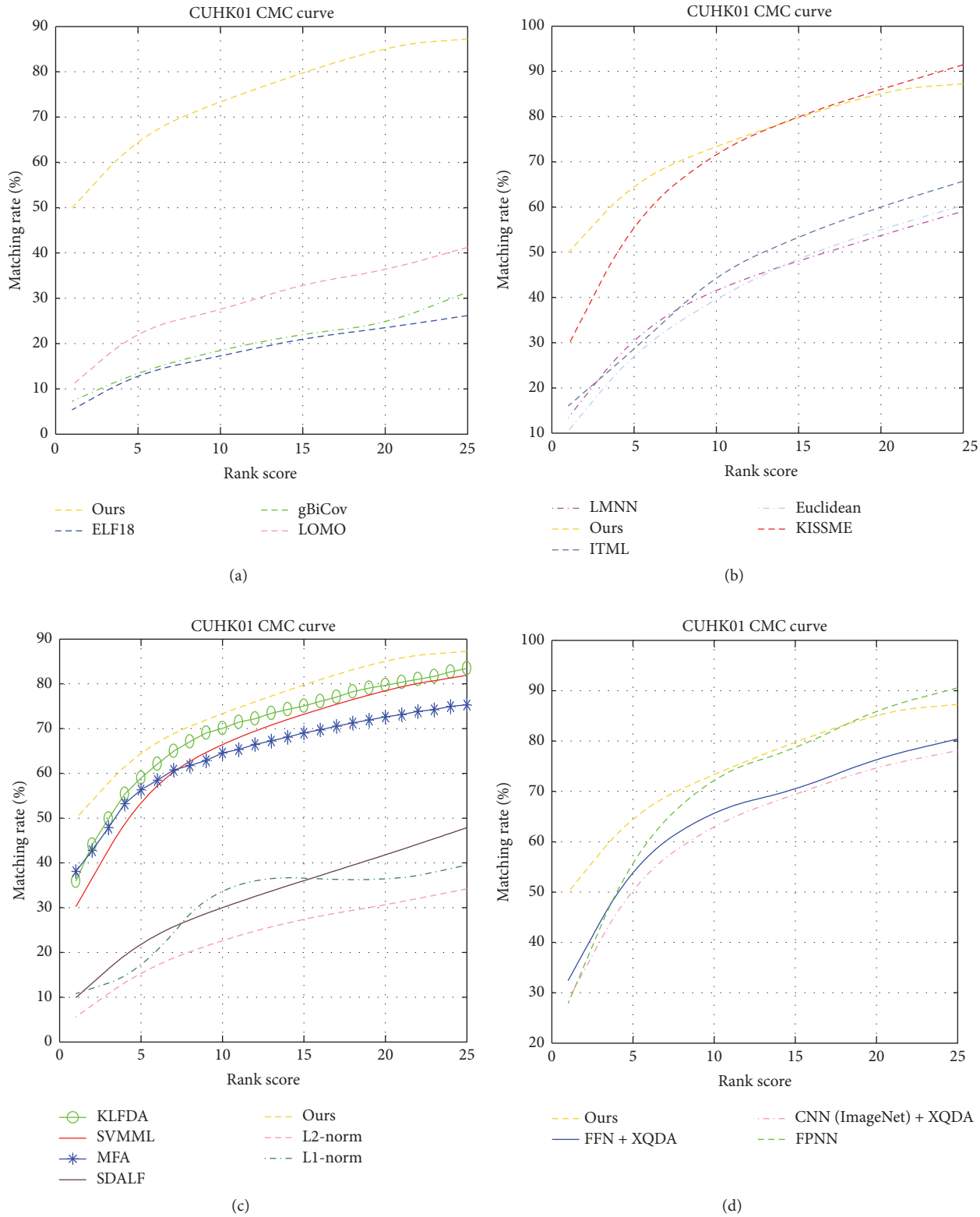

Figure 5: CMC curves on CUHK01 data set. (a) Performance comparison with feature representation algorithms; (b) performance comparison with metric learning algorithms; (c) performance comparison with other state-of-the-art algorithms; (d) performance comparison with other CNN-based algorithms.

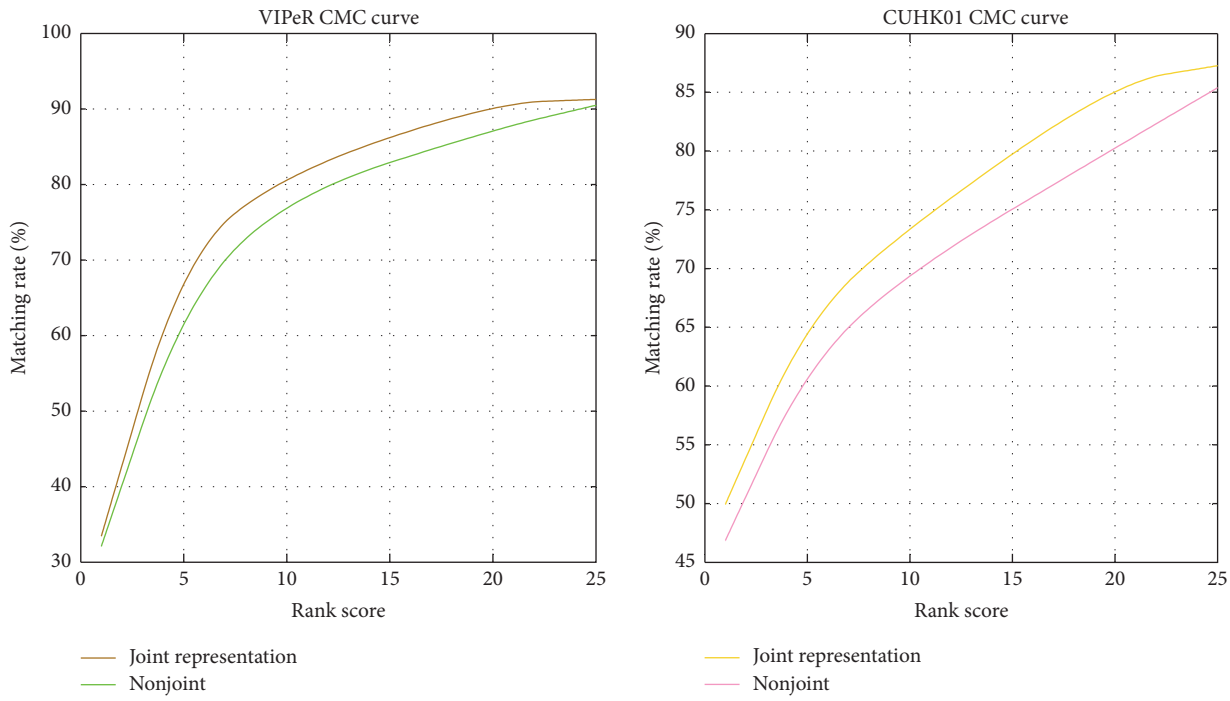

Figure 6: Performance comparison with two branches input method using CMC curves on VIPeR and CUHK01 datasets.

4.6. Superiority of Joint Representation Learning. Many previous works on deep learning of person reidentification share the common input framework that they extract features from two images separately. As mentioned above, joint representation learning is easier to avoid features ignored and hidden when they are extracted independently. To validate the effectiveness of our proposed framework, we compare it with two branches on VIPeR dataset and CUHK01 dataset. The CMC curves in Figure 6 show that joint representation learning method consistently surpasses methods which have two branches, thereby demonstrating the good performance of our method depending on joint representation learning.

4.7. Superiority of Linear SVM Layer. In this paper, we introduce linear SVM to replace the traditional softmax activation function to measure the similarity of the comparing pair. We also perform experiments to evaluate the contribution of our linear SVM layer. We employ a softmax layer to replace the last linear SVM layer with the other layers left unchanged. In this way, the deep network is used to assess whether two input images belonged to the same person. The experiments are conducted on the CUHK01 dataset. The results in Figure 7 show that the linear SVM layer is more suitable for person reidentification problem than softmax layer.

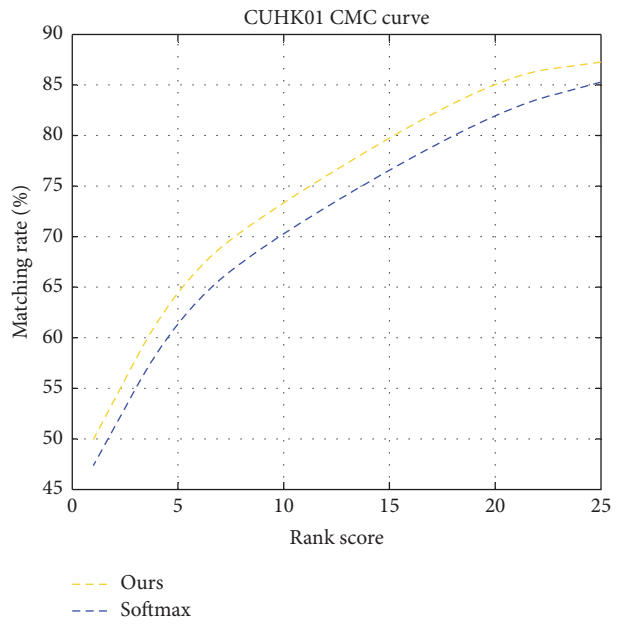

Figure 7: Performance comparison with softmax layer using CMC curves on CUHK01 datasets.

5. Conclusion

In this paper, we present an effective linear Support Vector Machines network based on joint representation for person reidentification. The proposed model introduces L2-SVM to replace traditional softmax layer to deal with rank-like comparison problem. Instead of using the Siamese network to train a pair of input images, we use joint representation learning strategy to avoid designing new network architecture with two entrances. Extensive experiments on two challenging person reidentification datasets (VIPeR and CUHK01) demonstrate the effectiveness of our proposed approach. In the future, we intend to adapt our method on video sequence data and promote the efficiency of reidentification.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Unmanned Equipment Intelligent Control Support Software System (Grant no. 2015ZX01041101).

References

[1] F. Shen, C. Shen, X. Zhou, Y. Yang, and H. T. Shen, "Face image classification by pooling raw features," *Pattern Recognition*, vol. 54, pp. 94–103, 2014.

[2] F. Shen, C. Shen, A. van den Hengel, and Z. Tang, "Approximate least trimmed sum of squares fitting and applications in image analysis," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1836–1847, 2013.

[3] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang, "Exploiting Web Images for Dataset Construction: A Domain Robust Approach," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1771–1784, 2017.

[4] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang, "A new web-supervised method for image dataset constructions," *Neurocomputing*, vol. 236, pp. 23–31, 2017.

[5] F. Shen, X. Zhou, Y. Yang, J. Song, H. T. Shen, and D. Tao, "A fast optimization method for general binary code learning," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5610–5621, 2016.

[6] F. Shen, C. Shen, Q. Shi, A. van den Hengel, Z. Tang, and H. T. Shen, "Hashing on nonlinear manifolds," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1839–1851, 2015.

[7] Y. Yao, X.-S. Hua, F. Shen, J. Zhang, and Z. Tang, "A domain robust approach for image dataset construction," in *Proceedings of the 24th ACM Multimedia Conference, MM 2016*, pp. 212–216, gbr, October 2016.

[8] Y. Yao, J. Zhang, F. Shen, X. Hua, J. Xu, and Z. Tang, "Automatic image dataset construction with multiple textual metadata," in *Proceedings of the 2016 IEEE International Conference on Multimedia and Expo, ICME 2016*, usa, July 2016.

[9] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," *Computer Vision and Pattern Recognition*, pp. 1335–1344, 2016.

[10] L. Ren, J. Lu, J. Feng, and J. Zhou, "Multi-modal uniform deep learning for RGB-D person re-identification," *Pattern Recognition*, vol. 72, pp. 446–457, 2017.

[11] S. Gong, M. Cristani, S. Yan, and C. C. Loy, "Person Re-identification," *Advances in Computer Vision and Pattern Recognition*, 2014.

[12] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.

[13] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8927, pp. 134–146, 2015.

[14] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 3594–3601, usa, June 2013.

[15] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proceedings of the 2012 23rd British Machine Vision Conference, BMVC 2012*, gbr, September 2012.

[16] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Advances in Neural Information Processing Systems*, pp. 1473–1480.

[17] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proceedings of the 2013 14th IEEE International Conference on Computer Vision, ICCV 2013*, pp. 2528–2535, aus, December 2013.

[18] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2288–2295, IEEE, Providence, RI, USA, June 2012.

[19] K. Jüngling, C. Bodensteiner, and M. Arens, "Person re-identification in multi-camera networks," in *Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2011*, usa, June 2011.

[20] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the 2009 20th British Machine Vision Conference, BMVC 2009*, gbr, September 2009.

[21] N. Gheissari, T. B. Sebastian, P. H. Tu, J. Rittscher, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, pp. 1528–1535, usa, June 2006.

[22] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 498–505, Kyoto, Japan, October 2009.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 580–587, Columbus, Ohio, USA, June 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.

[25] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 1988–1996, can, December 2014.

[26] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2353–2367, 2016.

[27] Y. Tang, *Deep Learning Using Support Vector Machines*, CoRR, 2013.

[28] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2360–2367, IEEE, San Francisco, Calif, USA, June 2010.

[29] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3586–3593, IEEE, Portland, Ore, USA, June 2013.

[30] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[31] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," *Pattern Recognition Letters*, vol. 33, no. 7, pp. 898–903, 2012.

[32] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8689, no. 1, pp. 536–551, 2014.

[33] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 2197–2206, usa, June 2015.

[34] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8695, no. 7, pp. 1–16, 2014.

[35] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 539–546, IEEE, Washington, DC, USA, June 2005.

[36] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 649–656, June 2011.

[37] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proceedings of the 2010 21st British Machine Vision Conference, BMVC 2010*, gbr, September 2010.

[38] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 152–159, usa, June 2014.

[39] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proceedings of the 22nd International Conference on Pattern Recognition, ICPR 2014*, pp. 34–39, swe, August 2014.

[40] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2428–2433, Cancun, December 2016.

[41] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 3908–3916, usa, June 2015.

[42] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision—ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12–18, 2008, Proceedings, Part I*, vol. 5302 of *Lecture Notes in Computer Science*, pp. 262–275, Springer, Berlin, Germany, 2008.

[43] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 209–216, June 2007.

[44] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[45] A. Mignon and F. Jurie, "PCCA: a new approach for distance learning from sparse pairwise constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2666–2672, June 2012.

[46] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 3610–3617, usa, June 2013.

[47] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 3550–3557, usa, June 2014.

[48] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: what features are important?" in *Computer Vision—ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7–13, 2012, Proceedings, Part I*, vol. 7583 of *Lecture Notes in Computer Science*, pp. 391–401, Springer, Berlin, Germany, 2012.

[49] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, ICCV*, bra, October 2007.

[50] G. Zhang, J. Kato, Y. Wang, and K. Mase, "People re-identification using deep convolutional neural network," in *Proceedings of the 9th International Conference on Computer Vision Theory and Applications, VISAPP 2014*, pp. 216–223, prt, January 2014.