Multi-Scale Triplet CNN for Person Re-Identification

Jiawei Liu[1], Zheng-Jun Zha[1], Qi Tian[2], Dong Liu[1], Ting Yao[3], Qiang Ling[1], Tao Mei[3]

[1]University of Science and Technology of China, China
[2]University of Texas at San Antonio, USA
[3]Microsoft Research Asia, China

{ljw368}@mail.ustc.edu.cn, {zhazj,dongeliu,qling}@ustc.edu.cn
{qi.tian}@cs.utsa.edu, {tiyao,tmei}@microsoft.com

ABSTRACT

Person re-identification aims at identifying a certain person across non-overlapping multi-camera networks. It is a fundamental and challenging task in automated video surveillance. Most existing researches mainly rely on hand-crafted features, resulting in unsatisfactory performance. In this paper, we propose a multi-scale triplet convolutional neural network which captures visual appearance of a person at various scales. We propose to optimize the network parameters by a comparative similarity loss on massive sample triplets, addressing the problem of small training set in person re-identification. In particular, we design a unified multi-scale network architecture consisting of both deep and shallow neural networks, towards learning robust and effective features for person re-identification under complex conditions. Extensive evaluation on the real-world Market-1501 dataset have demonstrated the effectiveness of the proposed approach.

Keywords: Person re-identification; Deep CNN

1. INTRODUCTION

Person re-identification is a pressing demand in automated video surveillance and attracts increasing attention from academia and industry. Given an image of a certain person taken from one camera, the task of re-identification is to identify the person images taken from different cameras [1, 2]. A person re-identification system could save a lot of human labour in exhaustively searching for a target person from a large number of video sequences.

Despite the encouraging progress in person re-identification, this task is still very challenging and remains unsolved due to camera setting variations, human pose changes, illumination changes as well as background clutter and occlusions [3]. In addition, persons may share similar visual appearance, making the re-identification more difficult. Figure 1 illustrates the matching of a probe image of a target person against a set of gallery images as well as several kinds of image variations.

Figure 1. Illustration of (a)matching a probe image against gallery images. (b) image variations (left to right): low resolution, human pose changes, camera view changes, occlusions, illumination variation, and persons share similar appearance.

To address these challenges, existing research focus on the development of effective features to describe visual appearance of persons and/or an appropriate metric to measure the similarity among person images. A number of handcraft low-level features have been proposed to tackle the various image variations, including color histogram [4, 5, 6, 7], Local Binary Pattern (LBP) [8], Gabor feature [6], and SIFT [9], etc. Among these features, color representation is an important cue in person re-identification. Many sophisticated features have been developed in recent years to boost the re-identification performance. For example, Yang et al. [10] proposed a sophisticated color descriptor, termed salient color names. Farenzena et al. [11] proposed the Symmetry-Driven Accumulation of Local Features (SDALF) to exploit the symmetry property of human body to handle variations of camera views. Gary and Tao [12] proposed to select optimal descriptors from color and texture features by AdaBoost. Zhao et al. [13] exploited salience information and emphasized rare colors by giving them higher weights in matching. Ma et al. [14] and Zheng et al. [15] encoded local features to a final global representation. Song et al. [16] proposed to learn human attributes for person re-identification. Liao et al. [17] analyzed the horizontal occurrence of local features and maximized the occurrence to improve the robustness of features.

Recently, deep learning has obtained great successes in various computer vision tasks [18]. It has shown great ability for learning inconceivable and effective visual representation. However, there are few works that exploit deep

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2967209

learning technique to facilitate person re-identification. Li et al. [19], Yi et al. [20], and Ahmed et al. [21] used a Convolutional Neural Network (e.g., "Siamese" network) to learn visual features from pedestrian image pairs. However, the number of generated negative pairs greatly outnumber the positive pairs. This may enforce the trained model biased towards negative pairs. And these networks were all designed with a relatively shallow architecture, which can not well exploit the effectiveness of deep learning.

Based on visual features, some existing works focus on learning an appropriate distance/similarity metric to compare the features across person images. The metric is expected to make the images of the same person close in feature space and separate the images from different persons far away from each other. For example, Dikmen et al. [22] proposed to learn a Mahalanobis distance metric through a maximum margin formulation. Sareesh et al. [23] learnt a distance metric by a Local Fisher Discriminant Analysis (LFDA) method. Martin et al. [24] introduced a simple, but effective strategy to learn a distance metric from equivalence constraints. Mignon et al. [5] developed a Pairwise Constrained Component Analysis (PCCA) method, which derives a projection from the original feature space into a low-dimensional space.

In this paper, we propose a multi-scale triplet convolutional neural network which captures visual appearance of a person at various scales. We design a unified multi-scale network architecture consisting of both deep and shallow neural networks, towards learning robust and effective features for person re-identification with various variations. We propose an integrated "end-to-end" learning procedure to learn deep features and similarity metric simultaneously. We optimize the network parameters based on a comparative similarity loss on massive sample triplets, addressing the problem of small training set in person re-identification. Extensive experiments on the real-world Market1501 data set have shown that the proposed approach achieves better performance than the state-of-the-art methods.

2. THE PROPOSED APPROACH

2.1 Comparative Similarity Loss

Person re-identification can be viewed as a person image matching problem. The objective is to learn an effective representation and an appropriate similarity metric to match images from the same person and distinguish images from different persons. Let $\mathcal{X} = \{x_i^p \mid i = 1, 2, ..., N\}$ denote a training set, where x_i^p represents the p-th image of i-th person and N is the number of pedestrians. For a probe image x of a person of interest, person re-identification system searches for the images from the same person by matching the probe image against a set of gallery images $\{\mathcal{G}\}$. The visual features and similarity metric are expected to give higher similarity score to the pair of images $\{x_i, x_i^+\}$ from the same person over the the pair $\{x_i, x_i^-\}$ from different persons. Hence, we design a comparative similarity loss on image triplets as follows:

$$L_{triplet} = \sum_{(x_i, x_i^+, x_i^-) \in \chi} \max\left(1 - \frac{\|f_w(x_i) - f_w(x_i^-)\|_2^2}{\|f_w(x_i) - f_w(x_i^+)\|_2^2 + C}, 0\right) \quad (1)$$

where $f_w(x)$ is the feature representation from a CNN model, which projects the images from raw pixel space into a feature space $\mathbb{R}^d$. The constant C is a predefined margin. The comparative triplet loss aims to enforce the distance among images from different persons larger than the images from the same person.

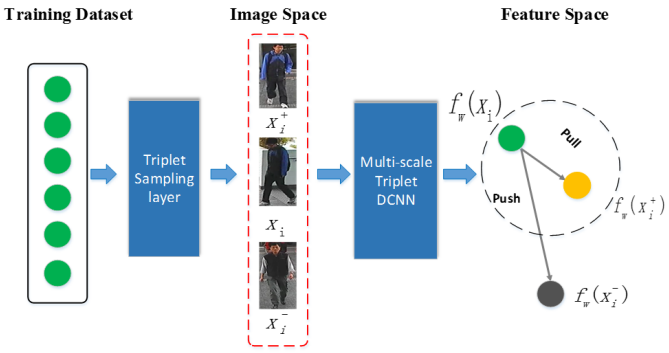

Figure 2. The minimization of comparative similarity loss on sample triplets leads to effective Deep CNN that generates similar features for the same person and dissimilar features for different ones.

The above triplet loss dose not constrain the matching pair $(f_w(x_i), f_w(x_i^+))$ close to each other. As a result, the images belonging to the same person may form a cluster with large intra-class divergence in the learnt feature space. In other words, the learnt feature may be susceptible to the variations caused by illumination, human pose, camera view and occlusions, in visual matching. To address this problem, we design a loss term on image pairs to further constrain the distance among the intra-class image features as follows:

$$L_{pair} = \sum_{(x_i, x_i^+) \in \chi} \|f_w(x_i) - f_w(x_i^+)\|_2^2 \quad (2)$$

By combining the above triplet and pair loss functions, the comparative similarity loss is obtained as follows:

$$L = \sum_{(x_i, x_i^+, x_i^-) \in \chi} \max\left(1 - \frac{\|f_w(x_i) - f_w(x_i^-)\|_2^2}{\|f_w(x_i) - f_w(x_i^+)\|_2^2 + C}, 0\right) + \mu \sum_{(x_i, x_i^+) \in \chi} \|f_w(x_i) - f_w(x_i^+)\|_2^2 \quad (3)$$

where μ is the weight parameter between the triplet and pair loss. This loss function together with the designed Deep CNN model as shown in Figure 2 is able to learn robust and effective visual features for person re-identification.

2.2 Multi-Scale Network

The proposed multi-scale triplet network is illustrated in Figure 3, consisting of three sub-networks: one deep convolutional neural network and two shallow networks. Due to the small scale of pedestrian image corpus, the traditional deep CNN model requiring a large amount of training samples can not be directly applied to person re-identification task. Therefore, we propose a triplet network architecture which shares same parameters with the proposed comparative loss on massive sample triplets instead of the traditional softmax loss. Given a training set $\{\chi\}$, large number of sample triplets can be formed as the network input addressing the problem of limited training samples. Each sample triplet

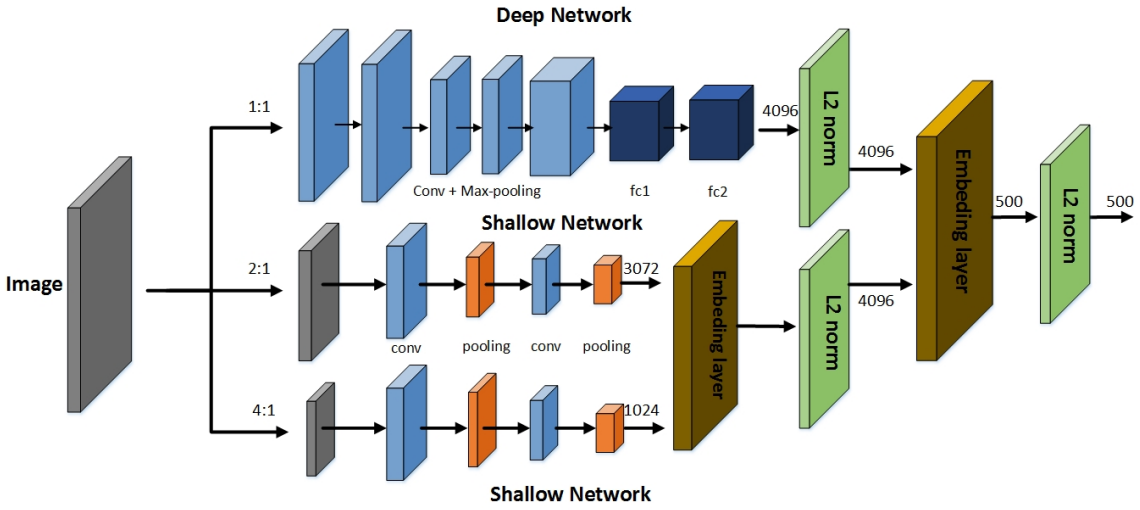

Figure 3. The multi-scale network architecture consist of one deep network and two shallow networks, which capture visual appearance of a person at various scales. The responses from the deep and shallow networks are embedded at an embedding layer to generate final feature representation.

includes a probe image x_i, an image of the same person x_i^+, as well as an image from a different person x_i^-.

As the Alex network [18] has shown good performance in many compute vision tasks, we design the deep CNN according to Alex network. The two shallow networks take down-sample images with the rates of 2:1 and 4:1 as input respectively. The shallow networks can produce less invariance and low-level appearance features from images. After normalizing the outputs from the three networks with L_2 norm, the three feature vectors are fused into an embedding layer, leading to a final 500-dimensional feature vector from the entire network. The features from the deep and shallow networks capture image content at different scales and have complementary advantages. The fused feature is more robust and effective for representing pedestrian appearance under various conditions. In particular, the deep network contains five convolutional layers, five max-pooling layers, two local normalization layers, and three fully-connection layers. Each shallow network contains two convolutional layers followed by two pooling layers.

2.3 Training Strategy

We optimize the parameters of our CNN model by using the mini-batch stochastic gradient descent algorithm. The collection of training sample triplets is divided into a number of mini-batches of triplets. Given each mini-batch, the training errors are calculated based on the proposed loss function and the derivatives are obtained by back-propagation along the networks to update the parameters at each layer. The multi-scale triplet CNN model can be represented as follows:

$$f(x) = h_n(h_{n-1}(h_{n-2}(\cdots h_1(x)\cdots))) \quad (4)$$

where the h_i is the transfer function of the ith layer. The parameters of the transfer function are denoted as w_i. Given the objective function, we calculate the gradient of the loss with respect to the parameters at each layer by the chain rule as follows:

$$\frac{\partial T}{\partial w_l} = \frac{\partial T}{\partial f_w(x_i)} \times \frac{\partial f_w(x_i)}{\partial g_n} \times \frac{\partial g_n}{\partial g_{n-1}} \times \cdots \times \frac{\partial g_{l+1}}{\partial g_l} \times \frac{\partial g_l}{\partial w_l} \quad (5)$$

It should be note that the gradients $(\frac{\partial T}{\partial f_w(x_i)})$ of the loss for different samples in a triplet (x_i, x_i^+, x_i^-) are different.

3. EXPERIMENT

Our multi-scale triplet convolutional neural network was implemented based on the **Caffe** [1] framework. The overall network was trained using two NVIDIA K40 GPUs, Intel i7 CPU, and 32G memory.

Datasets - There are multiple pubic data sets established for person re-identification, including **Market1501, CUHK Campus, VIPeR,** and **PRID2011** etc. Among them, **Market1501** is the most challenging and realistic one. **Market1501** was collected from a total of six cameras, placed in front of campus supermarket. It contained 32,643 bounding boxes of 1,501 identities with Deformable Part Model (DPM) detector. Each identity was captured by six cameras at most and two cameras at least. Following to [14], the data set is divided into two parts. One part has 750 identities as training set and the other has 751 identities as testing set, in which there are 3,363 query images belong to the 750 identities, and each query has 14.8 ground-truth images in average in the gallery set.

Training Setting - The trained Caffe model was used to initialize the weights of our deep sub-network. The two shallow networks were trained from scratch. We started the stochastic gradient descent (SGD) method with learning rate of 0.001 and the weight decay of 0.0005. The parameters μ, C in Eq.(3) are set to 0.002 and 0.008 respectively in the experiments. Each mini-batch contains 160 sample triplets. The entire framework was trained for 150,000 iterations in total.

[1] http://caffe.berkeleyvision.org/

Performance on the Market1501 dataset - We compare our approach with several state-of-the-art methods, including Bag of words model[14], SDALF [10], eSDC[15], and XQDA[16]. Experimental results are shown in Table 1. It can be seen that our approach achieves the best performance among all the methods. Given a single query image, our approach achieves 45.1% rank-1 recognition rates and improves SDALF[10], eSDC[13], BOW(singleQ)[14], and XQDA[16] by 109.8%, 40.1%, 26.0%, and 3.0%, respectively. Given multiple query images, our approach outperforms existing BOW(MutiQ) method[14] with significant performance improvement in terms of rank-1 recognition rate. It can be also observed that the performance of our approach with multi-scale network performs better than that with a single Alex network. This demonstrates the effectiveness of the combination of deep and shallow networks on person re-identification task.

Table 1. Performance comparison with the state-of-the-art methods on the Market-1501 dataset

Method	Rank-1	Rank-5	Rank-10	Rank-30
BOW (singleQ)**[14]**	35.8	52.4	60.3	71.9
BOW (multiQ)**[14]**	44.4	60.2	66.5	76.2
SDALF[10]	21.5	34.5	42.3	50.7
eSDC[13]	32.2	47.1	55.2	64.1
XQDA[16]	43.8	-	-	-
Our Method (sin-network)	42.3	68.5	76.2	84.6
Our Method (singleQ)	45.1	70.1	78.4	88.7
Our Method (multiQ)	**55.4**	**78.9**	**85.6**	**93.7**

4. CONCLUSION

In this paper, we proposed a multi-scale convolutional neural network for person re-identification. A multi-scale triplet network architecture integrating deep and shadow networks was developed to capture visual appearance of a person at various scales and learn robust and effective features under complex conditions. A network learning method was proposed to optimize network parameters based on a comparative similarity loss over massive sample triplets. We have conducted extensive experiments on the real-world Market1501 data set and the experimental results have shown that the proposed approach outperforms the state-of-the-art methods.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Contract 61472392 and 61429201. This work was also supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.

6. REFERENCES

[1] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.

[2] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*, volume 1. Springer, 2014.

[3] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015.

[4] Alexis Mignon and Frédéric Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2666–2672. IEEE, 2012.

[5] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *Computer Vision–ECCV 2014*, pages 1–16. Springer, 2014.

[6] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668, 2013.

[7] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *Image Processing, IEEE Transactions on*, 23(8):3656–3670, 2014.

[8] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, 2013.

[9] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[10] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *Computer Vision–ECCV 2014*, pages 536–551. Springer, 2014.

[11] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[12] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[13] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535, 2013.

[14] Bingpeng Ma, Yu Su, and Frédéric Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 413–422. Springer, 2012.

[15] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[16] Xiao Liu, Mingli Song, Qi Zhao, Dacheng Tao, Chun Chen, and Jiajun Bu. Attribute-restricted latent topic model for person re-identification. *Pattern recognition*, 45(12):4204–4213, 2012.

[17] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Dong Yi, Zhen Lei, and Stan Z Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014.

[20] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[21] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[22] Mert Dikmen, Emre Akbas, Thomas S Huang, and Narendra Ahuja. Pedestrian recognition with a learned metric. In *Computer Vision–ACCV 2010*, pages 501–512. Springer, 2010.

[23] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3318–3325, 2013.

[24] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.