Survey of Deep Learning In Face Recognition

FU Zhi-Peng
School of Computer Science, Northwestern Polytechnical University,
Xi'an
beiting_1983@163.com

ZhANG Yan-Ning*
School of Computer Science, Northwestern Polytechnical University,
Xi'an
ynzhang@nwpu.edu.cn

HOU Hai-Yan
College of Medical Technology and Engineering, Henan University of Science and Technology,
Luoyang
hyhou_2006@163.com

***Abstract*—Deep learning is the research focus in the recent years. Because of its excellent performance,it is widely used in the area of pattern recognition. Facial feature is useful for a variety of tasks, the application of deep learning in this area is also developing fast. We introduce some recent research work in this domain, and show the potential of it.**

***Key words*—Deep learning, face recognition, DBN, RBM, auto-encoder**

I. INTRODUCTION

Allowing computer to model our expressions well enough to exhibit the intelligence of machine has been the focus for decades. There emerged a lot of Algorithms to realize this task, such as eigenfaces, which use principal component analysis to efficiently represent pictures of faces[1][2]; Graph Matching, which use dynamic link structure for distortion invariant object recognition, and employ elastic graph matching to fine the closest stored graph[3]; Geometrical Feature Matching, which are based on the computation of a set of geometrical features from the picture of a face[4]; Template Matching, A simple version of template matching is that a test image represented as a two-dimensional array of intensity values is compared using a suitable metric, such as the euclidean distance, with a single template representing the whole face; Support Vector Machine, in witch face recognition problem is formulated as a problem in difference space, which models dissimilarities between two facial images[5]; Neural Networks, the first artificial neural networks (ANN)techniques used for face recognition is a single layer adaptive network called WISARD which contains a separate network for each stored individual[6] . But the none of their robustness to pose, background, and occlusions can not reach the state-of-the-art.

All kinds of databases are also designed to test the algorithms such as FERET, CMU PIE, AR, XM2VTS, ORL and so on[7][8][9].

Thanks to the development of deep learning, which make it possible for the face recognition algorithm to realize the robustness and accuracy at the same time, through the help of deep learning state-of-the-art results can be realized.

In this paper, we will introduce the basic principal of deep learning in section 2 and shows several typical algorithms used in face recognition in section III

II. RELATED WORK

Inspired by the depth of the brain[10], neural network researchers had devoted themselves to train deep multi-layer neural networks. In fact, the deep architecture is under research for a long time, but there was no successful methods were reported until 2006 because of the difficulty in training deep networks. Milestone was built in 2006, when Hinton and collaborators at U. of Toronto introduced Deep Belief Networks for short, using learning algorithm that greedily trains one layer at a time and exploiting an unsupervised learning algorithm for each layer[11].

A. Typical deep networks

The typical deep networks include Deep Generative Architectures, Convolutional Neural Networks, Auto-Encoders, as introduced below:

- **Deep generative architecture:** A sigmoid belief network is a typical deep generative architecture[12][13][14][15], in the sigmoid belief network, the layer above give values to the layer below, the conditional distributions parametrization as fallow:

$$P(\mathrm{h}_i^k = 1 \left| \mathrm{h}^{k+1}\right.) = \mathrm{sigm}(\mathrm{b}_i^k + \sum_j \mathrm{W}_{i,j}^{k+1}\mathrm{h}_j^{k+1}) \cdot$$

h_i^k is the binary activation of hidden node i in layer *k*, h^k is the vector (h_1^k, h_2^k, ...) and they denote the input vector $x=h^0$, P(...) always represents a probability distribution:

$$P(\mathrm{x}, \mathrm{h}^1, \ldots, \mathrm{h}^\ell) = P(\mathrm{h}^\ell)\left(\prod_{k=1}^{\ell-1} P(\mathrm{h}^k \left| \mathrm{h}^{k+1}\right.)\right) P(\mathrm{x} \left| \mathrm{h}^1\right.)$$

$$P(\mathrm{h}^{\ell-1}, \mathrm{h}^\ell) \infty e^{\mathrm{b}'\mathrm{h}^{\ell-1} + \mathrm{c}'\mathrm{h}^\ell + \mathrm{h}^{\ell\prime}\mathrm{W}\mathrm{h}^{\ell-1}}$$

- **Convolutional neural networks:** It is a exception of the hardness of the training of deep supervised networks.

Thanks to Education Reform Project (2012Y-098) and Youth Science Found (2013QN045) of Henan University of Science and Technology for founding

978-1-4799-6284-6/14/$31.00© 2014 IEEE

And now, pattern recognition systems based on convolutional neural networks are among the best performing systems[16].

- **Auto-encoder:** It have been used as building blocks to train deep networks, where each level is associated with an auto-encoder that can be trained separately[17]. An auto-encoder is trained to encode the input x into some representation c(x) so that the input can be reconstructed from that representation. The formulation that generalizes the mean squared error criterion to the minimization of the negative log-likelihood of the reconstruction, given the encoding c(x):

$$RE = -\log P(\mathrm{x}|\mathrm{c}(\mathrm{x})).$$

If $x|c(x)$ is Gaussian, it recovers the familiar squared error. If the inputs xi are either binary or considered to be binomial probabilities, then the loss function would be

$$-\log P(\mathrm{x}|\mathrm{c}(\mathrm{x})) = -\sum_i \mathrm{x}_i \log \mathrm{f}_i(\mathrm{c}(\mathrm{x})) + (1-\mathrm{x}_i)\log(1-\mathrm{f}_i(\mathrm{c}(\mathrm{x})))$$

f (•) is called the decoder, and f (c(x)) is the reconstruction produced by the network, and in this case should be a vector of numbers in (0, 1), e.g., obtained with a sigmoid.

There are other structures such as denoising auto-encoder[18], DCN[19], sun-product[20] and so on, witch get state-of-the-art results. Several applications in face recognition using the methods described above are mentioned in section III.

B. Restricted Boltzmann Machines

Deep Belief Networks (DBNs) are the core units of a deep architecture, and it is based on Restricted Boltzmann Machines(RBMs), RBMs is an energy based model:

$$Energy(\mathrm{x},\mathrm{h}) = -\mathrm{b}'\mathrm{x} - \mathrm{c}'\mathrm{x} - \mathrm{h}'\mathrm{W}\mathrm{x} - \mathrm{x}'\mathrm{U}\mathrm{x} - \mathrm{h}'\mathrm{V}\mathrm{h}$$

There are two types of the parameters, which denote by θ, the offsets b_i and c_i (each associated with a single element of the vector $\boldsymbol{x}$ or $\boldsymbol{h}$), and the weights W_{ij}, U_{ij} and V_{ij} (each associated with a pair of units).

The model will be more stable when the energy is the smallest, which means the maximum likelihood estimation of the parameters:

$$\begin{aligned}\frac{\partial \log P(\mathrm{x})}{\partial\theta} &= \frac{\partial \log \sum_{\mathrm{h}} e^{-\mathrm{Energy}(\mathrm{x},\mathrm{h})}}{\partial\theta} - \frac{\partial \log \sum_{\tilde{\mathrm{x}},\mathrm{h}} e^{-\mathrm{Energy}(\tilde{\mathrm{x}},\mathrm{h})}}{\partial\theta} \\ &= -\frac{1}{\sum_{\mathrm{h}} e^{-\mathrm{Energy}(\mathrm{x},\mathrm{h})}} \sum_{\mathrm{h}} e^{-\mathrm{Energy}(\mathrm{x},\mathrm{h})} \frac{\partial \mathrm{Energy}(\mathrm{x},\mathrm{h})}{\partial\theta} \\ &\quad + \frac{1}{\sum_{\tilde{\mathrm{x}},\mathrm{h}} e^{-\mathrm{Energy}(\tilde{\mathrm{x}},\mathrm{h})}} \sum_{\tilde{\mathrm{x}},\mathrm{h}} e^{-\mathrm{Energy}(\tilde{\mathrm{x}},\mathrm{h})} \frac{\partial \mathrm{Energy}(\tilde{\mathrm{x}},\mathrm{h})}{\partial\theta} \\ &= -\sum_{\mathrm{h}} P(\mathrm{h}|\mathrm{x}) \frac{\partial \mathrm{Energy}(\mathrm{x},\mathrm{h})}{\partial\theta} + \sum_{\tilde{\mathrm{x}},\mathrm{h}} P(\tilde{\mathrm{x}},\mathrm{h}) \frac{\partial \mathrm{Energy}(\tilde{\mathrm{x}},\mathrm{h})}{\partial\theta}\end{aligned}$$

III. DEEP ARCHITECTURE IN FACE RECOGNITION

All the above mentioned architectures can be used in the area of face recognition, some works had already gotten better results comparing to other algorithms.

A. Deep generative architecture in face recognition

G. Hinton et al. [21]trained DBNs using a greedy layer-wise procedure that guarantees to improve a lower bound on the log-likelihood of the data. The author argued that the generative models can learn low-level features without requiring feedback from the label, they also can learn more parameters than discriminative models without overfitting. And besides, this method can be extended to be used in face recognition.

Marc'Aurelio Ranzato et al[22] use a gated MRF(Markov Random Field) as the front-end of a DBNs to learn a deep generative model of images including human face, it can learn better features for facial expressions. It is also robust to occluded images. The model is computational expensive, but thanks to the data and computational power enlargement, the cost could be ignored.

In the first layer, the mPoT model is adopted as the front-end for the DBNs, the mPoT model is a higher-order MRF with potentials defined over triplets of variables: two input pixels and one latent variable, denoted by $h^c_{j.}$

In the higher layer, assume that the input to that layer consists of a binary vector denoted by h^{i-1}. This is modeled by a binary RBMs, the RBMs is also defined by energy function in which W^i is the i-th layer parameter matrix and bi is the i-th layer vector of biases:

$$E(\mathrm{h}^{i-1},\mathrm{h}^{i}) = -\sum_k h^i_k (W^{iT}_k \mathrm{h}^{i-1} + b^i_k)$$

$$p(h^i_k = 1|\mathrm{h}^{i-1}) = \sigma(W^{iT}_k \mathrm{h}^{i-1} + b^i_k)$$

$$p(h^{i-1}_k = 1|\mathrm{h}^{i}) = \sigma(W^{i}_k \mathrm{h}^{i})$$

The training procedure in this paper is shown below:

- First, train mPoT to fit the distribution of the input.
- Second, use mPoT to compute the expectation of the first layer latent variables conditioned on the input training images.
- Third, use these expected values as input to train the second layer of latent variables. Once the second layer is trained, use it to compute expectations of the second layer latent variables conditioned on the second layer input to provide inputs to the third layer.
- Repeat until to the last layer.

This method gained a better result as figure1[22] shows:

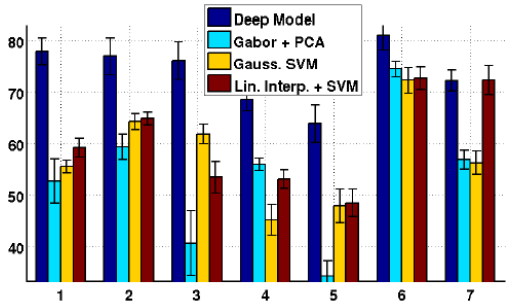

Figure 1 facial recognition accuracy on the TFD dataset when only the test images are subject to 7 types of occlusion.

B. Convolutional neural networks in face recognition

Osadchy M et al.[23]introduce a successful use of convolutional networks in face detection and pose estimation, they build a trainable system which can map raw images to points in a low-dimensional space and train the system to map face images with known poses to the corresponding points on

the manifold. They also train it to map non-face images to points far away from the manifold. Proximity to the manifold then tells whether or not an image is a face, and projection to the manifold yields an estimate of the pose. A Convolutional Network is employed as the basic architecture for the image-to-face-space mapping function.
Taable 1[23] shows the better result of their work

TABLE1: COMPARISONS OF THE RESULTS WITH OTHER MULTI-VIEW DETECTORS

sample	tilted		profile		MIT+CMU	
Osadchy M ‘s detector	90%	97%	67%	83%	83%	88%
Jones&Viola（tilted）	90%	95%	x		x	
Jones&Viola（profile）	x		70%	83%	x	
Rowley	89%	96%	x		x	
Sclmeiderman & Kanade	x		86%	93%	x	

Sun Y et al.[24]use three levels of convolutional networks to construct a cascaded regression structure. In the first level, they make accurate predictions that effectively avoids the local minimum. Use the whole face as input that can make the best use of texture context information. This will be effective even if the low-level features are ambiguous. The other two levels are used to refine the result of the first level, their inputs are constrained to a small local region around the initial position. Multiple convolutional networks are used to improver the reliability and accuracy. There are three characters of their work:

- First, the first level convolutional networks should be deep enough;
- Second, absolute value rectification after the hyperbolic tangent activation function is used to improve the robustness;
- Last, share the weights locally to enhance the performance.

The result shows that their method improved the results a lot compare with the method proposed by Belhumeur et al. [25] and Cao et al.[26] on LFPW test images, More than 20% relative accuracy improvement is achieved for nose tip and two mouth corners.

Huang G.B et al.[27] uses CDBN to learn hierarchical representations, develops local convolutional restricted Boltzmann machines to exploits the global structure. The networks are not only trained on the gray images, but also on the LBP images to get 59 dimension uniform LBP features, face in the images is divide into overlapped patches, which are related to the hidden units. Cosine Similarity Metric Learning (CSML) is applied in this method. The test is carried on LFW database, and get a state-of-the-art accuracy.

Nair and Hinton[28]applied deep learning to object recognition and face verification, using a modification to binomial units that they refer to as noisy rectified linear units. they subsample the face images to 32x32. In addition, their method was not translation invariant and had to rely on manual alignment through hand-corrected eye coordinates as preprocessing.

C. C.Other related works

Luo P, Wang X, Tang X[29] propose a hierarchial face parser, which is composed of part-based face detectors, component-based detectors, and component segmentators. For accurate face parsing, they recast segmentation of face components as the cross-modality data transformation problem, and solve it by a new deep learning strategy, which can output the label map given an image patch as input. By incorporating the deformable part-based detectors and the segmentators, the parser is very robust to occlusions, pose variations, and background clutters. They test their method on several applications and demonstrate great improvement.

Cai X et al.[30]use stacked ISA to train the deep networks, and then formulate the proposed method as an appropriate optimization problem, and employ discriminative pre-training and fine-tuning methods, which are widely used in deep learning. This method use ISA as the basic unit instead of RBM, which is similar to auto-encoder, the pooling units are used to confirm the invariance. The result of test on LFW is up to 88.75%.

IV. CONCLUSION

The development of parallel computing, internet and the booming of computer hardware technology make it possible to train deep networks on large amount of data. The deep architecture shows its superiority over other methods not only in the accuracy but also the robustness to all kinds of conditions. But the utilization of deep learning in face recognition is still at the beginning, so a lot of works are needed to be done, such as how to get a deep architecture that can accurate understand the human expressions, how to verify a persons’ face on the low resolution images, how to transform expressions to emotions and so on.

ACKNOWLEDGMENT

This paper is founded by Education Reform Project (2012Y-098) and Youth Science Found (2013QN045) of Henan University of Science and Technology.

REFERENCES

[1] L. Sirovich and M. Kirby, “Low-Dimensional procedure for the characterisation of human faces,” J. Optical Soc. of Am., vol. 4, pp.519-524, 1987.

[2] M. Kirby and L. Sirovich, “Application of the Karhunen-Loève procedure for the characterisation of human faces,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 12, pp. 831-835, Dec.1990.

[3] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C. Von Der Malsburg, R.P. Wurtz, and M. Konen, “Distortion Invariant object recognition in the dynamic link architecture,” IEEE Trans. Computers, vol. 42, pp. 300-311, 1993.

[4] S. Tamura, H. Kawa, and H. Mitsumoto, “Male/Female identification from 8_6 very low resolution face images by neural network,” Pattern Recognition, vol. 29, pp. 331-335, 1996.

[5] P.J. Phillips, “Support vector machines applied to face recognition,” Processing system 11, 1999.

[6] T.J. Stonham, “Practical face recognition and verification with

WISARD," Aspects of Face Processing, pp. 426-441, 1984.
[7] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(12):1615–1618, 2003.
[8] A. R. Martinez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center(CVC) Technical Report, Barcelona, Spain, June 1998.
[9] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In Second International Conference on Audio and Video-based Biometric Person Authentication, March 1999.
[10] Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research, Computational Neuroscience: Theoretical Insights into Brain Function*, *165*, 33–56.
[11] Hinton, G. E., Osindero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527–1554.
[12] Dayan, P., Hinton, G. E., Neal, R., & Zemel, R. (1995). The Helmholtz machine. *Neural Computation*, *7*,889–904.
[13] Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, *268*, 1558–1161.
[14] Saul, L. K., Jaakkola, T., & Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, *4*, 61–76.
[15] Titov, I., & Henderson, J. (2007). Constituent parsing with incremental sigmoid belief networks. In *Proc.45th Meeting of Association for Computational Linguistics (ACL'07)*, pp. 632–639 Prague, Czech Republic.
[16] T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research, Computational Neuroscience: Theoretical Insights into Brain Function*, *165*, 33–56.
[17] Vincent, P., Larochelle, H., Bengio, Y., &Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In Cohen, W. W., McCallum, A., & Roweis, S. T. (Eds.), *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, pp. 1096–1103. ACM.
[18] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. The Journal of Machine Learning Research, 2010, 9999: 3371-3408.
[19] Deng L, Yu D. Deep convex net: A scalable architecture for speech pattern classification[C]//Proceedings of the Interspeech. 2011.
[20] Poon H, Domingos P. Sum-product networks: A new deep architecture[C]//Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. IEEE, 2011: 689-690.
[21] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.
[22] Ranzato M, Susskind J, Mnih V, et al. On deep generative models with applications to recognition[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 2857-2864.
[23] Osadchy M, Cun Y L, Miller M L. Synergistic face detection and pose estimation with energy-based models[J]. The Journal of Machine Learning Research, 2007, 8: 1197-1215.
[24] Sun Y, Wang X, Tang X. Deep Convolutional Network Cascade for Facial Point Detection[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 3476-3483.
[25] Belhumeur P N, Jacobs D W, Kriegman D J, et al. Localizing parts of faces using a consensus of exemplars[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 545-552.
[26] Cao X, Wei Y, Wen F, et al. Face alignment by explicit shape regression[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2887-2894.
[27] Huang G B, Lee H, Learned-Miller E. Learning hierarchical representations for face verification with convolutional deep belief networks[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2518-2525.
[28] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010: 807-814.
[29] Luo P, Wang X, Tang X. Hierarchical face parsing via deep learning[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2480-2487.
[30] Cai X, Wang C, Xiao B, et al. Deep nonlinear metric learning with independent subspace analysis for face verification[C]//Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012: 749-752.