2014 22nd International Conference on Pattern Recognition

Deep Metric Learning for Person Re-Identification

(Invited Paper)

Dong Yi, Zhen Lei, Shengcai Liao and Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences (CASIA)

***Abstract*—Various hand-crafted features and metric learning methods prevail in the field of person re-identification. Compared to these methods, this paper proposes a more general way that can learn a similarity metric from image pixels directly. By using a "siamese" deep neural network, the proposed method can jointly learn the color feature, texture feature and metric in a unified framework. The network has a symmetry structure with two sub-networks which are connected by a cosine layer. Each sub-network includes two convolutional layers and a full connected layer. To deal with the big variations of person images, binomial deviance is used to evaluate the cost between similarities and labels, which is proved to be robust to outliers. Experiments on VIPeR illustrate the superior performance of our method and a cross database experiment also shows its good generalization.**

I. Introduction

The task of person re-identification is to judge whether two person images belong to the same subject or not. In practical applications, the two images are usually captured by two cameras with disjoint views. The performance of person re-identification is closely related to many other applications, such as cross camera tracking, behaviour analysis, object retrieval and so on. The algorithms proposed in this field are also overlapped with other fields in pattern recognition. In recent years, the performance of person re-identification has increased continuously and will increase further.

The essence of person re-identification is very similar to biometric recognition problems, such as face recognition. The core of them is to find a good representation and a good metric to evaluate the similarities between samples. Compared to biometric problems, person re-identification is more challenging due to the low quality and high variety of person images. Person re-identification usually needs to match the person images captured by surveillance cameras working in wide-angle mode. Therefore, the resolution of person images are low (*e.g.,* around 48×128 pixels) and the lighting conditions are unstable too. Furthermore, the direction of cameras and the pose of persons are arbitrary. These factors cause the person images under surveillance scenarios have two distinctive properties: large variations in intra class, and ambiguities between inter classes.

Since the pixels of person images are unstable, effective representations are important and needed for person re-identification. To this end, existing methods borrow many sophisticated features from other fields, such as HSV histogram, Gabor, HOG and so on. Based on the features, direct matching or discriminative learning are then used to evaluate the similarity. Existing methods mainly focus on the second step that is how to learn a metric to discriminate the persons. Many good metric learning methods have been proposed in this context, such as KISSME [1], RDC [2] and so on.

The majority of existing methods include two separate steps: feature extraction and metric learning. The features always come from two separate sources: color and texture, some of which are designed by hand, some of which are learned, and they are finally connected or fused by simple strategies. On the contrary, this paper proposes a new method to combine the separate modules together that is learning the color feature, texture feature and metric in a unified framework, which is called as "Deep Metric Learning" (DML).

The main idea of DML is inspired by a "siamese" neural network [3], which is originally proposed for signature verification. Given two person images $\mathbf{x}$ and $\mathbf{y}$, we want to use a siamese deep neural network to assess their similarity $s = DML(\mathbf{x}, \mathbf{y})$. Different from the original work [3], our DML does not need the two sub-networks share the same weights and biases. In this way, each sub-network in DML can adapt to its corresponding view, which makes DML more appropriate to person re-identification across views. By using cosine as the last layer, the similarity equation can be written as $s = DML(\mathbf{x}, \mathbf{y}) = Cosine(B_1(\mathbf{x}), B_2(\mathbf{y}))$, where B_1 and B_2 denote the two sub-networks of DML. If we want to construct a generic (don't consider view) deep metric, B_1 and B_2 should share their parameters.

Compared with existing methods, DML has three advantages:

1) DML can learn a similarity metric from image pixels directly. All layers in DML are optimized by the same objective function, which are more effective than the hand-crafted features in traditional methods.
2) The multi-channel filters learned in DML can capture the color and texture information simultaneously, which are more reasonable than the simple fusion strategies in traditional methods, *e.g.,* feature concatenation and sum rule.
3) The structure of DML is flexible that can easily switch between view specific and general person re-identification tasks by whether sharing the parameters of sub-networks.

DML is tested on the most popular person re-identification database, VIPeR [4], using the common evaluation protocol. The results show that DML outperforms most of existing methods and is on a par with the state-of-the-art [5]. To evaluate the generalization of DML, we conduct a cross database experiment, that is training on CUHK Campus [5] and testing on VIPeR. The results of the cross database experiment

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.16

are also better than the newest transfer learning method [6] under similar experimental setting. To our knowledge, this is the first strict cross database experiment in the field of person re-identification. For practical applications, cross database experiment is more significant than traditional experiments.

II. Related Work

This work uses deep learning to learn a metric for person re-identification. Related works in three aspects are reviewed in this section: feature representation, metric learning for person re-identification and siamese convolutional neural network.

Early papers mainly focus on how to construct effective feature representation. From 2005, numerous features are used or proposed for person re-identification [7]. The most popular features include HSV color histogram [8], [5], LAB color histogram [9], SIFT [9], LBP histogram [5], Gabor features [5] and their fusion. Among the features, color has the most contribution to the final results. On the other hand, [8] has proved that using the silhouette and symmetry structure of person can improve the performance significantly, therefore the color and texture features are usually extracted in a predefined grid or finely localized parts. The recent advances in this aspect are color invariant signature [10] and salience matching [11]. According to the history of biometrics research, the future directions of feature representation may be precise body parts segmentation, person alignment and pose normalization.

Based on the extracted features, naive feature matching or unsupervised learning methods have got moderate results, but state-of-the-art are achieved by supervised methods, such as Boosting [12], Rank SVM [13], PLS [13] and Metric learning [2], [1], [5]. In these methods, metric learning is the main stream due to its flexibility. Compared with standard distance measures, *e.g.,* L_1, L_2 norm, the learned metric is more discriminative for the task on hand and more robust to large variations of person images across view. Most papers use a holistic metric to evaluate the similarity of two samples, but [5] first divides the samples into several groups according to their pose and then learn a metric for each group. By using the pose information explicitly, [5] obtains the highest performance.

Early in 1993, a siamese neural network [3] was proposed to evaluate the similarity of two signature samples. In the same year, a neural network [14] with similar structure was proposed for fingerprint verification. Different from traditional neural networks, the siamese architecture is composed by two sub-networks sharing the same parameters. Each sub-network is a convolutional neural network. Then the siamese neural network was used for face verification [15] by the same research group. The best property of siamese neural network is its unified and clear objective function. Guided by the objective function, the end-to-end neural network can learn an optimal metric towards the target automatically. The responsibility of the last layer of the siamese neural network is to evaluate the similarity of the output of two sub-networks, which can be in any form [15], such as L_1, L_2 norm and cosine. [3] used cosine function because of its invariance to the magnitude of samples. Because of the good property of cosine function and it has been used widely in many pattern recognition problems [16], we choose it as the last layer of DML.

Although good experimental results have been obtained in [3], [14] and [15], their disadvantages are lacking implementation details and lacking comparison with other methods. This paper will remove the parameter sharing constraint of the siamese neural network and apply it in the person re-identification problem. In the following sections, the implementation details will be described and the comparisons will be reported.

III. Deep Metric Learning

Affected by various factors, the similarity of two person images is hard to evaluate. Under the joint influence of resolution, illumination and pose changes, the ideal metric for person re-identification maybe highly nonlinear. Deep learning is exact one of the most effective tool to learn the nonlinear metric function.

A. Architecture

For most of pattern recognition problems, neural network works in a standalone mode. The input of neural network is a sample and the output is a predicted label. This mode works well for handwritten digit recognition, object recognition and other classification problems when the labels of the training set are the same as the testing set. For the person re-identification problem, the subjects in the training set are generally different from those in the testing set, therefore the "sample $\rightarrow$ label" style neural network cannot apply to it. To deal with this problem, we construct a siamese neural network, which includes two sub-networks working in a "sample pair $\rightarrow$ label" mode.

The flowchart of our method is shown in Figure 1. Given two person images, they are first separated into three overlapped parts respectively and the image pairs are matched by three siamese convolutional neural network (SCNN). For two image patches $\mathbf{x}$ and $\mathbf{y}$, SCNN can predict a label $l = \pm 1$ to denote whether the image pair comes from the same subject or not. Because many applications need rank the images in the database based on their similarities with a probe image, our SCNN outputs a similarity score instead. The structure of the SCNN is shown in Figure 2, which is composed by two convolutional neural networks (CNN). And the two CNNs are connected by a cosine layer. The similarity of two image patches is calculated by

$$s = \frac{B_1(\mathbf{x})^T B_2(\mathbf{y})}{\sqrt{B_1(\mathbf{x})^T B_1(\mathbf{x})}\sqrt{B_2(\mathbf{y})^T B_2(\mathbf{y})}}, \quad (1)$$

where B_1 and B_2 are the functions of the two CNNs respectively.

Existing siamese neural networks have a constraint, that is their two sub-networks should share the same parameters, *i.e.,* weights and biases. In this paper, we remove this constraint in some conditions. Without parameters sharing, the network can deal with the view specific matching tasks more naturally. With parameters sharing, the network is more appropriate for general task, *e.g.,* cross database person re-identification. We call these two modes as "General" and "View Specific" SCNN.

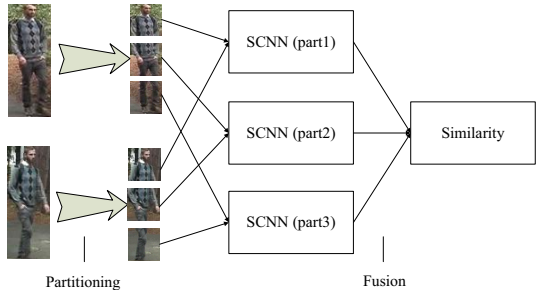

Fig. 1. The flowchart of the proposed method. Learning three SCNNs for each part independently and fusing them by sum rule.

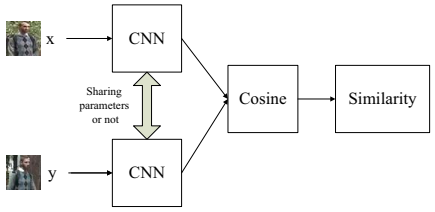

Fig. 2. The structure of the siamese convolutional neural network (SCNN). The SCNN can work in two modes: sharing parameters (General SCNN) and independent parameters (View Specific SCNN)

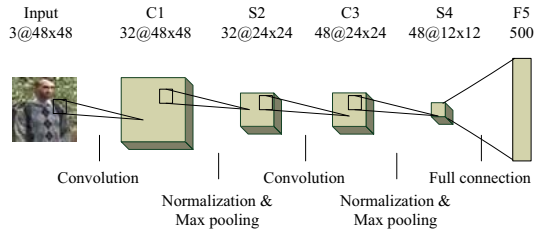

Fig. 3. The structure of the 5-layer CNN used in our method.

B. Convolutional Neural Network

The CNN in this paper (Figure 2) is composed by 2 convolutional layers, 2 max pooling layers and a full connected layer. As shown in Figure 3, the number of channels of convolutional and pooling layers are 32, 32, 48 and 48. The output of the CNN is 500 dimensions. Every pooling layer includes a cross-channel normalization unit. Before convolution the input data are padded by zero values, therefore the output have the same size with input. The filter size of C1 layer is 7×7 and the filter size of C2 layer is 5×5. ReLU neuron [17] is used as activation function for each layer.

C. Cost Function and Learning

Backpropagation (BP) [18] is used to learn the parameters of SCNN. For cost function, we propose three candidates as shown in Figure 4: Square loss, Exponential loss, and Binomial deviance [19]. Given a sample pair's similarity $-1 \leq s \leq 1$ and their corresponding label $l = \pm 1$, the three cost functions are written as

$$J_{square} = (s-l)^2, \quad (2)$$

$$J_{exp} = e^{-sl}, \quad (3)$$

$$J_{dev} = \ln(e^{-2sl}+1). \quad (4)$$

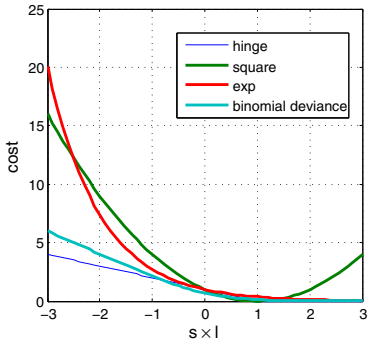

Fig. 4. Cost function candidates for the SCNN. The hinge cost is also drawn for reference, which is used in SVM classifier.

From Figure 4 we can see that Exponential loss give the largest cost when the similarity has incorrect sign and the shape of Deviance loss is very similar with Hinge loss which has been proved robust to outliers. By considering Hinge loss is not differentiable at $sl = 1$, we use Deviance as cost function to optimize the neural network.

By plugging Equ. (1) into Equ. (4), we can get the forward propagation function to calculate the cost from a sample pair.

$$J_{dev} = \ln(e^{-2Cosine(B_1(\mathbf{x}),B_2(\mathbf{y}))l}+1). \quad (5)$$

Differentiating the cost function with respect to the input samples $\mathbf{x}$ and $\mathbf{y}$, we can get

$$\frac{\partial J_{dev}}{\partial \mathbf{x}} = \frac{-2le^{-2Cosine(B_1(\mathbf{x}),B_2(\mathbf{y}))l}}{e^{-2Cosine(B_1(\mathbf{x}),B_2(\mathbf{y}))l}+1} \cdot \frac{1}{\|B_1(\mathbf{x})\|\|B_2(\mathbf{y})\|} \cdot (B_2(\mathbf{y}) - \frac{B_1(\mathbf{x})^T B_2(\mathbf{y}) B_1(\mathbf{x})}{B_1(\mathbf{x})^T B_1(\mathbf{x})}) \cdot \frac{dB_1}{d\mathbf{x}}, \quad (6)$$

and

$$\frac{\partial J_{dev}}{\partial \mathbf{y}} = \frac{-2le^{-2Cosine(B_1(\mathbf{x}),B_2(\mathbf{y}))l}}{e^{-2Cosine(B_1(\mathbf{x}),B_2(\mathbf{y}))l}+1} \cdot \frac{1}{\|B_1(\mathbf{x})\|\|B_2(\mathbf{y})\|} \cdot (B_1(\mathbf{x}) - \frac{B_2(\mathbf{y})^T B_1(\mathbf{x}) B_2(\mathbf{y})}{B_2(\mathbf{y})^T B_2(\mathbf{y})}) \cdot \frac{dB_2}{d\mathbf{y}}, \quad (7)$$

where $\|\cdot\|$ denotes the magnitude of vector.

Based on Equ. (5), Equ. (6) and Equ. (7), we can learn the parameters of SCNN by standard BP algorithms. In traditional neural network, the error is backward propagated from top to down through a single path. On the contrary, the error of SCNN is backward propagated through two branches by Equ. (6) and Equ. (7) respectively. When training samples from two different domains are sent to two branches, each branch can adapt to the corresponding domain well. In practice, we also can assign asymmetry label l to positive and negative sample pairs to tune the network, *e.g.,* 1 for positive pairs and -2 for negative pairs.

Our network is trained by batch based stochastic gradient descent. The size of batch is 128 including 64 positive and 64 negative image pairs. When converting a multi-class training set into binary-class, the number of negative pairs is far more than positive pairs. Therefore, we randomly select negative pairs from the whole negative sample pool for each batch. For

person re-identification problem, about 300 epoches are needed to obtain good results.

IV. Experiments

Five popular databases were built for person re-identification: VIPeR [4], i-LIDS [20], ETHZ [21], CIVAR [22] and CUHK Campus [5]. Among these databases, the evaluation protocol of VIPeR is the clearest one, therefore we compare our method with other methods on VIPeR. The experiments are conducted in two settings:

1) Single database person re-identification: training and testing both on VIPeR using view specific SCNN;
2) Cross database person re-identification: training on CUHK Campus and testing on VIPeR using general SCNN.

A. Single Database Person Re-Identification

Except for the newest paper [6], other papers all conduct experiments in this setting that is training and testing on the same database. VIPeR includes 632 subjects and 2 images per subject coming from 2 different camera views (camera A and camera B). We split VIPeR into disjoint training (316 subjects) and testing set (316 subjects) randomly, and repeat the process 11 times. The first split (Dev split) is used for parameter tuning, such as the number of epoch, learning rate, weight decay and so on. The other 10 splits (Test splits) are used for reporting the results.

In the training stage, all training images from camera A and camera B are grouped into pairs randomly and send to a view specific SCNN. The pairs from same subjects are labeled as positive, and those from different subjects are labeled as negative. Each camera corresponds a sub-network in the SCNN. In the testing stage, one image of each subject is used as gallery and the other one is used as probe.

Before evaluate the performance, we use the first part (head part) to tune the most two important parameters on the Dev split: the number of training epoch and the cost for negative sample pairs.

1) The Number of Epoch: Figure 5 shows the epoch-cost curve on the Dev split. Low cost reflects high performance approximately although there is no explicit relationship between the cost and the recognition rate. At the beginning of training, the cost of the test set drops significantly and it gradually becomes converged after epoch > 250. Finally, we set epoch $= 300$ based on our experience.

2) Asymmetric Cost: As described in Section III-C, the two class training set converted from multi-class is very asymmetric. The number of negative sample pairs is far more than positive pairs. In the training process, we cannot cover all negative pairs so just randomly select a portion of them to construct the training batch. This may cause the negative pairs prone to under-fitting.

To balance the weight of positive and negative sample pairs, we can assign asymmetric costs to them. While fixing the cost of positive pair to 1, we tune the cost of negative pair c from

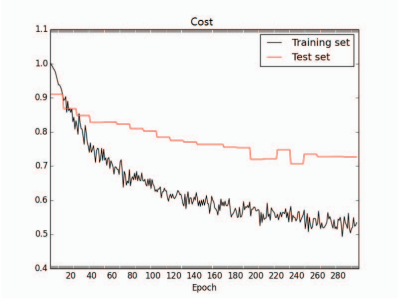

Fig. 5. The relationship between the number of epoch and the deviance cost on the Dev split. X axis is the number of training epoch, and Y axis is the deviance cost of the network.

TABLE I. The rank-30 recognition rates on the Dev split at different negative costs (part1: head).

Negative cost	Rank-30 recognition rate
c = 0.25	68.35%
c = 0.5	68.99%
c = 1	70.89%
c = 2	74.37%
c = 4	56.01%

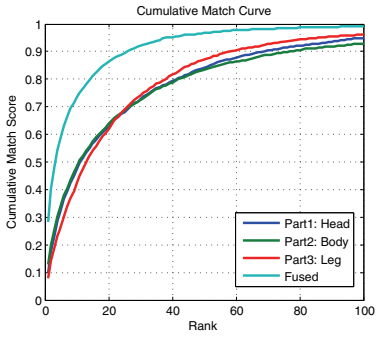

Fig. 6. The rank curves of the 3 parts and the whole image on VIPeR.

0.25 to 4. The asymmetric cost can apply easily on Equ. (5), Equ. (6) and Equ. (7) by setting

$$l = \begin{cases} 1 & \text{for positive pair} \\ -c & \text{for negative pair} \end{cases}. \quad (8)$$

Table I shows the relationship between the negative cost c and the rank-30 recognition rate of the first part (head part). On the Dev split, the highest performance is achieved at $c = 2$. This illustrates that the negative pairs should be paid more attention in each training batch.

3) Results: After tune the parameters, we keep them fixed in the following experiments, *i.e.,* epoch=300 and c=2. Because each person image is divided into three parts: head, body and leg, the training and testing are done for each part respectively and the three similarity scores are fused by sum rule. Repeating the experiments 10 times on the Test splits, the average rank curves are shown in Figure 6.

For precise comparison, we also list the recognition rates in Table II. The results of compared methods are copied from the original papers. If the results are unavailable, they are leaved as "-". From the table we can see that the proposed method outperforms most of compared methods. Especially, it is on a

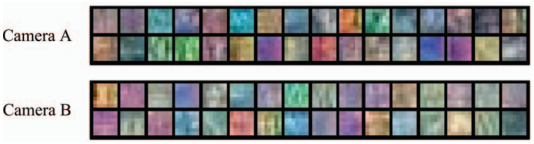

Fig. 7. Some convolutional filters learned on the VIPeR database. Top: the filters in the first sub-network for camera A; Bottom: the filters in the second sub-network for camera B.

par with the current state-of-the-art [5]. The rank-1 and rank-50 recognition rates of [5] are higher than ours slightly, and our method outperforms [5] in other situations. Among all methods, our method is nearly the most simple and elegant one. From the bottom to top layers in the network, every building block contributes to a common objective function and is optimized by BP algorithm simultaneously.

Figure 7 shows some filters learned by the first convolutional layer of our network. The filters of each sub-network have different colors and texture patterns, which means that they capture the information in the corresponding camera view efficiently. Because the filters in all layers are multi-channel, the color and texture information are fused in a very natural way.

B. Cross Database Person Re-Identification

In this section, we conduct a more challenging experiment which is coincide with practical applications. In practical systems, we usually collect a large dataset first and train a model on it. Then the trained model is applied to other datasets or videos for person image matching. A practical person re-identification algorithm should have good generalization with respect camera view changes and dataset changes. The previous works mainly focused on camera view changes but the cross database person re-identification problem was less studied.

[6] in the past ICCV 2013 has started to concern this problem. In [6], the authors proposed a transfer Rank SVM (DTRSVM) to adapt a model trained on the source domain (i-LIDS or PRID [25]) to target domain (VIPeR). All image pairs in the source domain and the negative image pairs in the target domain are used for training. Different from DTRSVM, our network trains in the source domain only and its generalization is tested in the target domain.

CUHK Campus database is used as training set, which includes 1816 subjects, 7264 images. Each subject has 4 images from 2 camera views. The resolution of CUHK Campus is 60×160. Before training, we scale them to 40×128 first. For the testing set, we use the same setting with the last experiment. A half of subjects and images are randomly selected from VIPeR to construct the testing set, including 316 subject, 632 images. The testing is repeated 10 times too, and the average rank curve is reported.

The images in CUHK Campus database are captured in 5 batches. The camera views between the batches are different and they are more different from the camera views of VIPeR. Due to the independence of camera views between the training and testing set, the view specific SCNN cannot be applied in this experiment. Therefore, a general SCNN is used for this task by letting the two sub-networks share their parameters. By grouping 7264 images into pairs, we get 10896 positive and 26368320 negative samples on CUHK Campus. Then the general SCNN is trained with the same number of epoch and negative cost as the previous experiment.

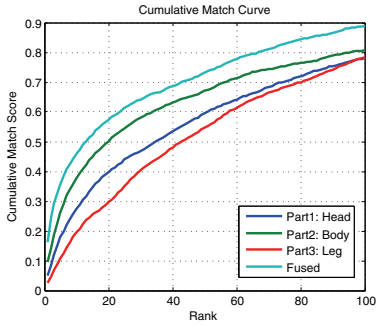

Fig. 8. The rank curves of the 3 parts and the whole image in the cross database experiment (CUHK Campus $\rightarrow$ VIPeR).

TABLE III. CROSS DATABASE EXPERIMENT: COMPARISON OF THE PROPOSED METHOD AND DTRSVM [6] ON VIPER.

Methods	Training Set	1	10	20	30
DTRSVM	i-LIDS	8.26%	31.39%	44.83%	53.88%
DTRSVM	PRID	10.90%	28.20%	37.69%	44.87%
Ours	CUHK Campus	**16.17%**	**45.82%**	**57.56%**	**64.24%**

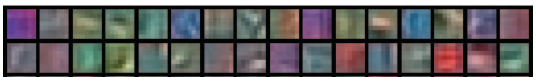

Fig. 9. Some convolutional filters learned on the CUHK Campus database.

Figure 8 shows the rank curve of three parts and their fusion. The performance differences of three parts are more obvious than Figure 6. From the figure, we can see that the order of performance is body>head>leg, which is consistent with our intuition. Generally, the body is the most stable part in the person images and the leg is most unstable. Parts fusion improves the performance significantly in both experiments.

The recognition rates are shown in Table III and the results of DTRSVM [6] are listed for comparison. Because the scale of i-LIDS and PRID is too small to train a good neural network, we only use CUHK Campus as training set. From the results we can see that our method outperforms DTRSVM significantly and even approach the performance of some methods in single database setting, such as ELF [12] and RDC [2].

In the visual sense, the images in CUHK Campus have richer texture and better quality than VIPeR. The filters learned on CUHK Campus verify this point. The filters in Figure 9 have clearer structure than those in Figure 7. But the VIPeR filters have higher contrast in color, this may be caused by the different capture environments of VIPeR and CUHK Campus.

V. CONCLUSIONS

This paper proposed a deep metric learning method by using siamese convolutional neural network. The structure and training process were described in detail. Two person re-identification experiments were conducted to illustrate the

TABLE II. SINGLE DATABASE EXPERIMENT: COMPARISON OF THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART METHODS ON VIPeR.

Method \ Rank	1	5	10	15	20	25	30	50
ELF [12]	12.00%	31.00%	41.00%	-	58.00%	-	-	-
RDC [2]	15.66%	38.42%	53.86%	-	70.09%	-	-	-
PPCA [23]	19.27%	48.89%	64.91%	-	80.28%	-	-	-
Salience [9]	26.74%	50.70%	62.37%	-	76.36%	-	-	-
RPML [24]	27%	-	69%	-	83%	-	-	95%
LAFT [5]	**29.6**%	-	69.31%	-	-	88.7%	-	**96.8**%
Ours	28.23%	**59.27**%	**73.45**%	**81.20**%	**86.39**%	**89.53**%	**92.28**%	96.68%

superiority of the proposed method. This is the first work to apply deep learning in the person re-identification problem and is also the first work to study the person re-identification problem in cross database setting. Extensive results illustrated that the network can switch flexibly between two modes to deal with the cross view and cross database person re-identification problems. In the future, we will apply DML to other applications; explore the way to pre-train the network; and investigate the effect of "dropout" in the applications. Moreover, we will continue to research how to train a general person matching engine with good generalization across database.

ACKNOWLEDGMENT

This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, Jiangsu Science and Technology Support Program Project #BE2012627, and AuthenMetric R&D Funds.

REFERENCES

[1] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2288–2295.

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

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *NIPS*, 1993, pp. 737–744.

[4] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*, 2007.

[5] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3594–3601.

[6] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 3567–3574.

[7] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, 2005, pp. 26–33 vol. 2.

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

[9] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3586–3593.

[10] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 7, pp. 1622–1634, 2013.

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

[12] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision C ECCV 2008*, ser. Lecture Notes in Computer Science, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin Heidelberg, 2008, vol. 5302, pp. 262–275.

[13] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *BMVC*, 2010, pp. 1–11.

[14] P. Baldi and Y. Chauvin, "Neural networks for fingerprint recognition," *Neural Computation*, vol. 5, no. 3, pp. 402–418, 1993.

[15] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR (1)*, 2005, pp. 539–546.

[16] H. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Computer Vision C ACCV 2010*, ser. Lecture Notes in Computer Science, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Springer Berlin Heidelberg, 2011, vol. 6493, pp. 709–720.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1106–1114.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] J. Friedman, R. Tibshirani, and T. Hastie, *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics. New York: Springer-Verlag, 2009.

[20] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 649–656.

[21] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proceedings of the XXII Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[22] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *British Machine Vision Conference (BMVC)*, 2011, p. 68.1C68.11.

[23] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2666–2672.

[24] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Computer Vision C ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7577, pp. 780–793.

[25] M. Hirzer, C. Beleznai, P. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, ser. Lecture Notes in Computer Science, A. Heyden and F. Kahl, Eds. Springer Berlin Heidelberg, 2011, vol. 6688, pp. 91–102.