MULTI-PART COMPACT BILINEAR CNN FOR PERSON RE-IDENTIFICATION

Jian Liu, Zhen Yang, Tao Zhang, Huilin Xiong

School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China
Institute for Sensing and Navigation, Shanghai Jiao Tong University, China

ABSTRACT

In paper, we present a novel multi-part compact bilinear convolutional neural network (CNN) model, which consists of a bilinear CNN and two part-networks aiming to learn the global features and the finer local features simultaneously. The bilinear operation is simplified with recently proposed compact bilinear pooling method, and bilinear vectors are averagely pooled to keep more local spatial information. The proposed model is trained by using a histogram loss function in order to reduce the distribution overlap of positive pairs and negative pairs. Experiments show that, the combination of compact bilinear CNN and histogram loss can significantly improve the original models, and performs favorably compared to the state of the art.

Index Terms— Bilinear CNN, multi-part, compact bilinear pooling, histogram loss

1. INTRODUCTION

Person re-identification aims to correctly match images of the same person taken from non-overlapping cameras or one single camera across time. It has been widely used in video surveillance and human-computer interaction, etc. However, it is still a challenging task, because varying illumination, viewpoints, poses and occlusions of body can make two images of different persons more similar than the same person.

In more detail, the person re-identification includes two aspects: extraction of discriminative features from person images, and design of the distance metric for comparing these features. To handle the imaging variations caused by different viewpoints, poses and illumination, many robust descriptors have been developed, including color histograms [1,2,3], local binary patterns [4,5], color names [6,7], and Gabor features [8], etc. Apart from these low-level features, mid-level semantic attributes have also been used to describe person [9]. Once the person features are obtained, similarity metric learning and ranking algorithms are applied, where the metric learning aims to map the feature space to distance space in which the same person is closer than different ones. There are many metric learning approaches in the literature, such as Mahalanobis Metric Learning (KISSME) [10], Local Fisher Discriminant Analysis (LFDA) [11], Marginal Fisher Analysis (MFA) [11], Locally Adaptive Decision Functions (LADF) [12], Large Margin Nearest Neighbor (LMNN) [11], and attribute consistent matching [13], etc.

Deep learning has also been applied to person re-identification, in view of its great success in various computer vision and pattern recognition tasks. Yi *et al.* [14] and Li *et al.* [15] first apply deep learning to solve the problem of person re- identification. To handle the problem of insufficient training data in person re-identification, image pairs or triplets are usually employed to compute the loss. In Yi's method, each input image is partitioned into three overlapping parts which pass through the part-specific convolutional network, and the outputs are connected to form the final representation. The representations of two images are further compared using a cosine similarity metric. Ahmed *et al.* [16] improve their deep learning architecture by computing cross-input neighborhood differences. Varior *et al.* [17] use the long short-term memory (LSTM) to learn the dependence between local regions, aiming to fully exploit the spatial relationships. Ustinova *et al.* introduce Bilinear CNN [18] to learn discriminative descriptors and histogram loss [19] to train deep network respectively. In this paper, we present a multi-part convolutional neural network model by combining a bilinear CNN and two part-networks to learn the global features and finer local features simultaneously. Experiment results show that the proposed model works favorably in comparison to the state-of-the-art.

2. RELATED WORK

2.1. Compact bilinear CNN

Bilinear CNN, introduced by Lin *et al.* [20], achieved satisfactory results on several fine-grained recognition and face recognition datasets. It consists of two CNNs to extract features. These features are multiplied using the outer product at each spatial location and pooled locally to obtain an image descriptor. They give a hint that two separated pathways may correspond to part and texture detector, and the bilinear operation can model local pairwise feature interactions, which partially explains the effectiveness of bilinear CNN.

However, bilinear features are typically of high dimension, which cost much computation and storage. Gao *et al.* [21] propose two compact bilinear pooling methods to reduce the feature dimensionality. Bilinear descriptor is obtained by comparing two image descriptors at every location and the comparison operator can be seen as a second order polynomial kernel, which performs discriminatively in linear classifier. In Gao's method, polynomial kernel is approximated by two low-dimensional methods: Random Maclaurin (RM) [22] and Tensor Sketch (TS) [23].

TS outperforms RM provided the output dimension is not extremely low, so we take TS as the approximation method in our experiment. Let $x \in R^c$ denote the feature, $f(x) \in R^d$ denote the mapped feature and they typically satisfy that $d \ll c^2$. TS first generates random and uniform distributed $h_k \in N^c$ and $s_k \in N^c$ where $h_k(i) \in \{1,2,\dots,d\}$, $s_k(i) \in \{+1,-1\}$ and k = 1, 2. Count Sketch is defined as $Cx(x,h,s) = \{(Cx)_1,\dots,(Cx)_d\}$, where $(Cx)_j = \sum_{i:h(i)=j} s(i)x_i$. Finally $f(x)$ is defined as $f(x) = FFT^{-1}(FFT(Cx(x,h_1,s_1)) \circ FFT(Cx(x,h_2,s_2)))$, where $\circ$ denotes element-wise multiplication.

2.2. Histogram loss

Histogram loss, achieved state-of-the-art results in person re-identification datasets and competitive results in other two datasets. The loss function was designed to avoid tuning hyper parameters such as margins or thresholds which largely affect result quality of deep network. Histogram loss is computed in two stages. Firstly, two distributions (histograms) of similarities corresponding to matching pairs and non-matching pairs are estimated. The similarity of two images is measured by scalar product. Secondly, the overlap of the two distributions, meaning the similarity of a positive pair is lower than a negative pair, is computed. The only tunable parameter is the number of bins in the histograms, which has little effect on performance. Let $X = \{x_1, x_2, \dots, x_N\}$ denote the batch of image representations and m_{ij} mark the relationship between example x_i and x_j. If x_i and x_j correspond to the same person then m_{ij} is $+1$, otherwise m_{ij} is -1. The sample distributions are $S^+ = \{s_{ij} = \langle x_i, x_j \rangle \mid m_{ij} = +1\}$ and $S^- = \{s_{ij} = \langle x_i, x_j \rangle \mid m_{ij} = -1\}$ where s_{ij} is bounded to $[-1;+1]$. The distributions are then matched to R-dimensional histograms H^+ and H^-, which are uniformly filled by nodes $t_1 = -1, t_2, \dots, t_R = +1$ with step $\Delta = \frac{2}{R-1}$. The node value h_r^+ of histogram H^+ is computed as:

$$h_r^+ = \frac{1}{|S^+|} \sum_{(i,j):m_{ij}=+1} \delta_{i,j,r} \tag{1}$$

Each pair similarity should be assigned to nodes adjacent to it, so the weights $\delta_{i,j,r}$ are defined as:

$$\delta_{i,j,r} = \begin{cases} (s_{ij} - t_{r-1})/\Delta, & if\ s_{ij} \in [t_{r-1}; t_r] \\ (t_{r+1} - s_{ij})/\Delta, & if\ s_{ij} \in [t_r; t_{r+1}] \\ 0, & otherwise \end{cases} \tag{2}$$

Now we have the histograms H^+ and H^- approximating to the positive pair distributions p^+ and negative pair distributions p^-. Histogram loss is computed as:

$$\mathrm{L} = \sum_{r=1}^{R} \left(h_r^- \sum_{q=1}^{r} h_q^+ \right) \tag{3}$$

The histogram loss is computed based on all the possible pairs in one batch, which is much different from triplet loss or pair loss.

3. THE PROPOSED METHOD

Person re-identification is partly similar to fine-grained recognition task, because different people may look similar from their appearance and can only be distinguished by subtle ingredients. There is already some research borrowing ideas from fine-grained recognition [18]. Ustinova *et al.* split one image to three parts, and pass to three separated convolutional network whose outputs are combined by bilinear operation. Our method mainly differs in that we pass the full image to network and replace the bilinear operation with compact bilinear pooling. Besides, we add two additional part-networks.

3.1. Multi-part compact bilinear CNN

The overall architecture is shown in Fig. 1 which consists of a compact bilinear CNN (CBC) and two part-networks. Hereinafter, we call it multi-part compact bilinear CNN (MCBC). The CBC has two streams, and each contains three convolution layers to extract features. Two stream outputs are combined by compact bilinear operation and average pooled locally. Then a fully connected layer is followed, whose output is fused with the output of two part-networks to form the final representation. In the first convolution layer, we pass RGB image of size 60 × 160 × 3 through 32 learned filters of size 7 × 7. The second convolution layer consists of 64 filters of size 5 × 5. They are all followed by a max-pooling layer to halve the features size. The third convolution layer also consists of 64 filters of size 3 × 3. The stride for these three convolution layers are respectively 2 × 2, 1 × 1 and 1 × 1. Rectified linear unit (ReLU) is the activation function for each convolution layer. We finally get 64 feature maps of size 20 × 8 from each stream. These two representations are combined by compact bilinear operation in each location and the result vector is of 400 dimensions. The feature maps are further average pooled on every 2 × 2 region with stride 2 × 2. The fully

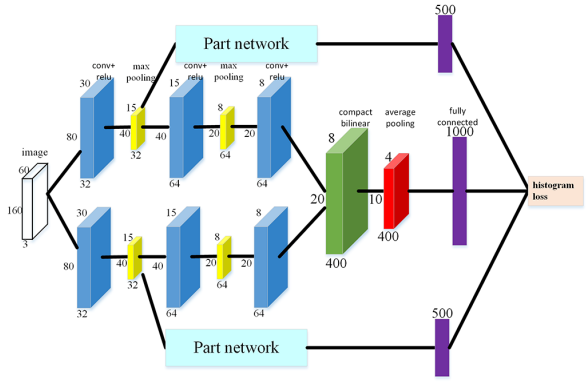

Fig. 1. The architecture of Multi-Part Compact Bilinear CNN.

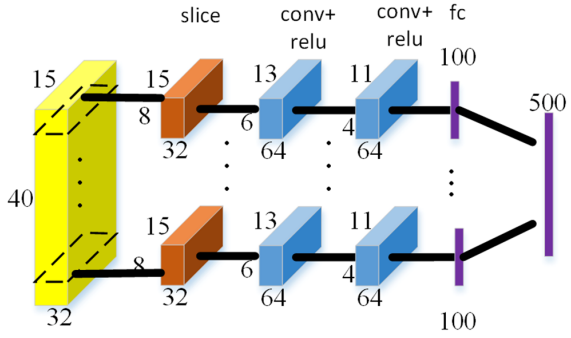

Fig. 2. The architecture of part-network.

connected layer of CBC generates an output of 1000 dimensions. It concatenates two outputs of part-networks and passes to the histogram loss layer.

3.2. Part-network

Person re-identification is often processed on parts or neighborhood [15,24] to keep more spatial information. Inspired by this, we also add additional part-network to learn finer representation which takes the feature maps of the first max-pooling layer as input. The part-network architecture is shown in Fig. 2. The feature maps of the first max-pooling layer of CBC are of size 40 × 15 and they are sliced to 5 equal non-overlapping parts. Each part passes to one sub-network. Each sub-network consists of two convolutional layer and a fully connected layer. To get some finer features, we do not apply max pooling after the convolution. The convolutional layers' channels are both 64. The filter sizes are both 3 × 3; the strides are 1 × 1. And there is no padding. Each convolutional layer is followed by ReLU as the activation function.

4. EXPERIMENTS

4.1. Datasets and evaluation protocols

We conducted experiments on the CUHK03 dataset [15] and only single-shot results are reported. We randomly choose one image for each person from the test set. On CUHK03 dataset, previous researches typically divide 1360 identities into non-overlapping train (1160), test (100) and validation (100) sets. We use the newest dataset that contains 1467 identities, and split it to train (1367) and test (100) set, which may result slightly higher accuracy. Results are averaged on 5 random splits. For each split, results are averaged on 100 random query-gallery sets.

The performance is evaluated with Recall@K metric namely the right match ratio among first K gallery candidates sorted by similarity.

4.2. Implementation details

We implement proposed method using the Caffe [25] deep learning framework. To check the performance of CBC and reveal the contribution of part-network, we implement experiments on two slightly changed models. One is the main structure CBC, directly removing the part-network from the architecture shown in Fig. 1. Another is the overall architecture MCBC. All images are first resized to 160 × 60. The batch size is 256 according to the batch size comparison result in [19]. Images are randomly shuffled according to their labels (identity) in every epoch. We did not do any data augmentation. The starting learning rate is 1e − 4, and decreased to 1e − 5 after 20K iterations. The training is stopped at the 30K iterations. We use Adam [26] for stochastic optimization and set other parameters as: momentums are 0.9 and 0.999, weight decay is 0.0005.

5. RESULTS AND DISCUSSION

We evaluate our method on both CUHK03 Labeled dataset and CUHK03 Detected dataset. Both CBC and MCBC are tested on each dataset.

The Recall@K values for the experiments on CUHK03 dataset are shown in Table 1 and Table 2. On Labeled and Detected dataset, we significantly outperform the related methods namely Multiregion Bilinear DML (MR B-DML) and deep metric learning with histogram loss: 70.28% vs 65.04% and 65.77% in rank 1 respectively on Labeled dataset, and 63.11% vs 60.22% in rank 1 on Detected dataset. When compared to the state-of-the-art approaches, that is Fused Model, our method achieves slightly lower accuracy in rank 1 especially on Labeled dataset: 71.62% vs 72.43%, it even gives a boost of 2.61% in rank 10 and 0.99% in rank 20. MCBC slightly outperforms CBC about 1.5% in rank 1 on Labeled and Detected dataset, which demonstrates the effectiveness of part-network, but the advantage decreases, even falls behind when recall value increases.

Table 1. Results on CUHK03 Detected.

Method	r = 1	r = 5	r = 10	r = 20
LOMO+XQDA[1]	46.00	82.50	88.55	94.25
Ahmed et al. [16]	44.96	76.40	83.47	93.15
Fused Model [28]	**72.04**	--	**96.00**	**98.26**
Varior et al. [29]	68.1	88.1	94.6	--
end-to-end CAN[30]	63.05	82.94	88.17	93.29
MR B-DML [18]	60.22	88.10	94.22	97.02
CBC (ours)	63.11	89.34	94.58	97.83
MCBC (ours)	64.69	**90.28**	95.2	98.02

Table 2. Results on CUHK03 Labeled.

Method	r = 1	r = 5	r = 10	r = 20
LOMO+XQDA [1]	52.20	85.80	92.14	96.25
Ahmed et al. [16]	54.74	86.30	93.88	98.10
Ensembles [27]	62.1	89.1	94.3	97.8
Fused Model [28]	**72.43**	--	95.51	98.40
end-to-end CAN[30]	65.65	91.28	96.29	98.17
MR B-DML [18]	65.04	91.60	95.84	98.21
Ustinova et al. [19]	65.77	92.85	97.62	99.43
CBC (ours)	70.28	**95.24**	**98.25**	**99.71**
MCBC (ours)	71.62	94.86	98.12	99.39

6. CONCLUSION

In this work, we propose a new convolutional neural network, called multi-part compact bilinear CNN. It combines two recently proposed methods: compact bilinear pooling and histogram loss. Besides, we add two part-networks to boost the capability. We largely improve the performance compared to methods used in original paper and most previous works, and performs favorably compared to the state of the art, which demonstrates the effectiveness of our new method.

7. ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation of China under Grant 61673274.

8. REFERENCES

[1] Liao S, Hu Y, and Zhu X, et al, "Person re-identification by local maximal occurrence representation and metric learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197-2206, 2015.

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

[3] Zhao R, Ouyang W, and Wang X, "Unsupervised salience learning for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586-3593, 2013.

[4] Ojala T, Pietikainen M, and Maenpaa T, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971-987, 2002.

[5] Xiong F, Gou M, and Camps O, et al, "Person re-identification using kernel-based metric learning methods," *European conference on computer vision*, Springer International Publishing, pp. 1-16, 2014.

[6] Yang Y, Yang J, and Yan J, et al, "Salient color names for person re-identification," *European Conference on Computer Vision*, Springer International Publishing, pp. 536-551, 2014.

[7] Zheng L, Shen L, and Tian L, et al, "Scalable person re-identification: A benchmark," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116-1124, 2015.

[8] Li W, and Wang X, "Locally aligned feature transforms across views," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3594-3601, 2013.

[9] Layne R, Hospedales T M, and Gong S, et al, "Person Re-identification by Attributes," *Bmvc*, Vol. 2, No. 3, p. 8, 2012.

[10] Koestinger M, Hirzer M, and Wohlhart P, et al, "Large scale metric learning from equivalence constraints," *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, IEEE, pp. 2288-2295, 2012.

[11] Weinberger K Q, and Saul L K, "Distance metric learning for large margin nearest neighbor classification," *Advances in neural information processing systems*, vol. 18, p. 1473, 2006.

[12] Li Z, Chang S, and Liang F, et al, "Learning locally-adaptive decision functions for person verification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3610-3617, 2013.

[13] Khamis S, Kuo C H, and Singh V K, et al, "Joint learning for attribute-consistent person re-identification," *European Conference on Computer Vision*, Springer International Publishing, pp. 134-146, 2014.

[14] Yi D, Lei Z, and Liao S, et al, "Deep metric learning for person re-identification," *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, pp. 34-39, 2014.

[15] Li W, Zhao R, and Xiao T, et al, "Deepreid: Deep filter pairing neural network for person re-identification," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152-159, 2014.

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

[17] Varior R R, Shuai B, and Lu J, et al, "A siamese long short-term memory architecture for human re-identification," *European Conference on Computer Vision*, Springer International Publishing, pp. 135-153, 2016.

[18] Ustinova E, Ganin Y, and Lempitsky V, "Multiregion Bilinear Convolutional Neural Networks for Person Re-Identification," *arXiv preprint arXiv:1512.05300*, 2015.

[19] Ustinova E, and Lempitsky V, "Learning deep embeddings with histogram loss," *Advances in Neural Information Processing Systems*, pp. 4170-4178, 2016.

[20] Lin T Y, RoyChowdhury A, and Maji S, "Bilinear cnn models for fine-grained visual recognition," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449-1457, 2015.

[21] Gao Y, Beijbom O, and Zhang N, et al, "Compact bilinear pooling," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 317-326, 2016.

[22] Kar P, and Karnick H, "Random Feature Maps for Dot Product Kernels," *AISTATS*, Vol. 22, pp. 583-591, 2012.

[23] Pham N, and Pagh R, "Fast and scalable polynomial kernels via explicit feature maps," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 239-247, 2013.

[24] Cheng D, Gong Y, and Zhou S, et al, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335-1344, 2016.

[25] Jia Y, Shelhamer E, and Donahue J, et al, "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, pp. 675-678, 2014.

[26] Kingma D, and Ba J, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] Paisitkriangkrai S, Shen C, and van den Hengel A, "Learning to rank in person re-identification with metric ensembles," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1846-1855, 2015.

[28] Subramaniam A, Chatterjee M, and Mittal A, "Dep Neural Networks with Inexact Matching for Person Re-Identification," *Advances in Neural Information Processing Systems*, pp. 2667-2675, 2016.

[29] Varior R R, Haloi M, and Wang G, "Gated siamese convolutional neural network architecture for human re-identification," *European Conference on Computer Vision*. Springer International Publishing, pp. 791-808, 2016.

[30] H. Liu, J. Feng, and M. Qi, et al, "End-to-end comparative attention networks for person re-identification," *arXiv preprint arXiv:1606.04404*, 2016.