BODY STRUCTURE BASED TRIPLET CONVOLUTIONAL NEURAL NETWORK FOR PERSON RE-IDENTIFICATION

Hong Liu, Weipeng Huang

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
{hongliu, wepon}@pku.edu.cn

ABSTRACT

Person re-identification remains a challenging problem due to large variations of poses, occlusions, illumination and camera views. To learn both feature representation and similarity metric simultaneously, deep metric learning methods using triplet convolutional neural network have been applied in person re-identification. In this paper, we propose a body structure based triplet convolutional neural network (BSTCNN) for person re-identification. Specifically, a four-branch CNN architecture is built to learn features from different body parts. Body-part features are then fused in score level with a novel weighted distance layer which learns weights for each body part. We further design an improved triplet loss function to speed up convergence and boost the performance. Experimental results on two challenging datasets (CUHK01 and PRID2011) demonstrate that our approach significantly outperforms the state-of-the-art methods.

Index Terms— Person Re-identification, Deep Metric Learning, Weighted Distance Layer, Triplet Loss

1. INTRODUCTION

Person re-identification aims to match pedestrian images from non-overlapping camera views, which is fundamental and essential for surveillance and security systems. Despite significant advances, it still remains a challenging problem due to camera view changes, pose variations, low resolutions, unstable illumination and occlusions.

The framework of existing methods typically consists of two components: a feature extraction method to describe the pedestrian images, and a corresponding similarity metric to compute the distance of image pairs. Many traditional methods consider these two problems separately, focusing either on improving suitable hand-crafted features [1–3], or finding good distance metrics for comparison [4–6].

This work is supported by National High Level Talent Special Support ProgramNational Natural Science Foundation of China (NSFC, No.61340046,61673030U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011), Natural Science Foundation of Guangdong Province (No.2015A030311034), Scientific Research Project of Guangdong Province (No.2015B010919004).

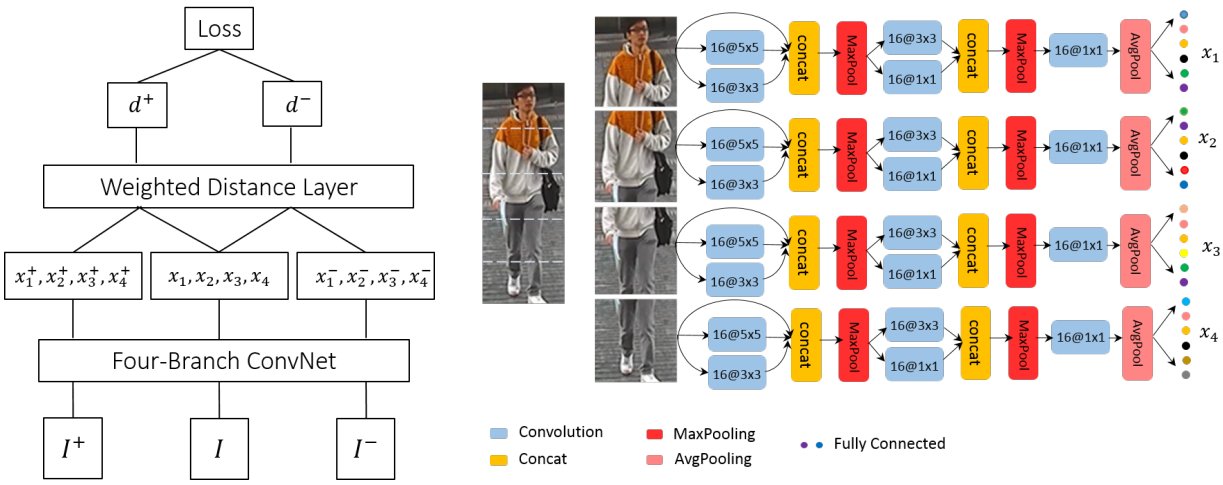
Recently, Convolutional Neural Network (CNN) based methods [7–12] gain increasing popularity in person re-identification. Among these methods, deep metric learning using triplet network has shown excellent performance by incorporating the two above-mentioned components into a unified framework. Shi *et al.* proposed a Constrained Deep Metric Learning (CDML) [9] method and built their CNN architecture in three branches to learn specific features from each body part. Similarly, Cheng *et al.* proposed a multi-channel parts-based convolutional neural network model [10] under the triplet framework. These two methods both trained the CNN in part based way so as to capture different statistical properties of body parts. However, high level features extracted from each body part are just simply concatenated in these methods, which can not effectively utilize the body structure information since different body parts are of different importance, as indicated by previous works [8, 10].

To deal with this problem, we propose a body structure based triplet convolutional neural network (BSTCNN) which consists of two main contributions:

(1) A novel weighted distance layer is designed to fuse body-part features in score level. Weights for each body-part distance are learned by back propagation algorithm.

(2) An improved loss function that combines both triplet loss and contrastive loss is utilized to speed up convergence and boost the performance.

Relation to prior work: Bromley *et al.* introduced siamese network [13] using contrastive loss which minimises intra-class distance and maximises inter-class distance independently. Schroff *et al.* proposed a triplet network called FaceNet [14] using triplet loss which forces intra-class distance to be less than inter-class distance. We borrow ideas from both of these two works and demonstrate that a combination of the triplet and contrastive loss produces better performance. Many previous works [9–11, 15] proposed multi-branch CNN architecture to learn specific features for each body part. However, these methods simply concatenate body-part features to construct a single vector and ignored the importance of different body parts. This situation motivates us to design a weighted distance layer to fuse features in score level and learn weights for each body part.



(a) Framework of BSTCNN

(b) Four-Branch ConvNet Architecture

Fig. 1: Framework of our method (BSTCNN) and the corresponding Four-Branch ConvNet architecture.

2. PROPOSED METHOD

In this section, we first describe the overall framework of our method (BSTCNN) and the corresponding architecture of the four-branch convolutional neural network, then introduce the weighted distance layer and the improved loss function. Finally, we present formulations of the optimization algorithm.

2.1. Overall Framework

The framework of our method is given in Fig.1(a). Let $< I, I^+, I^- >$ be a triplet example, where I and I^+ belong to the same person and I^- is from a different person. Each image is split into four overlapping parts and then fed into the Convolutional Network (ConvNet). Through the ConvNet we map image I into a learned feature space, where the *i-th* part is represented as $\boldsymbol{x}_i$. Then the extracted features are sent into a weighted distance layer, which learns weights for each part and forms a weighted distance for image pair. We train the network using an improved triplet-based loss function, which aims not only to separate the positive pair from the negative by a distance margin, but also to minimise the positive pair distance d^+ and maximise the negative pair distance d^-.

2.2. Four-Branch Convolutional Neural Network

As shown in Fig. 1(b), the ConvNet has four branches to learn specific features for each body part. The branches share the same structure but their weights are not tied. The input image is firstly divided into four overlapping parts and then fed into respective branch. Each branch has two inception modules [16], one convolution layer and one fully connected layer. The first inception module consists of 16 convolution kernels of 5x5 and 16 kernels of 3x3. The output feature maps are then concatenated together with a shortcut from the original input, followed by a 2x2 MaxPooling layer. The second inception module is similar to the first one, except that the convolution kernel sizes are 3x3 and 1x1, respectively. In these two inception modules, kernels of different sizes are adopted to capture features with different resolution. To reduce the depth of the feature maps, we employ another convolution layer with 16 kernels of 1x1. After a 2x2 AveragePooling layer, the fully-connected layer generates an output of 256 dimensions. Note that ReLU is used in all layers.

Overall, through our four-branch ConvNet, each input image is represented as $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4)$, where $\boldsymbol{x}_i$ indicates the 256-D features for the *i-th* body part. These features are then sent into weighted distance layer (see section 2.3). There are only 16 kernels in each convolution layer of our ConvNet, so it has fewer parameters compared with [9,10,15], while it still gains remarkable improvements in accuracy.

2.3. Weighted Distance Layer

Once high level features are obtained, Euclidean distance and triplet loss function are adopted to form the metric-cost part. We found that the previous works [9–11] simply concatenated body-part features into one vector, which can not make full use of the body structure information since different body parts are of different importance. This motivated us to design a novel weighted distance layer. The weighted distance layer fuses body-part features in score level by learning weights for each body-part distance.

Given a triplet $< I, I^+, I^- >$ and its corresponding body-part features, we first perform L2 normalization on $\boldsymbol{x}_i$ to constrain them to live on the 256-D hypersphere, then the square Euclidean distance of the *i-th* body part between I and I^+, I and I^-, can be formulated as:

$$d_i^+ = ||\boldsymbol{x}_i - \boldsymbol{x}_i^+||_2^2, \quad d_i^- = ||\boldsymbol{x}_i - \boldsymbol{x}_i^-||_2^2 \tag{1}$$

We further adopt a weighted combination method to produce the distance between I and I^+, I and I^-:

$$d^+ = \sum_{i=1}^{4} w_i ||\boldsymbol{x}_i - \boldsymbol{x}_i^+||_2^2, \;\; d^- = \sum_{i=1}^{4} w_i ||\boldsymbol{x}_i - \boldsymbol{x}_i^-||_2^2 \\ s.t. \;\; \sum_{i=1}^{4} w_i = 1 \tag{2}$$

where w_i is the weight parameter measuring the importance of the *i-th* body part. Usually, the parameter w_i can be tuned by grid search or random search. However, the extremely long training time of deep learning methods makes it computationally infeasible. That's why we design the weighted distance layer to learn parameter w_i automatically during training. In the weighted distance layer, w_i is updated by standard back propagation algorithm, just the same way as parameters updated in the convolution layer and fully connected layer. Detailed formulas are given in section 2.5.

2.4. Improved Loss Function

Triplet loss is proposed in [14] and it becomes popular in deep metric learning methods. Given a triplet and the corresponding intra-class distance d^+ , inter-class distance d^-, the triplet loss can be defined as:

$$L = max(0, d^+ + m - d^-) \tag{3}$$

where m is a margin that is enforced between positive and negative pairs. This loss serves to separate the positive pair from the negative by a distance margin m. However, the triplet loss function does not constrain how close the positive pairs or how far the negative pairs should be. It may lead to a situation that both d^+ and d^- are small (or large), even when the margin m has been ensured. This will slow down the convergence and drop the re-id performance.

To solve this problem, we add a term to the triplet loss function inspired by the contrastive loss [17]. The new loss function is defined as:

$$\hat{L} = L + \lambda(d^+ + max(0, t - d^-)) \tag{4}$$

where λ is a trade-off parameter. To avoid overfitting, a threshold t is added to the negative pair distance d^- , so that when d^- is larger than t, it makes no contribution to the loss. It should be noted that even when d^- is larger than d^+ by a predefined margin, i.e. $L = 0$, the learning process will continue. The secret lies in the second term $d^+ + max(0, t - d^-)$, which contributes to the loss when $L = 0$. Hence, this new loss function helps to speed up convergence.

In summary, the new loss function serves not only to separate the positive pair from the negative pair by a distance margin, but also to minimise positive pair distance and maximise negative pair distance.

2.5. Optimization

As a variant of triplet loss, our improved loss function $\hat{L}$ is convex, which can be optimized by the back propagation algorithm. Its gradients with respect to $\boldsymbol{x}_i, \boldsymbol{x}_i^+, \boldsymbol{x}_i^-$ are:

$$\frac{\partial \hat{L}}{\partial \boldsymbol{x}_i} = 2w_i(\boldsymbol{x}_i^- - \boldsymbol{x}_i^+) I_{d^+ + m - d^- > 0} + 2\lambda w_i(\boldsymbol{x}_i - \boldsymbol{x}_i^+) \\ + 2\lambda w_i(\boldsymbol{x}_i^- - \boldsymbol{x}_i) I_{t - d^- > 0} \tag{5}$$

$$\frac{\partial \hat{L}}{\partial \boldsymbol{x}_i^+} = 2w_i(\boldsymbol{x}_i^+ - \boldsymbol{x}_i)(\lambda + I_{d^+ + m - d^- > 0}) \tag{6}$$

$$\frac{\partial \hat{L}}{\partial \boldsymbol{x}_i^-} = 2w_i(\boldsymbol{x}_i - \boldsymbol{x}_i^-)(I_{d^+ + m - d^- > 0} + \lambda I_{t - d^- > 0}) \tag{7}$$

For updating the body-part weights w_i, the gradient with respect to w_i is computed by:

$$\frac{\partial \hat{L}}{\partial w_i} = d_i^+(\lambda + I_{d^+ + m - d^- > 0}) \\ - d_i^-(\lambda I_{t - d^- > 0} + I_{d^+ + m - d^- > 0}) \tag{8}$$

In the above equations, the indicator function $I_{condition} = 1$ if $condition$ is true; otherwise $I_{condition} = 0$. With the above derivations, the loss function in Eq.(4) can be easily integrated in back propagation during training.

3. EXPERIMENTS

In this section, we evaluate our method on CUHK01 [18] and PRID2011 [19] datasets. Based on the single-shot setting, we report the experimental results in the form of the average Cumulated Matching Characteristic (CMC) curve for 10 trials. The proposed method is implemented based on the Keras deep learning Framework[1] and our codes are open source[2].

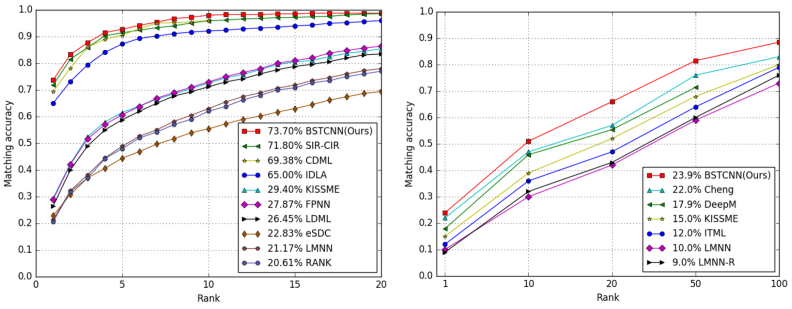
3.1. Experimental Settings

Datasets and Evaluation Protocol

CUHK01 dataset[3] contains 971 persons captured from two disjoint camera views. Each person has two images per camera view. Following the protocol of [8], we use 871 persons for training and the rest 100 persons for testing. $\lambda = 0.1, m = 0.2, t = 0.6$ is set for this dataset.

PRID2011 dataset[4] contains images recorded by two cameras. Camera view A and B contain 385 and 749 persons, respectively, with 200 persons appearing in both views. Following the protocol of [10], we randomly select half of the persons for training and half for testing. $\lambda = 0.2, m = 0.2, t = 0.6$ is set for this dataset.

[1] https://github.com/fchollet/keras
[2] https://github.com/wepe/BSTCNN
[3] http://www.ee.cuhk.edu.hk/~rzhao/
[4] https://lrs.icg.tugraz.at/datasets/prid/



(a) CUHK01 (b) PRID2011

Fig. 2: The CMC curves of different methods on (a) CUHK01 and (b) PRID2011.

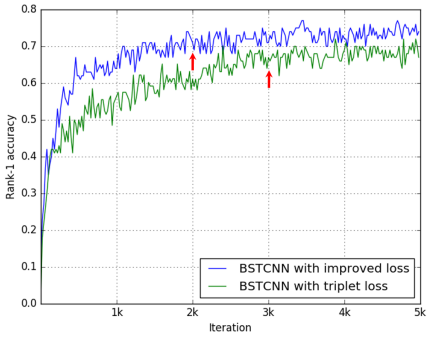


Fig. 3: Rank-1 accuracies vs Iteration with our improved loss and original triplet loss.

Data augmentation. Person re-id datasets are relatively small. To address the over-fitting problem, we augment the training sets by rotating each pedestrian image. Specifically, we rotate the original image with a random degree between [-15° , 15°]. We also horizontally flip each image.

Parameters setting. The weights of the ConvNet are initialized using the method proposed in [20], i.e. he_normal. Each of the four body-part weights is uniformly initialized to 0.25. As for the triplet selection, we adopt online mini-batch strategy which performs both moderate positive [9] and semi-hard negative mining [14]. Specifically, for each iteration, we randomly select 40 persons from the training set which generates 300 triplets in total. The training converge within 6K iterations with Adam gradient decent and it takes roughly 6-10 hours with a NVIDIA GeForce GTX 1080 GPU.

3.2. Experimental Results

Comparison with state-of-the-arts. Experiments are conducted on CUHK01 and PRID2011 datasets to compare our BSTCNN with some state-of-the-art methods, including F-PNN [7], ITML [21], LMNN [22], LDML [23], eSDC [24], KISSME [4], DeepM [11], CDML [9], SIR-CIR [12], IDLA [8] and the work by Cheng [10].

Fig.2 (a) illustrates the CMC curves and the rank-1 accuracies of these methods on CUHK01 dataset. We can see that the rank-1 accuracy of the proposed method reaches 73.7%, which is 1.9% higher than the previous best performance method SIR-CIR [12].

Fig.2 (b) illustrates the CMC curves and the rank-1 accuracies of these methods on PRID2011 dataset. Our method also outperforms other state-of-the-art methods, with a 23.9% rank-1 accuracy. Furthermore, our rank-20 accuracy reaches 66.0%, which is a remarkable improvement.

Evaluation of the improved loss function. To evaluate the effectiveness of the improved loss function, we compare the training process of using original triplet loss and using our improved loss on the CUHK01 dataset. Except the loss function, other settings are the same. As depicted in Fig.3, our model using original triplet loss converges after about 3k iterations and finally reaches about 70.0% rank-1 accuracy. When using our improved loss function, it converges after about 2k iterations and finally reaches about 75.0% rank-1 accuracy. This experimental result indicates that the improved loss function helps to speed up convergence and boost the performance.

Table 1: Body-part importance on CUHK01 and PRID2011.

Dataset	w_1	w_2	w_3	w_4
CUHK01	0.284	0.271	0.247	0.198
PRID2011	0.274	0.250	0.245	0.231

Analysis of the body-part importance. When the training converged, we can output the body-part weights w_1,w_2,w_3,w_4. We average them for 10 trials and the results are shown in Table 1. It is interesting to observe that both on CUHK01 and PRID2011 datasets, body part 1 gains the largest weight, and the value gradually decreases from body part 1 to part 4. This result is reasonable since body part 1 includes face and shoulders, where more discriminative features can be extracted. As we move down the body, less reliable features can be extracted so they contribute little to the person re-id task. This result prove the effectiveness of our weighted distance layer.

4. CONCLUSIONS

This paper introduces an improved triplet convolution neural network for person re-identification. Our network contains two novel elements: a weighted distance layer that learns weights for each body part, an improved loss function that speeds up convergence and boosts the performance. Experimental results demonstrate that our approach achieves better performance than the state-of-the-art methods both on CUHK01 and PRID2011 datasets. In the future, we will investigate other ways to utilize the body structure information, and study how to accelerate convergence of the triplet convolutional neural network.

5. REFERENCES

[1] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li, "Salient color names for person re-identification," *in Proceedings of ECCV*, pp. 536–551, 2014.

[2] Ziming Zhang, Yuting Chen, and Venkatesh Saligrama, "A novel visual word co-occurrence model for person re-identification," *in Proceedings of ECCV*, pp. 122–133, 2014.

[3] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Learning mid-level filters for person re-identification," *in Proceedings of CVPR*, pp. 144–151, 2014.

[4] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," *in Proceedings of CVPR*, pp. 2288–2295, 2012.

[5] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," *in Proceedings of CVPR*, pp. 3594–3601, 2013.

[6] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," *in Proceedings of ICCV*, pp. 2528–2535, 2013.

[7] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," *in Proceedings of CVPR*, pp. 152–159, 2014.

[8] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," *in Proceedings of the CVPR*, pp. 3908–3916, 2015.

[9] Hailin Shi, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Yang Yang, and Stan Z Li, "Constrained deep metric learning for person re-identification," *arXiv preprint arXiv:1511.07545*, 2015.

[10] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," *in Proceedings of CVPR*, pp. 1335–1344, 2016.

[11] Dong Yi, Zhen Lei, Shengcai Liao, Stan Z Li, et al., "Deep metric learning for person re-identification.," *in Proceedings of ICPR*, pp. 34–39, 2014.

[12] Faqiang Wang, Wangmeng Zuo, Liang Lin, David Zhang, and Lei Zhang, "Joint learning of single-image and cross-image representations for person re-identification," *in Proceedings of CVPR*, 2016.

[13] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," *IJPRAI*, vol. 7, no. 04, pp. 669–688, 1993.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," *in Proceedings of CVPR*, pp. 815–823, 2015.

[15] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng, "An enhanced deep feature representation for person re-identification," *in Proceedings of WACV*, pp. 1–8, 2016.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," *in Proceedings of CVPR*, pp. 1735–1742, 2006.

[18] Wei Li, Rui Zhao, and Xiaogang Wang, "Human reidentification with transferred metric learning," *in Proceedings of ACCV*, pp. 31–44, 2012.

[19] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *in Proceedings of ICCV*, pp. 1026–1034, 2015.

[21] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon, "Information-theoretic metric learning," *in Proceedings of ICML*, pp. 209–216, 2007.

[22] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," *NIPS*, pp. 1473–1480, 2005.

[23] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? metric learning approaches for face identification," *in Proceedings of ICCV*, pp. 498–505, 2009.

[24] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," *in Proceedings of CVPR*, pp. 3586–3593, 2013.