DeepID3: Face Recognition with Very Deep Neural Networks

Yi Sun[1] Ding Liang[2] Xiaogang Wang[3,4] Xiaoou Tang[1,4]
[1]Department of Information Engineering, The Chinese University of Hong Kong
[2]SenseTime Group
[3]Department of Electronic Engineering, The Chinese University of Hong Kong
[4]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
sy011@ie.cuhk.edu.hk liangding@sensetime.com
xgwang@ee.cuhk.edu.hk xtang@ie.cuhk.edu.hk

Abstract

The state-of-the-art of face recognition has been significantly advanced by the emergence of deep learning. Very deep neural networks recently achieved great success on general object recognition because of their superb learning capacity. This motivates us to investigate their effectiveness on face recognition. This paper proposes two very deep neural network architectures, referred to as DeepID3, for face recognition. These two architectures are rebuilt from stacked convolution and inception layers proposed in VGG net [10] and GoogLeNet [16] to make them suitable to face recognition. Joint face identification-verification supervisory signals are added to both intermediate and final feature extraction layers during training. An ensemble of the proposed two architectures achieves 99.53% *LFW face verification accuracy and* 96.0% *LFW rank-1 face identification accuracy, respectively. A further discussion of LFW face verification result is given in the end.*

1. Introduction

Using deep neural networks to learn effective feature representations has become popular in face recognition [12, 20, 17, 22, 14, 13, 18, 21, 19, 15]. With better deep network architectures and supervisory methods, face recognition accuracy has been boosted rapidly in recent years. In particular, a few noticeable face representation learning techniques are evolved recently. An early effort of learning deep face representation in a supervised way was to employ face verification as the supervisory signal [12], which required classifying a pair of training images as being the same person or not. It greatly reduced the intra-personal variations in the face representation. Then learning discriminative deep face representation through large-scale face identity classification (face identification) was proposed by DeepID [14] and DeepFace [17, 18]. By classifying training images into a large amount of identities, the last hidden layer of deep neural networks would form rich identity-related features. With this technique, deep learning got close to human performance for the first time on tightly cropped face images of the extensively evaluated LFW face verification dataset [6]. However, the learned face representation could also contain significant intra-personal variations. Motivated by both [12] and [14], an approach of learning deep face representation by joint face identification-verification was proposed in DeepID2 [13] and was further improved in DeepID2+ [15]. Adding verification supervisory signals significantly reduced intra-personal variations, leading to another significant improvement on face recognition performance. Human face verification accuracy on the entire face images of LFW was surpassed finally [13, 15]. Both GoogLeNet [16] and VGG [10] ranked in the top in general image classification in ILSVRC 2014. This motivates us to investigate whether the superb learning capacity brought by very deep net structures can also benefit face recognition.

Although supervised by advanced supervisory signals, the network architectures of DeepID2 and DeepID2+ are much shallower compared to recently proposed high-performance deep neural networks in general object recognition such as VGG and GoogLeNet. VGG net stacked multiple convolutional layers together to form complex features. GoogLeNet is more advanced by incorporating multi-scale convolutions and pooling into a single feature extraction layer coined inception [16]. To learn efficiently, it also introduced 1x1 convolutions for feature dimension reduction.

In this paper, we propose two deep neural network architectures, referred to as DeepID3, which are significantly deeper than the previous state-of-the-art DeepID2+ architecture for face recognition. DeepID3 networks are rebuilt from basic elements (*i.e.*, stacked convolution or inception

layers) of VGG net [10] and GoogLeNet [16]. During training, joint face identification-verification supervisory signals [13] are added to the final feature extraction layer as well as a few intermediate layers of each network. In addition, to learn a richer pool of facial features, weights in higher layers of some of DeepID3 networks are unshared. Being trained on the same dataset as DeepID2+, DeepID3 improves the face verification accuracy from 99.47% to 99.53% and rank-1 face identification accuracy from 95.0% to 96.0% on LFW, compared with DeepID2+. The "true" face verification accuracy when wrongly labeled face pairs are corrected and a few hard test samples will be further discussed in the end.

2. DeepID3 net

For the comparison purpose, we briefly review the previously proposed DeepID2+ net architecture [15]. As illustrated in Fig. 1, DeepID2+ net has three convolutional layers followed by max-pooling (neurons in the third convolutional layer share weights in only local regions), followed by one locally-connected layer and one fully-connected layer. Joint identification-verification supervisory signals [13] are added to the last fully-connected layer (from which the final features are extracted for face recognition) as well as a few fully connected layers branched out from intermediate pooling layers to better supervise early feature extraction processes.

The proposed DeepID3 net inherits a few characteristics of the DeepID2+ net, including unshared neural weights in the last few feature extraction layers and the way of adding supervisory signals to early layers. However, the DeepID3 net is significantly deeper, with ten to fifteen non-linear feature extraction layers, compared to five in DeepID2+. In particular, we propose two DeepID3 net architectures, referred to as DeepID3 net1 and DeepID3 net2, as illustrated in Fig. 2 and Fig. 3, respectively. The depth of DeepID3 net is due to stacking multiple convolution/inception layers before each pooling layer. Continuous convolution/inception helps to form features with larger receptive fields and more complex nonlinearity while restricting the number of parameters [10].

The proposed DeepID3 net1 takes two continuous convolutional layers before each pooling layer. Compared to the VGG net proposed in previous literature [10, 19], we add additional supervisory signals in a number of full-connection layers branched out from intermediate layers, which helps to learn better mid-level features and makes optimization of a very deep neural network easier. The top two convolutional layers are replaced by locally connected layers. With unshared parameters, top layers could form more expressive features with a reduced feature dimension. The last locally connected layer of our DeepID3 net1 is used to extract the final features without an additional fully

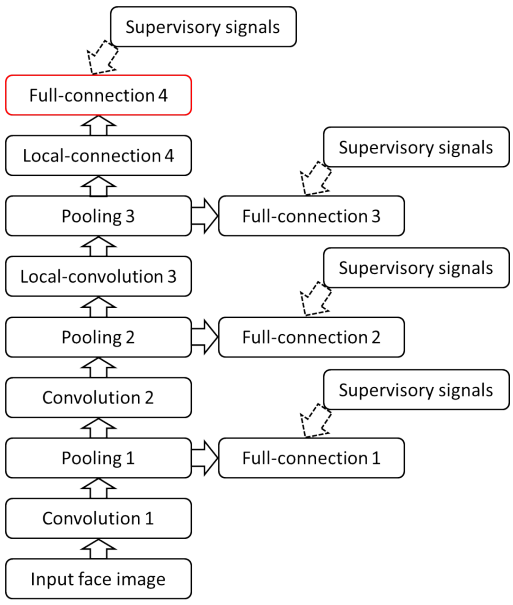


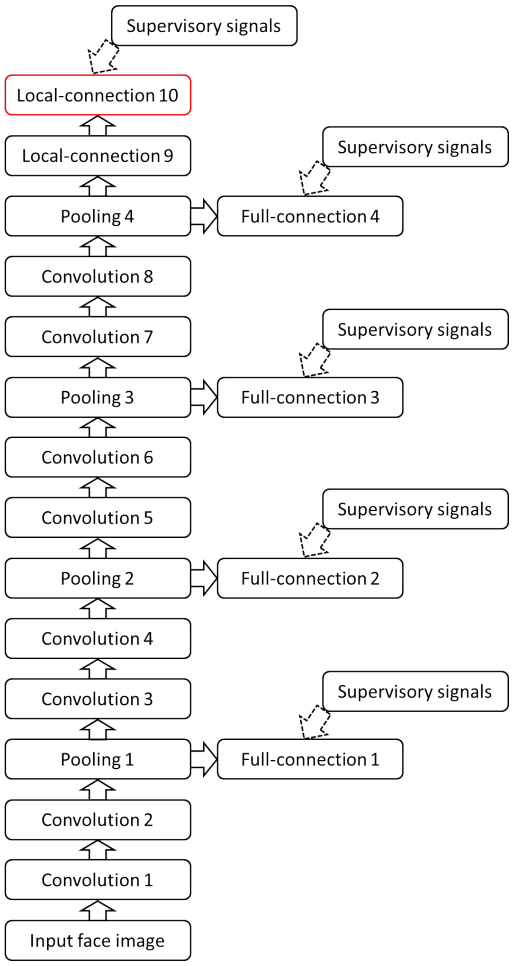
Figure 1: Architecture of DeepID2+ net [15]. Solid arrows show forward-propagation directions. Dashed arrows point the layers on which joint face identification-verification supervisory signals are added. The final feature extraction layer in red box is used for face recognition.

connected layer.

DeepID3 net2 starts with every two continuous convolutional layers followed by one pooling layer as does in DeepID3 net1, while taking inception layers [16] in later feature extraction stages: there are three continuous inception layers before the third pooling layer and two inception layers before the fourth pooling layer. Joint identification-verification supervisory signals are added on fully connected layers following each pooling layer.

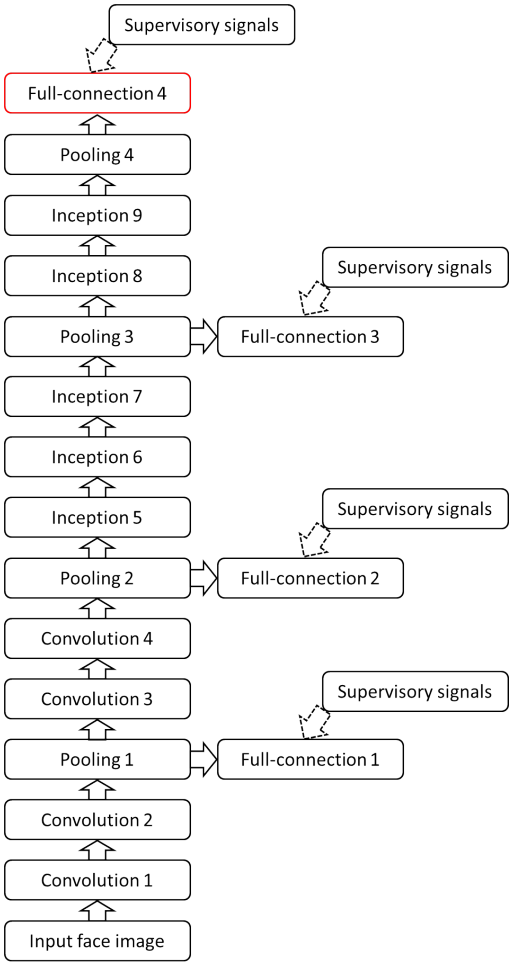
In the proposed two network architectures, rectified linear non-linearity [9] is used for all except pooling layers, and dropout learning [5] is added on the final feature extraction layer. Although with significant depth, our DeepID3 networks are much smaller than VGG net or GoogLeNet proposed in general object recognition due to a restricted number of feature maps in each layer.

The proposed DeepID3 nets are trained on the same 25 face regions as DeepID2+ nets [15], with each network taking a particular face region as input. These face regions are selected by feature selection in the previous work [13], which differ in positions, scales, and color channels such that different networks could learn complementary information. After training, these networks are used to extract features from respective face regions. Then an additional Joint Bayesian model [3] is learned on these



DeepID3 net1

Figure 2: Architecture of DeepID3 net1. Figure description is the same as Fig. 1.



DeepID3 net2

Figure 3: Architecture of DeepID3 net2. Figure description is the same as Fig. 1.

features for face verification or identification. All the DeepID3 networks and Joint Bayesian models are learned on the same approximately 300 thousand training samples as used in DeepID2+ [15], which is a combination of CelebFaces+ [14] and WDRef [3] datasets, and tested on LFW [6]. People in these two training data sets and the LFW test set are mutually exclusive. The face verification performance on LFW of individual DeepID3 net is compared to DeepID2+ net in Fig. 4 on the 25 face regions (with horizontal flipping), respectively. On average, DeepID3 net1 and DeepID3 net2 reduce the error rate by 0.81% and 0.26% compared to DeepID2+ net, respectively.

3. Experiments

To reduce redundancy, DeepID3 net1 and net2 are used to extract features on either the original or the horizontally

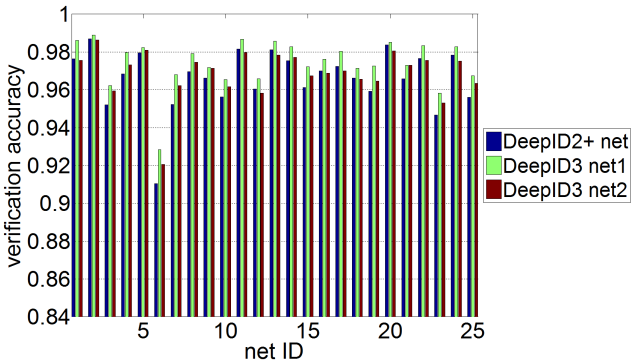


Figure 4: LFW face verification accuracy of individual DeepID2+ and DeepID3 net trained on the same face regions in [15].

Table 1: Face verification on LFW.

method	accuracy (%)
High-dim LBP [4]	95.17 ± 1.13
TL Joint Bayesian [2]	96.33 ± 1.08
DeepFace [17]	97.35 ± 0.25
DeepID [14]	97.45 ± 0.26
GaussianFace [7, 8]	98.52 ± 0.66
DeepID2 [13, 11]	99.15 ± 0.13
DeepID2+ [15]	99.47 ± 0.12
DeepID3	$\mathbf{99.53 \pm 0.10}$

flipped face region but not both. In test, feature extraction takes 50 times of forward propagation with half from DeepID3 net1 and the other half from net2. These features are concatenated into a long feature vector of approximately $30,000$ dimensions. With PCA, it is reduced to 300 dimensions on which a Joint Bayesian model is learned for face recognition.

We evaluate DeepID3 networks under the LFW face verification [6] and LFW face identification [1, 18] protocols, respectively. For face verification, 6000 given face pairs are verified to tell if they are from the same person. We achieve a mean accuracy of $\mathbf{99.53}\%$ under this protocol. Comparisons with previous works on mean accuracy and ROC curves are shown in Tab. 1 and Fig. 5, respectively.

For face identification, we take one closed-set and one open-set identification protocols. For closed-set identification, the gallery set contains 4249 subjects with a single face image per subject, and the probe set contains 3143 face images from the same set of subjects in the gallery. For open-set identification, the gallery set contains 596 subjects with a single face image per subject, and the probe set contains 596 genuine probes and 9494 imposter ones. Table 2 compares Rank-1 identification accuracy of closed-set identification and Rank-1 Detection and Identification rate (DIR) at a 1% False Alarm Rate (FAR) of open-set identification, respectively. We achieve $\mathbf{96.0}\%$ closed-set and $\mathbf{81.4}\%$ open-set face identification accuracies, respectively.

4. Discussion

There are three test face pairs which are labeled as the same person but are actually different people as announced on the LFW website. Among these three pairs, two are classified as the same person while the other one is classified as different people by our DeepID3 algorithm. Therefore, when the label of these three face pairs are corrected, the actual face verification accuracy of DeepID3 is 99.52%. For DeepID2+ [15], its face

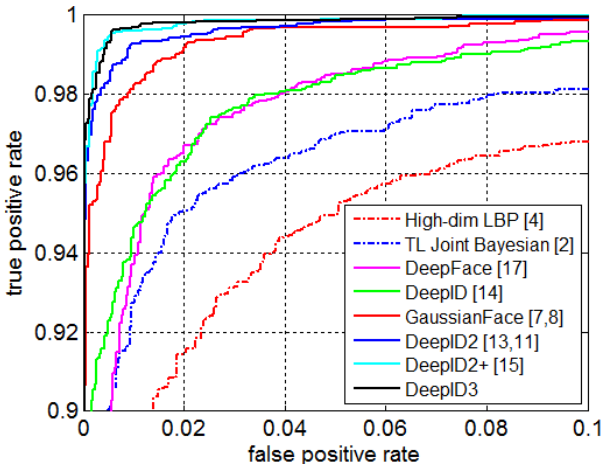


Figure 5: ROC of face verification on LFW.

Table 2: Closed- and open-set identification tasks on LFW.

method	Rank-1 (%)	DIR @ 1% FAR (%)
COTS-s1 [1]	56.7	25
COTS-s1+s4 [1]	66.5	35
DeepFace [17]	64.9	44.5
WST Fusion [18]	82.5	61.9
DeepID2+ [15]	95.0	80.7
DeepID3	**96.0**	**81.4**

verification accuracy before correcting the three wrong labels is 99.47%. However, DeepID2+ classified all the three wrongly labeled positive face pairs as different people. When these three wrong labels are corrected, the true face verification accuracy of DeepID2+ is also 99.52% [15]. DeepID3, although taking similar very deep architectures as VGG and GoogLeNet, does not improve over DeepID2+, with significantly shallower architecture, on the LFW face verification task. Whether those very deep architectures would take advantage of more training face data and finally surpass shallower architectures like DeepID2+ remains an open question.

We examine the test face pairs in LFW which are wrongly classified by all the DeepID series algorithms including DeepID [14], DeepID2 [13, 11],DeepID2+ [15], and DeepID3. There are nine common false positives and three common false negatives in total, around half of all wrongly classified face pairs by DeepID3. The three face pairs labeled as the same person but being classified as different people are shown in Fig. 6. The first pair of faces show great contrast of ages. The second pair is actually different people due to errors in labeling. The third one

is an actress with significantly different makeups. Fig. 7 shows the nine face pairs labeled as different people while being classified as the same person by algorithms. Most of them look similar or have interference such as occlusions.

Figure 6: Common false negatives in DeepID series algorithms.

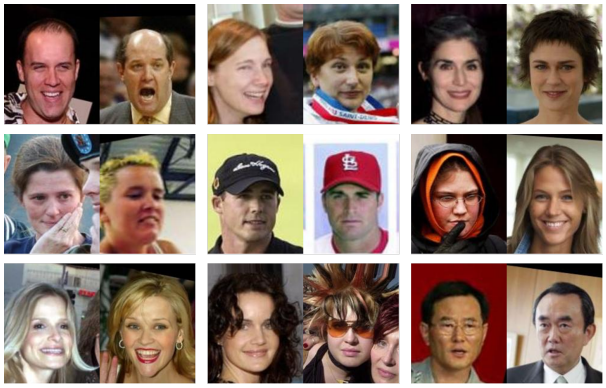


Figure 7: Common false positives in DeepID series algorithms.

5. Conclusion

This paper proposes two significantly deeper neural network architectures, coined DeepID3, for face recognition. The proposed DeepID3 networks achieve the state-of-the-art performance on both LFW face verification and identification tasks. However, when a few wrong labels in LFW are corrected, the improvement of DeepID3 over DeepID2+ on LFW face verification vanished. The effectiveness of very deep neural networks would be further investigated on larger scale training data in the future.

References

[1] L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *TR MSU-CSE-14-1*, 2014.

[2] X. Cao, D. Wipf, F. Wen, and G. Duan. A practical transfer learning algorithm for face verification. In *Proc. ICCV*, 2013.

[3] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Proc. ECCV*, 2012.

[4] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proc. CVPR*, 2013.

[5] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[7] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. Technical report, arXiv:1404.3840, 2014.

[8] C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. In *Proc. AAAI*, 2015.

[9] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. ICML*, 2010.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. Technical report, arXiv:1409.1556, 2014.

[11] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. NIPS*, 2014.

[12] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Proc. ICCV*, 2013.

[13] Y. Sun, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. Technical report, arXiv:1406.4773, 2014.

[14] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Proc. CVPR*, 2014.

[15] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. Technical report, arXiv:1412.1265, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. Technical report, arXiv:1409.4842, 2014.

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proc. CVPR*, 2014.

[18] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. Technical report, arXiv:1406.5266, 2014.

[19] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. Technical report, arXiv:1411.7923, 2014.

[20] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *Proc. ICCV*, 2013.

[21] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning and disentangling face representation by multi-view perceptron. In *Proc. NIPS*, 2014.

[22] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. Technical report, arXiv:1404.3543, 2014.