Deep Embedding for Face Recognition in Public Video Surveillance

Guan Wang[1,2], Yu Sun[2], Ke Geng[3](✉), Shengguang Li[1], and Wenjing Chen[4]

[1] First Research Institute of the Ministry of Public Security of PRC,
Beijing 100048, China
[2] School of Information Science and Technology, Beijing Forestry University,
Beijing 100083, China
[3] School of Electronic and Information Engineering, Beihang University,
Beijing 100191, China
gengk@avic.com
[4] School of New Media, Beijing Institute of Graphic Communication,
Beijing 102600, China

Abstract. Face recognition is essential to the surveillance-based crime investigation. The recognition accuracy on benchmark datasets has been boosted by deep learning, while there is still large gap between academic research and practical application. This work aims to identify few suspects from the crowd in real time for public video surveillance, which is a large-scale open-set classification task. The task specific face dataset is built from security surveillance cameras in Beijng subway. The state-of-the-art deep convolutional neural networks are trained end-to-end by triplet supervisory signal to embed faces into 128-dimension feature spaces. The Euclid distances in the embedding space directly correspond to face similarity, which enables real time large scale recognition in embedded system. Experiments demonstrate a 98.92% ± 0.005 pair-wise verification accuracy, which indicates the automatic learned features are highly discriminative and generalize well to new identities. This method outperforms other state-of-the-art methods on the suspects identification task, which fills the application gap in public video surveillance.

Keywords: Face recognition · Convolutional neural networks · Public video surveillance · Triplet loss

1 Introduction

Face recognition has been studied extensively for decades due to its practical applications. The key challenge of face recognition is to develop effective feature representations for reducing intra-personal variations while enlarging inter-personal

Guan Wang and Yu Sun contributed equally to this work.

J. Zhou et al. (Eds.): CCBR 2017, LNCS 10568, pp. 31–39, 2017.
https://doi.org/10.1007/978-3-319-69923-3_4

differences. Recently, deep learning has achieved great success on vision community [1–3], significantly improving the state-of-the-art in classification problems.

The face recognition accuracy on the *de-facto* benchmark Labeled Faces in the Wild (LFW) [4] dataset has been boosted rapidly. The DeepFace [5] uses an ensemble of deep convolutional neural networks (CNNs) trained on 4 million 3D aligned face images spanning 4000 unique identities. It employs the Siamese network architecture to minimize the distance between congruous pairs of faces and maximize the distance between incongruous pairs. The DeepFace is then extended by the DeepID series methods [6–9]. Compared to DeepFace, the DeepID methods only uses 2D affine face alignment. The DeepID1 [6] incorporates multiple CNNs to learn more discriminative features at different face patches. The Deep ID2 [7] employs a Bayesian learning framework to train a metric, multi-task learning using joint identification-verification supervisory signal. The Deep ID2+ [8] updates the CNN architectures with fully connected branches after each convolution layers to generate more joint supervisory signal. The Deep ID3 [9] uses very deep networks inspired by the VGGnet [1] and GoogLeNet [2]. More recently, the Facenet [10] directly embeds faces into compact vectors using CNNs end-to-end trained with triplet distance loss. The Facenet currently achieves state-of-the-art performance in LFW and YouTube Faces DB [11]. The open source project OpenFace [12] is developed based on triplet loss for real-time face recognition. The model achieved competitive accuracy on the LFW benchmark with a small training dataset. The Ministry of Public Security of the People's Republic of China sponsors the skynet project, which deploys surveillance cameras in stations, airports, streets, *etc.* for public security. But the surveillance-based crime investigation heavily depends on human labor. Despite significant recent advances, implementing face recognition in public video surveillance presents serious challenges to current approaches. Models trained on web images of celebrities suffer from low accuracy in practical surveillance scenarios.

To bridge between the literature and the industry, this work employs the deep embedding method for suspects identification in public video surveillance. The modified Inception-ResNet-v2 models are trained on the public surveillance images. The triplet loss is employed as supervisory signal to extract discriminative features. The experiment results demonstrate that the proposed method offers higher accuracy and validation rate with lower false accept value than OpenFace in surveillance scenarios.

2 Materials and Methods

2.1 Task Specification and Face Dataset

The images are captured by public surveillance cameras in Beijing subway. Four cameras are set up at both sides of the subway security check sensor gate, with two cameras on each side. When a person walks through the sensor gate, four images are captured simultaneously by cameras fixed at different angles of the gate, as indicated in Fig. 1a. Figure 1b shows the position of two cameras at one side of the gate. The other two cameras are located at the same position on the other side. The typical 4 images of one identity are shown in Fig. 1c.

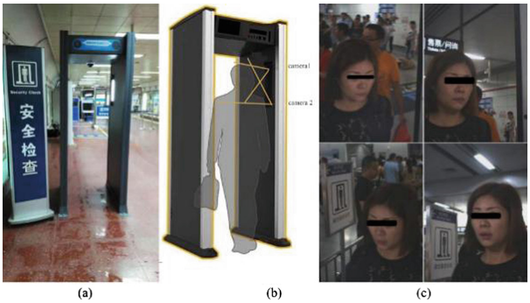

Fig. 1. (a) The face checkpoint at Beijing subway. (b) Position of the two cameras at one side of the gate. (c) Typical images of one identity captured simultaneously by 4 cameras.

The LFW is the *de-facto* testbed for face verification, which contains 13,233 images of 5,750 celebrities [1]. The LFW faces are web images of all human races captured at uncontrolled condition. Compared with LFW, our dataset manifests unique characters: faces are captured at relatively controlled condition, with fixed focal length under indoor illumination from almost the same distance. After human elimination of disqualified images, our dataset includes 1,806 Chinese citizens, ranging from teenage students, adults to aged people. What's more, the images captured in subway suffer from low illumination and high ISO noise.

Following the LFW's unrestricted with labeled outside data evaluation protocol [13], we constructed two test sets. The first test set is a balanced one like LFW, consisting of 1055 congruous face pairs and 1055 incongruous face pairs. The pairs are split into 5 folds. The balanced test set aims to assess the threshold deciding whether a pair is congruous or not, and evaluate the performance of the trained model using the benchmark same as LFW. The second test set is an imbalanced one that contains 6000 pairs. The 6000 pairs are split into 10 folds. All the identities in test set are held out from training. The other identities are left in the training set without pair labels, *i.e.* only labeled by an ID number. Compared with the LFW test set, there are two differences: First, the congruous pairs only account for 1% in the imbalanced test set. In each fold, only one person is the suspect and constitutes 6 congruous pairs. The other 594 pairs are incongruous pairs of the suspect with the other 594 identities. The incongruous pairs are used as varieties of distractors. Second, all images except the ones of the suspects are used only once in each fold. The imbalanced test set aims to evaluate the performance of the model for recognizing few suspects from the crowd. Thus the task is an open set classification problem, which demands that recognition model could generalize to massive identities beyond the training set.

2.2 Deep Embedding with Triplet Loss

Most CNNs use a softmax classification layer at training stage and then take an intermediate bottleneck layer as the representation at test stage, while the indirect bottleneck representation could not generalize to new faces. What's more, the parameters in the classification layer increase exponentially with identities in the training set. To circumvent the above problems, deep embedding trained by triplet loss is employed in our task. The Euclid distance in the embedding space directly corresponds to face similarity. A triplet consists of three face images of two identities: an anchor image x^a, a positive image x^p of the same identity, and a negative image x^n of another identity. The deep CNN followed by L2 normalization embeds each image x into a 128 dimensional hypersphere. The triplet loss is defined as:

$$L = \sum_{i=1}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right], \tag{1}$$

where $f(x)$ is the output embedding of the model, and N is the mini-bath size. As shown in Fig. 2, the training aims to separate the positive pair from the negative by a distance margin α for any triplet.

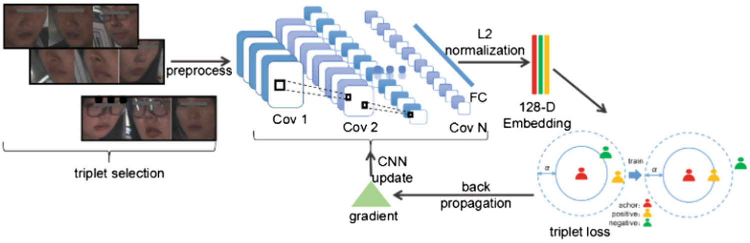

Fig. 2. End to end deep embedding training with triplet loss

As shown in Fig. 2, a deep CNN followed by L2 normalization is directly trained to output a compact embedding supervised by a triplet loss. The triplet loss is minimized by back propagation at training stage. To convergence the training algorithm in feasible time on our dataset, the selected triplets are which violate the margin α as the top of Fig. 2 indicates. For a certain anchor, the hard positive x^p which $\mathrm{argmax}_{x^p}(\|f(x^a) - f(x^p)\|_2^2)$ and the hard negative x^n that $\mathrm{argmin}_{x^n}(\|f(x^a) - f(x^n)\|_2^2)$ should be selected from the training set. Our dataset contains massive identities, while each identity has 4 images at most. So we choose the hard positives across all intra-identity images, while select the hard negatives within a random subset. Each mini-batch of triplets is generated online by the latest network checkpoint. To tradeoff between training efficiency and mini-batch coverage, the hard negatives in one mini-batch are searched in the range of 600 identities.

3 Experiments

3.1 Data Preprocessing

Both the training and test set are detected and aligned by Multi-task CNN (MTCNN) [14]. The original MTCNN select all faces in one image. But each image captured by the public video surveillance often includes more than one person, and the largest face are often far from the image center. So the MTCNN is revised in this paper to ignore the face position and only select the largest face. And then aligned with 5 face landmarks (two eyes, nose and mouth corners) by affine transformation. The faces in the training set are resized to 182 × 182 resolution with a maximal margin of 22 pixels. Before fed into CNNs, the aligned faces are randomly cropped to 160 × 160 resolution, randomly flipped and then normalized to standard distribution. The faces in test set are directly resized to 160 × 160 resolution and normalized to standard distribution.

3.2 Models Architecture

The Inception-ResNet-v2 [15] network is employed for feature extraction, which achieves a state-of-the-art in terms of accuracy on the ILSVRC image classification benchmark [16]. The network combines two most recent ideas, the inception blocks and the residual connections. The model uses three types of inception blocks, inception-a, inception-b, and inception-c. Residual connections allow shortcuts in the model and enable deeper neural networks, which also significantly simplify the inception blocks by reducing parallel towers. We modify the model by stacking different number of inception blocks. Three modified versions of the Inception-ResNet-v2 model are evaluated systemically in this work. For each version, the number of the three types of inception blocks are 1-2-1, 5-10-5 and 10-20-10, respectively. The final softmax layer for close set classification is replaced with a L2-normalization layer for deep embeddings.

Both the MTCNN and the modified Inception-ResNet-v2 models are implemented in the open source deep learning framework TensorFlow 1.0 [17] with TF-Slim library. All the experiments are conducted on an Ubuntu 16.04 Linux server with an Intel i7-6700 CPU with 64 GB memory and one NVIDIA TitanX GPU with 12 GB memory.

3.3 Training Algorithm

Deep neural networks are trained aiming to minimize the triplet loss function L, *i.e.* finding the value of parameters W that minimizes the loss. We use the Adagrad algorithm [18] where W is updated iteratively as:

$$W - = lr * \frac{g}{\sqrt{\sum g^2} + \varepsilon}, \tag{2}$$

where lr is the learning rate, g is the gradient in each iteration and ε is a small value of 10^{-6}.

All the models are trained for 80 epochs or qualified triplets are exhausted. We use an initial learning rate of 0.1, multiplying it by 0.1 every 35 epochs. To avoid model collapse, *i.e.* $f(x) = 0$, we initiate the distance margin value $\alpha = 0$ and increase α progressively. As training progresses, the difficulty of triplets increases accordingly.

3.4 Evaluation Protocol

The method is evaluated on the face verification task. Given a pair of face images (x_i, x_j) of identities (i, j), a squared L2 distance threshold d is used to determine the pair is same or different. The set of true accept TA(d) and false accept FA(d) is define as:

$$\mathrm{TA}(d) = \left\{ (i,j) \in P_{same},\ \text{with} \quad \left\| f(x_i) - f(x_j) \right\|_2^2 \leq d \right\}, \tag{3}$$

$$\mathrm{FA}(d) = \left\{ (i,j) \in P_{diff},\ \text{with} \quad \left\| f(x_i) - f(x_j) \right\|_2^2 \leq d \right\}, \tag{4}$$

where P_{same} denotes the congruous pairs and P_{diff} denotes the incongruous pairs. The validation rate VAL(d), the false accept rate FAR(d) and verification accuracy ACC (d) are defined as:

$$\mathrm{VAL}(d) = \frac{TA(d)}{P_{same}}, \quad \mathrm{FAR}(d) = \frac{FA(d)}{P_{diff}}, \quad \mathrm{ACC}(d) = \frac{\mathrm{TA}(d) + P_{diff} - \mathrm{FA}(d)}{P_{same} + P_{diff}}. \tag{5}$$

3.5 Threshold Learning

The squared L2 distance threshold d is important to the performance of the method. The threshold is estimated using the method in [19], *i.e.* it is learned separately on each fold to maximize the accuracy on the test pairs. As the second test set is highly imbalanced, 5 thresholds are learned on the 5 folds of the first test set, and then the average of the learned thresholds is used on the second test set.

4 Results and Discussion

We use the ACC value on the balanced test set for the comparison of results of the three models. All the three models, 1-2-1, 5-10-5 and 10-20-10, obtain similar high performance, with an accuracy of 92.94%, 93.27% and 94.36% respectively. It is noted that the accuracy slightly improves with the increase of model depth, and the best performance is achieved by the 10-20-10 model. We compare our best model with OpenFace on the two test sets and report the results in Table 1. The Receiver Operating Characteristic (ROC) curves are plotted in Fig. 3. From the results we have the following observations. First, the proposed model improves the accuracy of OpenFace model by a wide margin of 34.74% and 18.77% on the balanced and imbalanced test set, respectively. The ACC of 94.36% ± 0.01 and 98.92% ± 0.005 are achieved on the two test sets, respectively. The excellent performance indicates that our model is robust to the low quality of the images. But the OpenFace model trained on web images cannot generalize well. Second, the VAL value on the imbalanced test set of our model is 98.33% ± 0.05, *i.e.* only one pair images of one suspect is not recognized. Last, a low FAR is obtained by our model, with only 3.77% ± 0.01 and 1.08% ± 0.005 on

Table 1. Comparison with the OpenFace model

Test set	Method	VAL	FAR	ACC
Balanced	OpenFace	70.24% ± 0.08	50.73% ± 0.09	59.62% ± 0.04
	Proposed method	92.48% ± 0.02	3.77% ± 0.01	94.36% ± 0.01
Imbalanced	OpenFace	38.33% ± 0.29	19.43% ± 0.08	80.15% ± 0.08
	Proposed method	98.33% ± 0.05	1.08% ± 0.005	98.92% ± 0.005

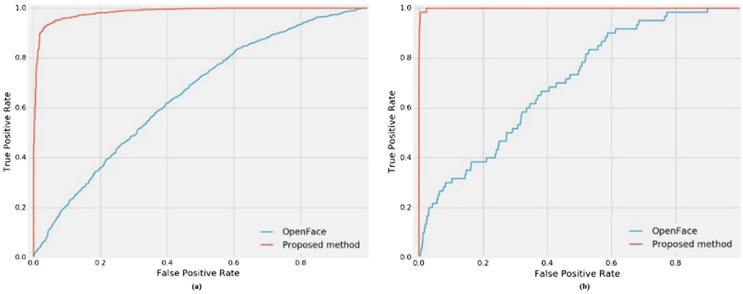

Fig. 3. The ROC curves on the (a) balanced and (b) imbalanced test set

the balanced and imbalanced test set, respectively. The varieties of distractors have very limited influence on our model.

Though OpenFace is developed for mobile applications, the model has not been implemented on any embedded system. Our trained models are deployed to the Jetson TX2 embedded system. The time taken on the device to process 100 pairs of images for the 1-2-1, 5-10-5 and 10-20-10 models are 13.75, 22.36 and 65.10 s respectively, with the accuracy consistent with that on the x86 workstation. Thus the face recognition can be conducted offline in real time at a low electrical power. The model with the best trade-off between accuracy and time cost could be selected according to practical application scenarios.

5 Conclusion

In this work, the first Chinese citizen face image dataset is built from public surveillance cameras in Beijng subway. The modified Inception-ResNet-v2 deep learning models are trained end-to-end by triplet supervisory signal to embed faces into 128-dimension feature spaces. The trained models are deployed to the Jetson TX2 embedded system for real time offline identification. The classification accuracy of 94.36% ± 0.01 and 98.92% ± 0.005 are achieved on the balanced and imbalanced test set, respectively. With the threshold learned from the balanced test set, a VAL of 98.33% ± 0.05 with 1.08% ± 0.005 FAR is achieved on the imbalanced test set. The experiments demonstrate that the automatic learned features are highly discriminative and generalize well to new identities. The proposed model is capable of identifying suspects from crowd in real time for public video surveillance, which bridges the gap between academic research and practical application.

Acknowledgment. The authors thank Shiwei Zhao and Li Li at First Research Institute of the Public Security Ministry for images collection. This work was supported by the Fundamental Research Funds for the Central Universities: 2017JC02. Yu Sun and Guan Wang contributed equally to this work.

References

1. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
2. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition, pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
3. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition, pp. 770–778 (2015)
4. Huang, G.B., Ramesh, M., Berg, T., et al.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst (2007)
5. Taigman, Y., Yang, M., Ranzato, M., et al.: DeepFace: closing the gap to human-level performance in face verification. In: Computer Vision and Pattern Recognition, pp. 1701–1708 (2014). https://doi.org/10.1109/CVPR.2014.220
6. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Computer Vision and Pattern Recognition, pp. 1891–1898 (2014). https://doi.org/10.1109/CVPR.2014.244
7. Sun, Y., Chen, Y., Wang, X., et al.: Deep learning face representation by joint identification-verification. In: Neural Information Processing Systems, pp. 1988–1996 (2014)
8. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: Computer Vision and Pattern Recognition, pp. 2892–2900 (2015). https://doi.org/10.1109/CVPR.2015.7298907
9. Sun, Y., Liang, D., Wang, X., et al.: DeepID3: face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
10. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Computer Vision and Pattern Recognition, pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
11. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Computer Vision and Pattern Recognition, pp. 529–534 (2011). https://doi.org/10.1109/CVPR.2011.5995566
12. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: OpenFace: a general-purpose face recognition library with mobile applications. Carnegie Mellon University (2016)
13. Huang, G.B., Learned-miller, E.: Labeled faces in the wild : updates and new reporting procedures. University of Massachusetts, Amherst (2014)
14. Zhang, K., Zhang, Z., Li, Z., et al.: Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE J. Solid-State Circuits **23**, 1161–1173 (2016). doi:10.1109/LSP.2016.2603342
15. Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv preprint arXiv:1602.07261 (2016)

16. Russakovsky, O., Deng, J., Su, H., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**, 211–252 (2015). doi:10.1007/s11263-015-0816-y
17. Abadi, M., Agarwal, A., Barham, P., et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (2016)
18. Duchi, J.C., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015). https://doi.org/10.5244/C.29.41