Deep Hybrid Similarity Learning for Person Re-Identification

Jianqing Zhu, Huanqiang Zeng, Shengcai Liao, Zhen Lei, Canhui Cai, and Lixin Zheng

***Abstract*— Person re-identification (Re-ID) aims to match person images captured from two non-overlapping cameras. In this paper, a deep hybrid similarity learning (DHSL) method for person Re-ID based on a convolution neural network (CNN) is proposed. In our approach, a light CNN learning feature pair for the input image pair is simultaneously extracted. Then, both the elementwise absolute difference and multiplication of the CNN learning feature pair are calculated. Finally, a hybrid similarity function is designed to measure the similarity between the feature pair, which is realized by learning a group of weight coefficients to project the elementwise absolute difference and multiplication into a similarity score. Consequently, the proposed DHSL method is able to reasonably assign complexities of feature learning and metric learning in a CNN, so that the performance of person Re-ID is improved. Experiments on three challenging person Re-ID databases, QMUL GRID, VIPeR, and CUHK03, illustrate that the proposed DHSL method is superior to multiple state-of-the-art person Re-ID methods.**

***Index Terms*— Metric learning, convolution neural network, deep hybrid similarity learning, person re-identification (Re-ID).**

I. Introduction

PERSON Re-IDentification (Re-ID) aims to match person images captured from two non-overlapping cameras. It plays an important role in video surveillance for public safety [1]. In practical scenarios, person images are usually with low resolution and partial occlusion and contain large intra-class variations of illumination, viewpoint and pose, as shown in Fig. 1. Therefore, how to develop an effective person Re-ID method becomes a very challenging and desirable topic.

Fig. 1. Example pairs of images from the QMUL GRID [1], VIPeR [2] and CUHK03 [3] databases. Images in the same column represent the same person captured under different cameras. (a) QMUL GRID. (b) VIPeR. (c) CUHK03.

Note that the two fundamental problems critical for person Re-ID are feature representation and similarity metric. For feature representations, there are Fisher vectors (LDFV) [4], bio-inspired features (kBiCov) [5], symmetry-driven accumulated local features (SDALF) [6], structural constraints enhanced feature accumulation (SCEFA) [7], color invariant signature [8], salience matching [9], [10], ensemble of local features (ELF16) [11], ensemble of invariant features [12], mid-level learning features [13] and convolutional neural network learning features [14], and so on. These either hand-crafted or machine learning feature representations have promoted the research of person Re-ID.

For similarity metrics, many machine learning algorithms are developed to calculate the similarity between an image pair, such as ranking support vector machine (Ranking SVM) [15], partial least square (PLS) [16], Boosting [17], multi-task learning [18], [19], metric learning [20]–[28] and convolutional neural networks (CNNs) [29]–[40]. Note that the metric learning and CNN based person Re-ID methods are the most popular methods, which will be highlighted as follows.

The metric learning based person Re-ID methods [23]–[28] learn a matrix with $d \times d$ parameters to calculate the Mahalanobis distance between two d dimensional hand-crafted features as the similarity measurement between two person images. However, they are prone to over-fitting on a small database. Because the number of parameters in a learned Mahalanobis matrix is square of the feature dimension and this is a large number. For this, the principal component analysis (PCA) is usually used for feature dimension compression before metric learning. However, it is not optimal for metric learning, since PCA is not jointly optimized

Manuscript received October 1, 2016; revised February 17, 2017; accepted July 20, 2017. Date of publication August 1, 2017; date of current version November 5, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61602191, Grant 61672521, Grant 61375037, Grant 61473291, Grant 61572501, Grant 61572536, Grant 61502491, Grant 61372107, and Grant 61401167, in part by the Natural Science Foundation of Fujian Province under Grant 2016J01308, in part by the Scientific and Technology Founds of Xiamen under Grant 3502Z20173045, in part by the Promotion Program for Young and Middleaged Teacher in Science and Technology Research of Huaqiao University under Grant ZQN-PY418 and Grant ZQN-YX403, and in part by the Scientific Research Funds of Huaqiao University under Grant 16BS108. This paper was recommended by Associate Editor D. Bull. *(Corresponding author: Huanqiang Zeng.)*

J. Zhu, C. Cai, and L. Zheng are with the College of Engineering, Huaqiao University, Quanzhou 362021, China, and also with the Fujian Provincial Academic Engineering Research Centre in Industrial Intellectual Techniques and Systems, Quanzhou 362021, China (e-mail: jqzhu@hqu.edu.cn; chcai@hqu.edu.cn; zlxgxy@qq.com).

H. Zeng is with the College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China (e-mail: zeng0043@hqu.edu.cn).

S. Liao and Z. Lei are with the Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: scliao@cbsr.ia.ac.cn; zlei@cbsr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2734740

with metric learning. Another solution is proposed in [26], which is able to jointly learn a discriminant low dimensional subspace and a similarity metric by the cross-view quadratic discriminant analysis. In [26], although the number of parameters in the projection matrix is reduced to be $d \times s$, it is still large, where s is the compressed feature dimension.

Most existing CNN based person Re-ID methods [29], [31], [33]–[38] undervalue the similarity learning, which only apply the simple cosine or Euclidean distance function to measure the similarity between a CNN learning feature pair. Moreover, recent CNN based person Re-ID methods [34], [36], [38] pay more attentions on the feature learning via a deeper feature learning module, in which several specified layers are also designed for further emphasizing feature learning, such as the attention component [34], the matching gate architecture [36] and the domain guided dropout [38]. The tendency of undervaluing similarity learning and emphasizing feature learning leads to two deficiencies. (1) The simple cosine or Euclidean distance function is not very discriminative for measuring the similarity between a CNN learning feature pair. (2) The deeper feature learning module requires the larger scale training dataset. As a result, there is still ample room for the improvement on CNN based person Re-ID methods.

In this paper, an effective deep hybrid similarity learning (DHSL) method for person Re-ID is proposed. As a CNN based person Re-ID method, the major contribution of this paper is to improve person Re-ID performance by reasonably assigning complexities of metric learning and feature learning modules in the CNN model. For the metric learning module, a hybrid similarity function with reasonable parameters is designed to measure person similarity. For the feature learning module, a light convolution neural network only with three convolution layers is applied to extract features. The hybrid similarity function is realized by learning a group of weight coefficients to project the element-wise absolute difference and multiplication of a CNN learning feature pair into a similarity score. Since both the element-wise difference and multiplication of a CNN learning feature pair are considered, the hybrid similarity function is more discriminative than the simple cosine or Euclidean distance based similarity metric. Note that the number of parameters in the hybrid similarity function is only 2 times of the feature dimension, which is much smaller than that of the Mahalanobis distance based similarity metric. Consequently, it has been verified from extensive experiments on three challenging person Re-ID databases (i.e., QMUL GRID [1], VIPeR [2] and CUHK03 [3]) that the proposed DHSL method outperforms multiple state-of-the-art person Re-ID methods.

The rest of this paper is organized as follows. Section II introduces the details of the proposed deep hybrid similarity learning method. Section III presents experimental results and analyses. Section IV concludes this paper.

II. Deep Hybrid Similarity Learning (DHSL) for Person Re-ID

Fig. 2 shows the framework of the proposed DHSL method. The corresponding details will be introduced in the following three subsections, respectively.

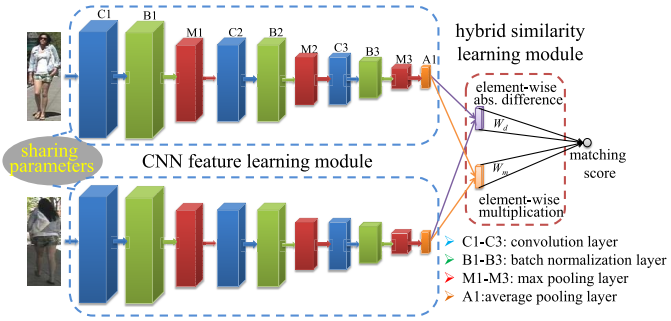

Fig. 2. The framework of the proposed DHSL method for person Re-ID.

A. Hybrid Similarity Learning Module

Assuming that the feature pair produced by the CNN feature learning module is $X_1, X_2 \in \Re^d$. Now, the question boils down to the similarity measurement between a feature pair. In this work, by analyzing the relationship among the Mahalanobis, Euclidean and cosine distances, a hybrid similarity function learned on the element-wise absolute differences and multiplications of feature pairs is proposed as below.

Firstly, the Mahalannobis and Euclidean distances are formulated in Eq. (1) and Eq. (2), respectively.

$$\begin{aligned} d_M^2(X_1, X_2) &= (X_1 - X_2)^T M (X_1 - X_2) \\ &= (vec(M))^T vec((X_1 - X_2)(X_1 - X_2)^T). \end{aligned} \tag{1}$$

$$d_E^2(X_1, X_2) = (X_1 - X_2)^T (X_1 - X_2). \tag{2}$$

The $vec(\cdot)$ function in Eq. (1) is used to rearrange a $d \times d$ dimensional matrix into a $d^2 \times 1$ dimensional vector. From these two equations, one can see that the Euclidean distance is the summation of d square differences, in which the k-th square difference is calculated at the k-th dimension. On the contrary, the Mahalanobis distance includes not only a linear combination of the d square differences, but also a linear combination of $d^2 - d$ correlations of the differences and each correlation is calculated between a feature pair at different dimensions. Hence, it can be easily observed that the Mahalanobis distance formulation is much more complex than the Euclidean distance formulation.

In addition to the Mahalannobis and Euclidean distances, the cosine distance is also one of commonly-used similarity metrics. Assuming that X_1 and X_2 have been ℓ_2 normalized, then cosine distance between them can be formulated as follows:

$$d_{cos}(X_1, X_2) = X_1^T \cdot X_2 = (X_1 . * X_2)^T \cdot \mathbf{1}, \tag{3}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^T \in \Re^d$ is a constant vector. To take a deep insight to the cosine distance, the Mahalanobis distance in Eq. (1) is further expanded as follows:

$$\begin{aligned} d_M^2(X_1, X_2) &= (vec(M))^T vec((X_1 - X_2)(X_1 - X_2)^T) \\ &= (vec(M))^T vec(X_1 X_1^T + X_2 X_2^T \\ &\quad - X_1 X_2^T - X_2 X_1^T). \end{aligned} \tag{4}$$

From Eqs. (3) and (4), one can see that the cosine distance is the summation of d correlations, in which the k-th correlation is only for the cross-correlation between X_1 and X_2 at the

TABLE I
THE NUMBERS OF PARAMETERS IN THE MAHALANNOBIS DISTANCE, THE EUCLIDEAN DISTANCE, THE COSINE DISTANCE AND THE PROPOSED HYBRID SIMILARITY, WHERE d IS THE FEATURE DIMENSION

Methods	Mahalannobis Eq. (1)	Euclidean Eq. (2)	Cosine Eq. (3)	Hybrid similarity Eq. (5)
Parameter number	$d \times d$	0	0	$2d$

k-th dimension. On the contrary, the Mahalanobis distance considers the cross-correlation of X_1 and X_2 and that of X_2 and X_1, the self-correlation of X_1 and that of X_2, in which both cross-correlations and self-correlations are calculated at the same dimension and different dimensions. Therefore, it can be concluded that the cosine distance is a simplification of the Mahalanobis distance.

Based on the above analysis, the Euclidean or cosine distance functions that are commonly-used in CNN are simpler but with lower discriminative ability to measure the similarity between a CNN learning feature pair. While the Mahalanobis distance is more discriminative but requires a large number of parameters ($d \times d$), which is thus not suitable to be integrated with a CNN directly. Therefore, we propose a hybrid similarity function that has not only a strong discriminative ability but also a reasonable number of parameters. The proposed hybrid similarity function is a linear combination of the element-wise absolute difference and multiplication between a feature pair as follows:

$$d_H(X_1, X_2) = W_d^T |X_1 - X_2| + W_m^T (X_1 . * X_2), \tag{5}$$

where $W_d \in \Re^d$ and $W_m \in \Re^d$ are two group of coefficients used to project the element-wise absolute difference and multiplication, respectively.

Table I summarizes the number of parameters in the Mahalannobis distance, the Euclidean distance, the cosine distance and the proposed hybrid similarity. Compared with the Mahalanobis distance in Eq. (4), the proposed hybrid similarity function is much simpler, since it only needs $2d$ parameters to take the difference and correlation information between pair features at each feature dimension into account. Compared with the Euclidean distance in Eq. (2), the proposed hybrid similarity function utilizes the absolute difference to replace the square difference for further simplifying the computation. In addition, the proposed hybrid similarity function considers both the element-wise difference and multiplication of a CNN learning feature pair by learning based coefficients W_d, $W_m \in \Re^d$, it therefore has a stronger discriminative ability.

Furthermore, to learn the proposed hybrid similarity function, the element-wise difference and multiplication layers are designed and integrated with the CNN feature learning module, as shown in Fig. 2. The forward and backward propagations of these two layers, and the corresponding objective functions are designed as follows.

1) Element-wise Absolute Difference Layer: The forward and backward propagations of the element-wise absolute difference layer are formulated as follows:

$$Diff(X_1, X_2) = |X_1 - X_2|, \tag{6}$$

$$\frac{\partial Diff}{dX_1^i} = \begin{cases} 1, & if\ X_1^i > X_2^i, \\ 0, & if\ X_1^i = X_2^i, \\ -1, & otherwise, \end{cases}$$

$$\text{and } \frac{\partial Diff}{dX_2^i} = -\frac{\partial Diff}{dX_1^i}, \tag{7}$$

where X_1^i and X_2^i represent i-th dimensions of X_1 and X_2, respectively.

2) Element-Wise Multiplication Layer: The forward and backward propagations of the element-wise multiplication layer are formulated as follows:

$$Mult(X_1, X_2) = X_1 . * X_2, \tag{8}$$

$$\frac{\partial Mult}{dX_1} = X_2 \text{ and } \frac{\partial Mult}{dX_2} = X_1. \tag{9}$$

B. CNN Feature Learning Module

As discussed before, the proposed hybrid similarity function is more discriminative than the cosine or Euclidean distance based similarity metric. Therefore, we apply a light CNN feature learning module for balancing the complexities between metric learning and feature learning modules.

As shown in Fig. 2, the proposed CNN feature learning module consists of two parameter sharing feature extraction branches. The so-called parameter sharing feature extraction branches means that the parameters of each branch are the same, which can be referred in [29], [31], [34], and [36]. In each branch, there are three convolution layers (i.e., C1, C2 and C3), three batch normalization [41] layers (i.e., B1, B2 and B3), three max pooling layers (i.e., M1, M2 and M3) and one average pooling layer (i.e., A1).

For C1, C2 and C3 layers, the zero padding operation is applied and 3×3 tiny sized filters are applied for saving filter parameters. For M1, M2 and M3 layers, the 3×3 max pooling operation is used. The strides of three convolution layers and three max pooling layers are set as 1 and 2, respectively. The A1 layer uses a 1×6 average pooling operation to calculate an average feature map on the horizontal direction to produce the feature representation of an input image. This strategy is inspired by our previous work [26], to improve the viewpoint robustness of the learned features. Moreover, the stride of the A1 layer is set as 1 to avoid compressing features excessively. The recommended parameter configuration for the CNN feature learning module is listed in Table II. This configuration is a commonly-used setting and selected based on experiments, following the way used in typical CNN based person Re-ID methods [29], [33], [35], [36], [38].

C. Objective Function Construction

Similar to [29] and [30], the person Re-ID problem is transformed into a classification problem: if a pair of person images holds the same ID, it will be a positive sample. Otherwise, it is a negative sample. To find a discriminative projection vector W to ensure the classification accuracy,

TABLE II
THE PARAMETER DETAILS OF THE CNN FEATURE LEARNING MODULE

Name	Output size ($h \times w \times c$)	Neuron	Filter ($h \times w \times c \times g$)	Stride	Pooling operation ($h \times w$)
C1	128×48×32	-	3×3×3×32	1	-
B1	128×48×32	ReLU	-	-	-
M1	64×24×32	-	-	2	3×3
C2	64×24×64	-	3×3×32×64	1	-
B2	64×24×64	ReLU	-	-	-
M2	32×12×64	-	-	2	3×3
C3	32×12×128	-	3×3×64×128	1	-
B3	32×12×128	ReLU	-	-	-
M3	16×6×128	-	-	2	3×3
A1	16×1×128	-	-	1	1×6
element-wise abs. difference	16×1×128	-	-	-	-
element-wise multiplication	16×1×128	-	-	-	-

Parameters h, w, c and g represent height, width, channel and group sizes, respectively.

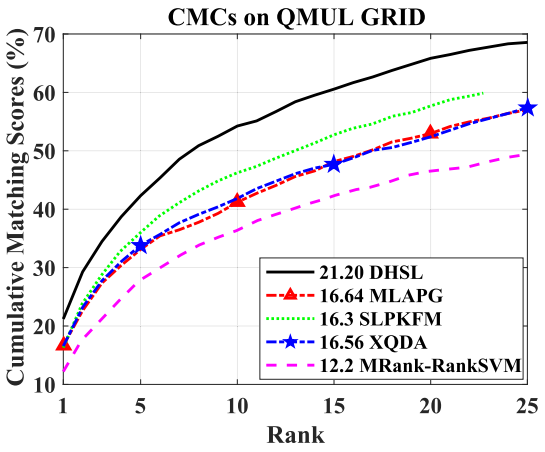

Fig. 3. CMC curves and rank-1 identification rates on QMUL GRID [1] (gallery: 125 individuals + 775 background images).

we apply a log-logistic model [42] to construct the objective function for the proposed hybrid similarity function learning:

$$W = \arg\min_{W}\{\frac{1}{K}[\sum\nolimits_{k=1}^{K} \log(1 + e^{-y_k W^T Z_k})] + \frac{1}{2}\alpha||W||_2^2\}, \tag{10}$$

where $W = [W_d, W_m] \in \Re^{d+d}$ is the hybrid project vector, α is a constant used to control the contribution of the regularization item, $\{(y_k, Z_k), k = [1, 2, \ldots, K]\}$ is the training dataset with K samples, $y_{(.)} \in \{-1, 1\}$ represents a class label, $Z_{(.)} = [Diff, Mult] \in \Re^{d+d}$ is an integration feature consisting of element-wise absolute difference ($Diff \in \Re^d$) and element-wise multiplication ($Mult \in \Re^d$). Note that this equation is optimized by using the stochastic gradient descent algorithm [43].

III. EXPERIMENT AND ANALYSIS

To validate the performance, the proposed DHSL method is evaluated and then compared with multiple state-of-the-art person Re-ID methods on three challenging datasets, QMUL GRID [1], VIPeR [2] and CUHK03 [3].

A. Dataset and Evaluation Protocol

QMUL GRID [1] contains 250 pedestrian image pairs, and each pair contains two images of the same person captured from 8 disjoint camera views in an underground station. Besides, there are 775 background images that do not belong to the 250 persons and can be used to enlarge the gallery set. The experimental setting of 10 random trials is provided by the GRID dataset. For each trial, 125 image pairs are used for training, and the remaining 125 image pairs as well as the 775 background images are exploited for testing. The average of cumulative match characteristic (CMC) [2] curves calculated on the 10 random trials is employed as the final result.

VIPeR [2] includes 632 person image pairs captured by a pair of cameras in an outdoor environment. Images in VIPeR have large variations in background, illumination, and viewpoint. Our experiments follow the widely adopted experimental protocol on VIPeR, which randomly divides 632 image pairs into 2 parts: half for training and the other half for testing, and repeats the procedure 10 times to obtain the average CMC as the final result.

CUHK03 [3] has 13,164 images of 1,360 pedestrians. These images are captured by 6 cameras over months, in which each person is observed by two disjoint camera views and has 4.8 images on average in each view. Both manually labeled and auto-detected pedestrian images are provided in CUHK03. Our experiments follow the same protocol in [3] as below. The CUHK03 is split into a training set with 1,160 persons and a test set with 100 persons. The experiments are conducted with 20 random trials and all the CMC curves are computed with the single-shot setting.

B. Implementation Detail

The implementation details of the proposed DHSL method can be described as below. All images in QMUL GRID, VIPeR and CUHK03 are scaled to 128×48 pixels. All the three datasets are augmented by the horizontal mirror operation. In addition, the small datasets QMUL GRID and VIPeR are further augmented by randomly rotating each image in the range $[-3°, 0°]$ and $[0°, 3°]$. For the network training, we initialize the weights in each layer from a normal distribution $N(0, 0.01)$ and the biases as 0. The regularization weight α in Eq. (10) for the two small datasets QMUL GRID and VIPeR is set as 5×10^{-2}, while that for the dataset CUHK03 is set as 5×10^{-4}. The size of mini-batch is 128 including 64 positive and 64 negative image pairs, and both positive and negative pairs are randomly selected from the whole training dataset. The momentums are set as 0.9. A base learning rate is started from 0.001 for QMUL GRID and VIPeR, while a larger learning rate (i.e., 0.01) is started for CUHK03 to accelerate the training process. The learning rates are gradually decreased as the training progress. That is, if the objective function is convergent at a stage, the learning rates are reduced to 1/10 of the original values, and the minimum learning rate is 10^{-4}.

TABLE III

PERFORMANCE COMPARISON OF THE PROPOSED DHSL METHOD AND MULTIPLE STATE-OF-THE-ART METHODS ON QMUL GRID [1] (GALLERY: 125 INDIVIDUALS + 775 BACKGROUND IMAGES)

Method	Rank=1 (%)	Rank=10 (%)	Rank=20 (%)	Reference
DHSL	**21.20**	**54.24**	**65.84**	proposed
SSDAL	19.1	45.8	58.1	2016 ECCV [37]
MLAPG	16.64	41.20	52.96	2015 ICCV [27]
SLPKFM	16.3	46.0	57.6	2015 CVPR [44]
XQDA	16.56	41.84	52.40	2015 CVPR [26]
MRank-RankSVM	12.2	36.3	46.6	2013 ICIP [45]

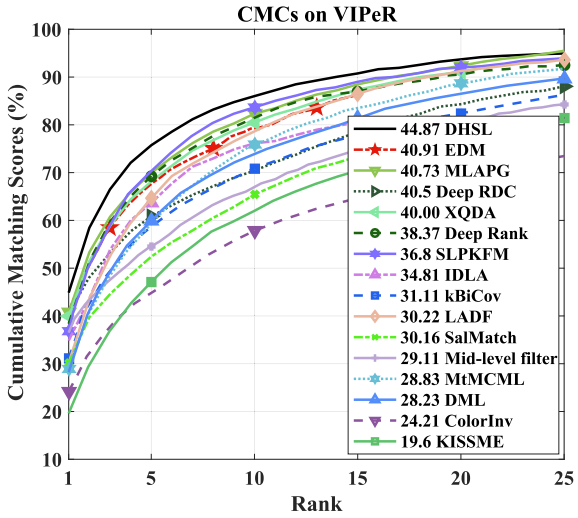

Fig. 4. CMC curves and rank-1 identification rates on VIPeR [2] (gallery: 316 individuals).

Moreover, the hard negative mining [30] is performed on CUHK03, since negative pairs on this big dataset are desirable to be used as fully as possible.

C. Comparison With State-of-the-Art Methods

1) Result on QMUL GRID: Fig. 3 and Table III show the performance comparison between the proposed DHSL method and the state-of-the-art person Re-ID methods on QMUL GRID [1]. It can be observed that the proposed DHSL method outperforms the semi-supervised deep attribute learning (SSDAL) [37] method by 2.1% rank-1 recognition rates, without using the assistance of human attributes. Moreover, the proposed DHSL method consistently outperforms the state-of-the-art metric learning based methods MLAPG [27], the polynomial kernel feature map (SLPKFM) method [44] and XQDA [26] at different ranks. This study shows that even for the small dataset QMUL GRID, the proposed DHSL method is able to obtain a promising performance.

2) Result on VIPeR: Fig. 4 and Table IV show the performance comparisons of the proposed DHSL method and the state-of-the-art person Re-ID methods on VIPeR [2]. One can see that the proposed DHSL method outperforms CNN based person Re-ID methods (i.e. EDM [33], Deep RDC [31], FT + JSTL + DGD [38], Deep Rank [35], SSDAL [37], GSCNN [36], IDLA [30] and DML [29]) and metric learning

TABLE IV

PERFORMANCE COMPARISON BETWEEN OUR PROPOSED DHSL METHOD AND MULTIPLE STATE-OF-THE-ART METHODS ON VIPeR [2] (GALLERY: 316 INDIVIDUALS). RED, GREEN AND BLUE COLORS REPRESENT THE 1st, 2nd AND 3rd BEST RESULTS, RESPECTIVELY

Method	Rank=1 (%)	Rank=10 (%)	Rank=20 (%)	Reference
DHSL	**44.87**	**86.01**	**93.70**	proposed
MTL-LORAE	42.30	81.60	89.60	2015 ICCV [19]
EDM	40.91	N/A	N/A	2016 ECCV [33]
MLAPG	40.73	82.34	92.37	2015 ICCV [27]
Deep RDC	40.5	70.4	84.4	2015 PR [31]
XQDA	40.00	80.51	91.08	2015 CVPR [26]
FT-JSTL+DGD	38.6	N/A	N/A	2016 CVPR [38]
Deep Rank	38.37	81.33	90.43	2016 TIP [35]
SSDAL	37.9	75.6	85.4	2016 ECCV [37]
SCNCD	37.8	81.2	90.4	2014 ECCV [46]
GSCNN	37.8	77.4	N/A	2016 ECCV [36]
SLPKFM	36.8	83.7	91.7	2015 CVPR [44]
IDLA	34.81	76.12	N/A	2015 CVPR [30]
kBiCov	31.11	70.71	82.44	2014 IVC [5]
LADF	30.22	78.82	90.44	2013 CVPR [25]
SalMatch	30.16	65.54	79.15	2013 ICCV [9]
Mid-level filter	29.11	65.95	79.87	2014 CVPR [13]
MtMCML	28.83	75.82	88.51	2014 TIP [18]
DML	28.23	73.45	86.39	2014 ICPR [29]
ColorInv	24.21	57.09	69.65	2013 TPAMI [8]
KISSME	19.6	62.2	77.0	2012 CVPR [22]

based person Re-ID methods (i.e., MTL-LORAE [19], MLAPG [27] and XQDA [26]). For example, the proposed DHSL method improves the rank-1 identification rate by 3.96%, 4.37% and 6.27% over EDM [33], Deep RDC [31] and FT + JSTL + DGD [38], respectively. The proposed DHSL method beats MTL-LORAE [19], MLAPG [27] and XQDA [26] by 2.57%, 4.14% and 4.87%, respectively, at rank 1. Moreover, the training of the proposed DHSL method is much simpler than FT + JSTL + DGD [38] and SSDAL [37]. Because the proposed DHSL method does not require a large database to pre-train the CNN, compared with FT + JSTL + DGD [38]. The proposed DHSL method also does not need to use additional human attributes to pre-train the CNN as that in SSDAL [37].

3) Result on CUHK03: Fig. 5 shows the performance comparison between the proposed DHSL method and the state-of-the-art person Re-ID methods CUHK03 [3]. As shown in Fig. 5(a), the proposed DHSL method beats the CNN based person Re-ID methods (i.e. CAN [34], PersonNet [32], EDM [33], IDLA [30], DeepReID [3]) and the metric learning based person Re-ID methods (i.e. MLAPG [27], XQDA [26] and KISSME [22]) on the manually labeled CUHK03 setting. As shown in Fig. 5(b), on the auto-detected CUHK03 setting, the similar comparison result is obtained, although the rank-1 identification rate of the proposed DHSL method is bit lower than that of the CAN [34] method.

In [38], both the domain individually CNN and domains jointed in single-task learning with domain guided dropout (JSTL + DGD) person Re-ID models are trained on a mixture setting. The mixture setting is constructed by mixing the manually labeled and auto-detected settings together. Since the mixture setting is large enough, we train

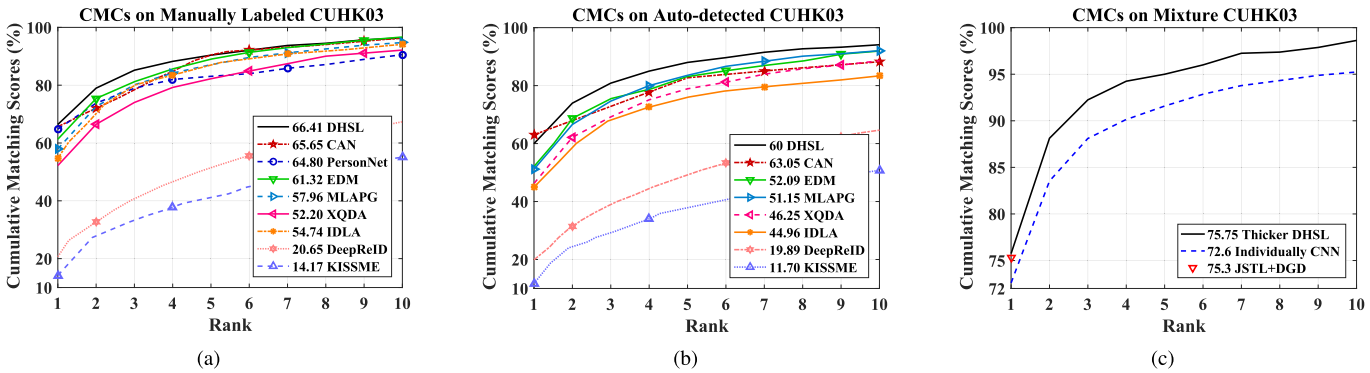

Fig. 5. CMC curves and rank-1 identification rates on CUHK03 [3] (gallery: 100 individuals). (a) Manually labeled setting. (b) Auto-detected setting. (c) Mixture setting, which mixes the manually labeled and auto-detected settings together for obtaining a large training set.

a thicker DHSL model. The number of channels in each layer in the thicker DHSL model is twice of that in the original DHSL model. For example, the C1, C2 and C3 layers in the thicker DHSL model hold 64, 128 and 256 channels, respectively. The thickest layers in individually CNN and JSTL + DGD models hold 1536 channels, thus the thicker DHSL model is still thinner than individually CNN and JSTL + DGD. As shown in Fig. 5 (c), the thicker DHSL obviously outperforms the individually CNN [38]. Moreover, the thicker DHSL obtains a bit higher rank 1 identification rate, compared with JSTL + DGD.

Based on the extensive experiments on either small datasets QMUL GRID, VIPeR or large dataset CUHK03, one can find that the proposed DHSL method outperforms both the state-of-the-art CNN and metric learning based person Re-ID methods. These results validate the effectiveness of the proposed DHSL method.

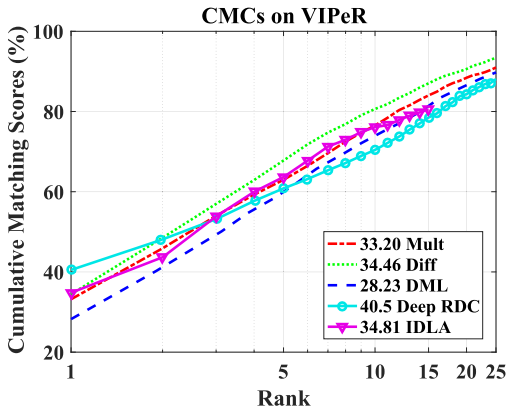

Fig. 6. CMC curves and rank-1 identification rates on VIPeR [2] (gallery: 316 individuals) resulted from three state-of-the-art CNN based person Re-ID methods and the proposed DHSL method under the element-wise absolute difference (Diff) and the element-wise multiplication (Mult) learning configurations, respectively.

D. Analysis of the Proposed DHSL Method

In this subsection, we further make a comprehensive performance analysis to show the effectiveness of each module in the proposed DHSL method, as follows.

1) Role of the Proposed CNN Feature Learning Module: To validate the effectiveness of the proposed CNN feature learning module, experiments are conducted under either the element-wise absolute difference (Diff) or the element-wise multiplication (Mult) configurations rather than the proposed hybrid similarity function. By assigning different values to W_d and W_m in Eq. (5), we can obtain the performance by considering only Diff ($W_m \equiv 0$) or Mult ($W_d \equiv 0$). The results are shown in Fig. 6. Note that with the Diff or Mult configuration, the trained CNN based person Re-ID model is similar to that used in the state-of-the-art CNN based person Re-ID models (i.e., IDLA [30], Deep RDC [31] and DML [29]), since these methods train their person Re-ID models on a single feature space constructed by calculating differences or multiplications of feature pairs.

From Fig. 6, one can see that the result by using the proposed CNN feature learning module under the Diff configuration is better than Deep RDC and IDML. Moreover, the result by using the proposed CNN feature learning module under the Mult configuration is also a bit better than that of DML. This study demonstrates the effectiveness of the proposed CNN feature learning module.

2) Role of Proposed Hybrid Similarity Module: In addition to the specially designed CNN feature learning module, extensive experiments are also conducted to evaluate the effectiveness of the proposed hybrid similarity learning module. The experiments consist of the following two parts: (1) the performance comparison between the proposed hybrid similarity function and the Mahalanobis distance based similarity metrics learned by the state-of-art metric learning methods, i.e., KISSME [22], ITML [20], MLAPG [27], and XQDA [26], using the same CNN feature (CNNFeat) that used in the proposed hybrid similarity function; (2) the performance comparison to the similarity metric by incorporating the Mahalanobis distance function into the proposed specially designed CNN feature learning module. Moreover, since the dimension of the feature (i.e., the output of the A1 layer in Fig. 2) extracted by the proposed CNN feature learning module is $16 \times 128 = 2048$ and the number of parameters in the Mahalanobis matrix is square of the feature dimension, a feature compression operation is necessary before the above-mentioned Mahalannobis distance based similarity metric learning. For that, in experimental part (1), for the methods, KISSME [22], ITML [20] and MLAPG [27], the PCA is used to compress the feature

TABLE V

TOP RANKED IDENTIFICATION RATES UNDER DIFFERENT FEATURE DIMENSIONS ON QMUL GRID [1]. DIM REPRESENTS THE COMPRESSED FEATURE DIMENSION

Method	Dim	Rank=1 (%)	Rank=10 (%)	Rank=20 (%)	Rank=30 (%)
DHSL	-	**21.20**	**54.24**	**65.84**	**71.28**
CNNFeat+PCA+KISSME	16	12.24	41.04	54.96	63.68
	32	16.96	47.68	58.64	66.48
	64	17.20	44.32	54.80	61.60
	128	15.04	39.36	50.48	57.12
	256	17.12	40.96	49.68	56.40
CNNFeat+PCA+ITML	16	5.84	25.44	34.00	38.96
	32	7.92	29.20	40.16	46.80
	64	9.76	35.92	47.44	54.72
	128	13.60	42.08	52.96	61.60
	256	14.32	44.40	55.68	63.60
CNNFeat+PCA+MLAPG	16	9.76	37.52	50.24	59.12
	32	15.76	43.12	56.64	64.24
	64	14.96	44.08	57.36	66.08
	128	15.60	42.80	57.04	65.28
	256	15.36	42.40	56.24	64.96
CNNFeat+XQDA	16	12.88	41.36	54.48	63.20
	32	13.52	43.28	55.76	62.96
	64	16.08	41.68	54.32	61.12
	128	15.60	39.84	49.76	55.04
	256	15.60	39.76	49.76	55.04
CNNStru+FC+Mah	16	2.72	22.56	37.76	47.36
	32	3.36	24.48	38.64	47.52
	64	3.92	26.00	39.04	48.80
	128	4.16	25.84	38.56	49.04
	256	4.00	26.48	40.00	47.60

TABLE VI

TOP RANKED IDENTIFICATION RATES UNDER DIFFERENT FEATURE DIMENSIONS ON VIPeR [2]. DIM REPRESENTS THE COMPRESSED FEATURE DIMENSION

Method	Dim	Rank=1 (%)	Rank=10 (%)	Rank=20 (%)	Rank=30 (%)
DHSL	-	**44.87**	**86.01**	**93.70**	**95.89**
CNNFeat+PCA+KISSME	16	19.24	63.73	80.03	88.20
	32	26.68	72.59	85.82	91.87
	64	32.25	74.56	86.55	91.74
	128	30.76	71.08	82.94	89.27
	256	27.37	65.09	77.59	83.26
CNNFeat+PCA+ITML	16	5.82	22.34	30.89	37.50
	32	11.93	38.35	49.08	55.79
	64	20.19	58.54	71.87	79.30
	128	17.15	51.08	63.86	70.76
	256	25.57	67.53	80.70	87.18
CNNFeat+PCA+MLAPG	16	16.39	60.28	77.75	85.98
	32	23.07	69.40	83.29	89.30
	64	29.18	73.13	86.17	91.33
	128	28.99	73.48	85.95	91.77
	256	28.10	72.69	85.66	91.33
CNNFeat+XQDA	16	23.92	68.26	82.72	89.78
	32	28.39	70.09	82.41	89.21
	64	28.67	67.34	79.75	86.27
	128	26.33	61.42	74.43	81.11
	256	23.51	55.32	67.31	74.21
CNNStru+FC+Mah	16	15.16	58.89	75.85	83.96
	32	16.90	63.86	79.49	86.93
	64	17.34	64.65	79.59	87.03
	128	18.64	63.83	79.08	86.90
	256	17.91	64.78	79.24	86.33

dimension, while the XQDA [26] can automatically realize the feature dimension compression. Hence, these methods are denoted as CNNFeat + PCA + KISSME, CNNFeat + PCA + ITML, CNNFeat + PCA + MLPAG, and CNNFeat + XQDA, respectively. In experimental part (2), the bottom CNN feature learning module (see Fig. 2, C1→...→A1) is kept the same with that is used for the proposed hybrid similarity function. Differently, an additional full connection (FC) layer is integrated after the A1 layer of the proposed CNN feature learning module to realize the feature compression. Hence, this method is denoted as CNNStru + FC + Mah. Moreover, considering that the feature compression degree has an important influence on the performance, the experiments are performed under different compressed feature dimensions. The corresponding results are shown in Tables V, VI and Fig. 7.

As shown in Tables V and VI, the proposed method obtains higher recognition rates than CNNFeat + PCA + KISSME, CNNFeat + PCA + ITML, CNNFeat + PCA + MLAPG, and CNNFeat + XQDA both on QMUL GRID and VIPeR datasets. This is because these methods independently optimize the feature learning and metric learning, while the proposed hybrid similarity function is able to jointly optimize the feature learning and metric learning. More specifically, for CNNFeat + PCA + KISSME, CNNFeat + PCA + ITML and CNNFeat + PCA + MLAPG, the feature compression via PCA does not consider the metric learning in general. For CNNFeat + XQDA, although XQDA can find an optimal subspace for metric learning, it still requires a large number of parameters in the feature compression, which is prone to over-fitting on a small dataset. On the contrary, the proposed

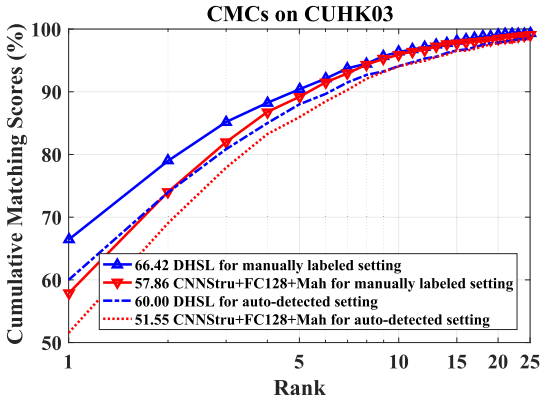

Fig. 7. CMC curves and rank-1 identification rates on CUHK03 [3] (gallery: 100 individuals).

method does not require a feature compression operation. Consequently, the proposed hybrid similarity function achieves a superior performance.

In addition, the proposed hybrid similarity function also obtains obvious improvement both on small (i.e., QMUL GRID and VIPeR) and large (i.e., CHUK03) datasets, compared with CNNStru + FC + Mah, as shown in Tables V and VI, and Fig. 7. Note that for the large dataset CUHK03, only the compressed feature dimension 128 is learned and denoted as CNNStru + FC128 + Mah, as it has better result on QMUL and VIPeR datasets. The superior performance achieved by the proposed DHSL method further indicates that the proposed hybrid similarity function is more

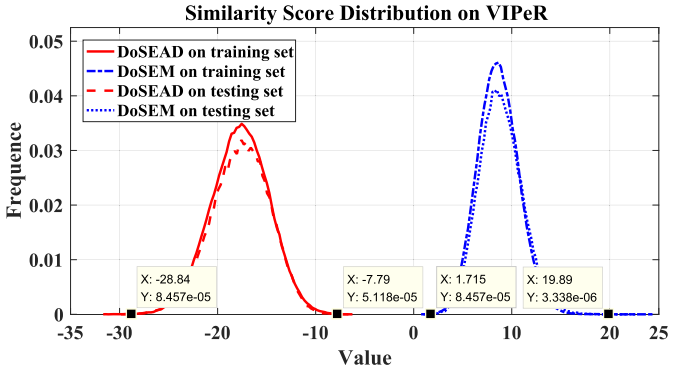

Fig. 8. Similarity score distributions on VIPeR [2]. DoSEAD represents the Distribution of similarity Scores calculated by projecting Element-wise Absolute Differences with W_d in Eq. (5). DoSEM represents the Distribution of similarity Scores calculated by projecting Element-wise Multiplications with W_m in Eq. (5).

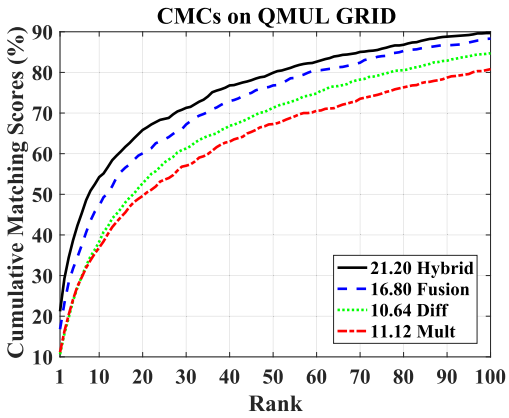

Fig. 9. CMC curves and rank-1 identification rates on QMUL GRID [1] (gallery: 125 individuals + 775 background images) resulted from the proposed DHSL method under the element-wise absolute difference (Diff), the element-wise multiplication (Mult), the simple score fusion (Fusion) of Diff and Mult, and the hybrid similarity learning configurations, respectively.

suitable to be integrated with a CNN for person Re-ID than the Mahalannobis distance function, since it can save a lot of parameter in the similarity metric learning so as to relieve the over-fitting problem.

3) Role of the Element-Wise Absolute Difference and Multiplication Complementary Behavior: First, we validate that it is able to learn a hybrid similarity function by projecting element-wise absolute differences and multiplications into similarity scores simultaneously. As shown in Fig. 2, there are three parameter sharing batch normalization layers in each feature extraction branch of the proposed CNN feature learning module, which are able to limit the scale of output features. Moreover, the element-wise absolutions and multiplications are further projected by the learned parameters (i.e., W_d and W_m in Eq. (5)). The Distribution of similarity Scores calculated by projecting Element-wise Absolute Differences (DoSEAD) with W_d and the Distribution of similarity Scores calculated by projecting Element-wise Multiplications (DoSEM) with W_m are evaluated, as shown in Fig. 8. One can see that both on the training and testing settings of VIPeR [2], the rough ranges of DoSEAD and DoSEM are $[-28.84, -7.79]$ and $[1.715, 19.89]$, respectively. This study indicates that the element-wise absolute differences and multiplications have no large scale difference. Therefore, it is able to learn a hybrid similarity function that considers element-wise absolute differences and multiplications simultaneously.

Second, we validate the proposed hybrid similarity function simultaneously learned on element-wise absolute differences and multiplications is helpful for improving the performance of person Re-ID. We evaluate the proposed DHSL method under the element-wise absolute difference (Diff), the element-wise multiplication (Mult), the simple score fusion of Diff and Mult (Fusion), and the proposed hybrid similarity learning (Hybrid) configurations, respectively. The fusion configuration is to independently train Diff and Mult person Re-ID models under the Diff and Mult configurations and then simply summarize the similarity scores from the Diff and Mult person Re-ID models as the final similarity score.

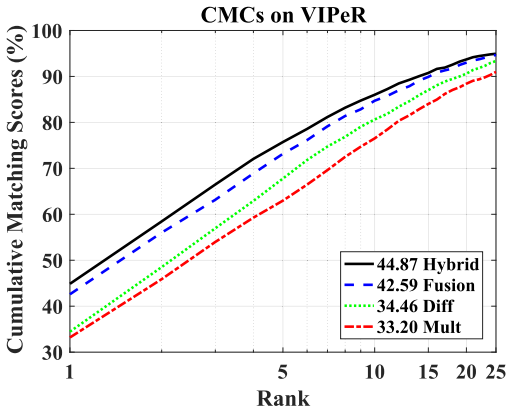

Fig. 10. CMC curves and rank-1 identification rates on VIPeR [2] (gallery: 316 individuals) resulted from the proposed DHSL method under the element-wise absolute difference (Diff), the element-wise multiplication (Mult), the simple score fusion (Fusion) of Diff and Mult, and the hybrid similarity learning configurations, respectively.

It can be observed from the results shown in Figs. 9 and 10 that the CMC resulted from the Fusion method is obvious better than that resulted from either the Diff or Mult method. This indicates the element-wise absolute difference and multiplication play an effective complementary role to each other for improving the person Re-ID performance. Moreover, it can be further found that the proposed hybrid similarity method (Hybrid) outperforms the Fusion method. This is due to the fact that the proposed deep hybrid similarity learning method is able to maximize the complementary effectiveness between Diff and Mult.

4) Role of Training Dataset Scale: In this experiment, the number of training person individuals is changed, and the number of testing person individuals is kept the same (i.e., 125 testing person individuals on QMUL GRID [1], 316 testing person individuals on VIPeR [2]). Moreover, the above-mentioned data augmentation operation is applied.

From the results shown in Tables VII and VIII, one can observe that the performance of the proposed DHSL

TABLE VII

TOP RANKED IDENTIFICATION RATES UNDER DIFFERENT TRAINING DATASET SCALES ON QMUL GRID [1]

Training Scale (Individuals)	Rank=1 (%)	Rank=10 (%)	Rank=20 (%)	Rank=30 (%)
50	14.48	42.72	54.96	63.36
75	18.48	46.64	58.64	66.48
100	**21.20**	51.44	63.84	70.40
125	**21.20**	**54.24**	**65.84**	**71.28**

TABLE VIII

TOP RANKED IDENTIFICATION RATES UNDER DIFFERENT TRAINING DATASET SCALES ON VIPER [2].

Training Scale (Individuals)	Rank=1 (%)	Rank=10 (%)	Rank=20 (%)	Rank=30 (%)
100	30.79	72.78	83.92	88.70
150	34.72	78.86	88.51	92.91
200	41.61	81.90	91.23	94.56
250	41.77	83.54	92.25	95.35
300	44.08	84.97	93.01	95.70
316	**44.87**	**86.01**	**93.70**	**95.89**

method is increased with the training dataset scale on both QMUL GRID [1] and VIPeR [2] datasets. Moreover, from the results on QMUL GRID in Tables III and VII, it can be seen that the proposed DHSL method trained by only using 75 training person individuals obtains a better performance than MLAPG [27], SLPKFM [44], XQDA [26] and MRank-RankSVM [45]. Similarly, from the results on VIPeR in Tables IV and VIII, it can be found that the proposed DHSL method trained by only using 200 training person individuals is comparable to MTL-LORAE [19], MLAPG [27], XQDA [26] and Deep RDC [31], and is better than the rest methods, such as LADF [25], MtMCML [18], KISSME [22], IDLA [30] and DML [29], and so on. These results illustrate that the proposed DHSL method is less dependent on the training data scale. This is because the proposed DHSL method has a reasonable number of parameters.

5) Running Time Analysis: To validate the running time advantage of the proposed DHSL method, we evaluate the feature extraction time (FET) per image and similarity calculation time (SCT) per image pair. The software tools are Matconvnet [47], CUDA 8.0, CUDNN V5.1, MATLAB 2016 and Visual Studio 2015. We re-implement LOMO[1] [26] and EFL16[2] [11] feature extraction codes provided by their authors to evaluate FETs. Both for LOMO and EFL16, the FET excludes the feature compression running time, since the initial code does not provide feature compression codes. The pretrained JSTL + DGD [3] [38] model is implemented in the Caffe [48] deep learning framework. The running time results are shown in Table IX.

As shown in Table IX, the summation of FET and SCT resulted from the proposed DHSL method is less than half of the FET of LOMO [26], using the same CPU setup. Moreover, the FETs of the proposed DHSL method are only 4.48% and 6.13% of those of JSTL + DGD [38] under the same CPU and GPU settings, respectively. Even for the thicker DHSL, its FETs are only 9.28% and 14.26% of those of JSTL + DGD [38] under the same CPU and GPU settings, respectively. This study clearly illustrates that the proposed DHSL method is much more efficient than the state-of-the-art methods.

[1]http://www.cbsr.ia.ac.cn/users/scliao/projects/lomo_xqda/index.html

[2]http://isee.sysu.edu.cn/~chenyingcong/code/demo_feat.zip

[3]https://github.com/Cysu/dgd_person_reid

TABLE IX

RUNNING TIME COMPARISON OF DIFFERENT METHODS. FET AND SCT REPRESENT FEATURE EXTRACTION TIME PER IMAGE AND SIMILARITY CALCULATION TIME PER IMAGE PAIR, RESPECTIVELY

Method	Mex	CPU: I7-6820HQ @2.7 Hz		GPU: NVIDIA Quadro M1000M	
		FET (msec/image)	SCT (msec/pair)	FET (msec/image)	SCT (msec/pair)
DHSL	Yes	**4.9781**	**0.0373**	**0.3640**	**0.0028**
LOMO [26]	Yes	10.1297	N/A	N/A	N/A
ELF16 [11]	No	254.6033	N/A	N/A	N/A
Thicker DHSL	Yes	10.3066	0.0749	0.8462	0.0053
JSTL+DGD [38]	Yes	111.0322	N/A	5.9347	N/A

IV. CONCLUSION

In this paper, a deep hybrid similarity learning (DHSL) method for person Re-IDentification (Re-ID) is proposed. The superior performance of DHSL is achieved by reasonably assigning complexities of metric learning and feature learning modules in the CNN model. In the metric learning module, the hybrid similarity function is proposed and learned on the element-wise absolute differences and multiplications of the CNN learning feature pairs simultaneously, which yields a more discriminative similarity metric. In the feature learning module, a light CNN only including three convolution layers is applied. We further examine the effectiveness of each role in the proposed DHSL method, e.g., complementary behavior of element-wise absolute differences and multiplications, training dataset scale and running time analysis. Experiments on three challenging person Re-ID databases, QMUL GRID, VIPeR and CUHK03, show the proposed DHSL method consistently outperforms multiple state-of-the-art person Re-ID methods.

REFERENCES

[1] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. CVPR*, Jun. 2009, pp. 1988–1995.

[2] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. Workshop Perform. Eval. Tracking Surveill.*, 2007, pp. 1–7.

[3] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, Jun. 2014, pp. 152–159.

[4] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proc. ECCV Workshop*, 2012, pp. 413–422.

[5] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image Vis. Comput.*, vol. 32, nos. 6–7, pp. 379–390, 2014.

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. CVPR*, Jun. 2010, pp. 2360–2367.

[7] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Exploring structural information and fusing multiple features for person re-identification," in *Proc. CVPR Workshop*, Jun. 2013, pp. 794–799.

[8] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, Jul. 2013.
[9] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. ICCV*, Dec. 2013, pp. 2528–2535.
[10] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, Jun. 2013, pp. 3586–3593.
[11] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. Yuen, "An asymmetric distance model for cross-view feature mapping in person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1661–1675, 2017.
[12] Y.-G. Lee, S.-C. Chen, J.-N. Hwang, and Y.-P. Hung, "An ensemble of invariant features for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 470–483, Mar. 2017.
[13] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. CVPR*, Jun. 2013, pp. 144–151.
[14] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li, "Cross dataset person re-identification," in *Proc. ACCV Workshop*, 2014, pp. 650–664.
[15] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *Proc. BMVC*, 2010, vol. 1. no. 3, p. 5.
[16] W. R. Schwartz and L. S. Davis, "Learning discriminative appearance-based models using partial least squares," in *Proc. Brazilian Symp. Comput. Graph. Image Process.*, Oct. 2009, pp. 322–329.
[17] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008, pp. 262–275.
[18] L. Ma, X. Yang, and D. Tao, "Person re-identification over camera networks using multi-task distance metric learning," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3656–3670, Aug. 2014.
[19] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proc. ICCV*, Dec. 2015, pp. 3739–3747.
[20] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. ICML*, Jun. 2007, pp. 209–216.
[21] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. NIPS*, 2005, pp. 1473–1480.
[22] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, Jun. 2012, pp. 2288–2295.
[23] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR*, Jun. 2011, pp. 649–656.
[24] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Proc. ECCV*, 2012, pp. 780–793.
[25] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. CVPR*, Jun. 2013, pp. 3610–3617.
[26] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, Jun. 2015, pp. 2197–2206.
[27] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. ICCV*, Dec. 2015, pp. 3685–3693.
[28] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1846–1855.
[29] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, Aug. 2014, pp. 34–39.
[30] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. CVPR*, Jun. 2015, pp. 3908–3916.
[31] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.
[32] L. Wu, C. Shen, and A. van den Hengel. (2016). "PersonNet: Person re-identification with deep convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1601.07255
[33] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. ECCV*, 2016, pp. 732–748.
[34] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. (2016). "End-to-end comparative attention networks for person re-identification." [Online]. Available: https://arxiv.org/abs/1606.04404
[35] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.
[36] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. ECCV*, 2016, pp. 791–808.
[37] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. ECCV*, 2016, pp. 475–491.
[38] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1249–1258.
[39] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.
[40] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4004–4012.
[41] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167
[42] R. C. Gupta, O. Akman, and S. Lvin, "A study of log-logistic model in survival analysis," *Biometrical J.*, vol. 41, no. 4, pp. 431–443, 1999.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[44] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. CVPR*, Jun. 2015, pp. 1565–1573.
[45] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *Proc. ICIP*, Sep. 2013, pp. 3567–3571.
[46] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. ECCV*, 2014, pp. 536–551.
[47] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 689–692.
[48] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

Jianqing Zhu received the B.S. degree in communication engineering and the M.S. degree in communication and information system from the School of Information Science and Engineering, Huaqiao University, Xiamen, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Assistant Professor with the College of Engineering, Huaqiao University, Quanzhou, China. His current research interests include computer vision and pattern recognition, with a focus on image and video analysis, particularly person re-identification, object detection, and video surveillance. He received the Best Biometrics Student Paper Award at the International Conference on Biometrics in 2015.

Huanqiang Zeng received the B.S. and M.S. degrees in electrical engineering from Huaqiao University, Xiamen, China, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore. He was a Post-Doctoral Fellow with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, from 2012 to 2013, and a Research Associate with Temasek Laboratories, Nanyang Technological University, Singapore, in 2008. He is currently a Professor with the School of Information Science and Engineering, Huaqiao University. He has authored over 50 papers in well-known international journals and conferences. His research interests include visual information processing and analysis, image/video communication, and computer vision. He has been actively serving as an associate/guest editor of multiple international journals, a General Co-Chair of the 2017 IEEE International Symposium on Intelligent Signal Processing and Communication Systems, a Technical Co-Chair of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, an Area Chair of the 2015 IEEE International Conference on Visual Communications and Image Processing, and a technical program committee member of multiple international conferences.

Shengcai Liao received the B.S. degree in mathematics and applied mathematics from Sun Yatsen University, Guangzhou, China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He was a Post-Doctoral Fellow with the Department of Computer Science and Engineering, Michigan State University, from 2010 to 2012. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include computer vision, pattern recognition, with a focus on image and video analysis, particularly face recognition, object detection, person re-identification, metric learning, and video surveillance. He received the Motorola Best Student Paper award and the First Place Best Biometrics Paper Award at the International Conference on Biometrics in 2006 and 2007, respectively, for his work on face recognition, and the Best Reviewer Award in IJCB 2014.

Zhen Lei received the B.S. degree in automation from the University of Science and Technology of China, in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. He has authored over 100 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an Area Chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, and 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.

Canhui Cai received the B.S. degree from Xidian University, Xi'an, China, in 1982, the M.S. degree from Shanghai University, Shanghai, China, in 1985, and the Ph.D. degree from Tianjin University, Tianjin, China, in 2003, all in electronic engineering. Since 1984, he has been with the Faculty of Huaqiao University, Quanzhou, China. He was a Visiting Professor with the Delft University of Technology, Delft, The Netherlands, from 1991 to 1992, and a Visiting Professor with the University of California at Santa Barbara, Santa Barbara, CA, USA, from 1999 to 2000. He is currently a Professor with the College of Engineering. He has authored or co-authored four books, and has authored over 150 papers in journals and conference proceedings. His research areas include video communications, image and video signal processing, and computer vision. He was a General Co-Chair of the Intelligent Signal Processing and Communication Systems in 2007.

Lixin Zheng received the B.S. degree in electronic technology application and the M.S. degree in machinery manufacturing from Huaqiao University, Quanzhou, China, in 1987 and 1990, respectively, and the Ph.D. degree in measurement technology and instrument from Tianjin University in 2002. He is currently a Professor and the Dean of the College of Engineering, Huaqiao University. His current research interests include motion control, power supply technology, computer vision, and pattern recognition.