# MODULAR HIERARCHICAL FEATURE LEARNING WITH DEEP NEURAL NETWORKS FOR FACE VERIFICATION

*Xue Chen*[1], *Baihua Xiao*[1], *Chunheng Wang*[1], *Xinyuan Cai*[1], *Zhijian Lv*[2], *Yanqin Shi*[2]

[1]State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]Beijing Institute of Science and Technology Information

## ABSTRACT

Feature representations play a crucial role in modern face recognition systems. Most hand-crafted image descriptors usually provide low-level information. In this paper, we propose a novel feature learning method based on deep neural networks to obtain high-level, hierarchical representations for face verification. Learning proceeds in two phases. In the pre-training phase, we train Restricted Boltzmann Machine(RBM) networks for each modular region in the image separately. In the fine-tuning phase, in order to develop good discriminative ability, we stack the RBM networks of each region in deep architecture and combine deep learning with side information constraints in the whole image scale. Finally, we formulate the proposed method as an appropriate optimization problem and adopt gradient descent algorithm to get the optimal solution. We evaluate our method on the LFW dataset. Representations learned from the networks achieve comparable performance (93.11%) to the state-of-art method.

***Index Terms***— deep neural networks, feature learning, face verification

## 1. INTRODUCTION

The two primary face recognition tasks are identification and verification. The goal of identification is to recognize the person by searching the database to find who he/she is. In this paper, we focus on verification, which is to validate the identity claimed by a person. When a pair of images are presented, the task is to indicate whether the pair contain the same person or not. Face verification poses a huge challenge due to variations caused by complex background, lighting, pose, and occlusions. The difference brought by these factors could be larger than that caused by identity changes, making the unconstrained face verification problem very difficult.

Most face recognition methods generally rely on hand-crafted descriptors such as SIFT [1], Local Binary Patterns (LBP) [2] and its variant, Hierarchical Local Binary Pattern (HLBP) [3]. These features usually fail to give good performance, because they only provide low-level and limited information. Recently, there has been much interest in developing unsupervised or supervised feature learning algorithms for image representations. Nair and Hinton [4] applied deep neural networks to object recognition and face verification, using a modification to binomial units. Honglak Lee et al. [5, 6] introduced convolutional deep networks to learn complementary hierarchical representations. Besides, S. M. Ali Eslami et al. [7] proposed an new network named as Shape Boltzmann Machine to model object shape and achieved good performance. The deep network attracts people's attention mainly because of its powerful learning ability. Unfortunately, training becomes hard when the input is of high dimension. To make learning computationally tractable, people usually downsample the image to a smaller size [4]. However, this operation may lose much useful information for subsequent discriminative training.

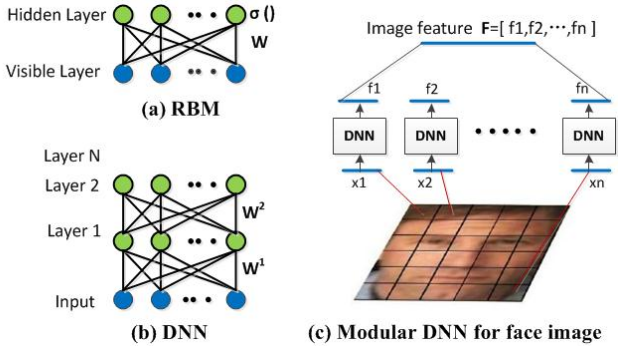


**Fig. 1**. (a)Restricted Boltzmann Machine (b)Deep Neural Network (c)Modular DNN. A modular region $x$ of the image is used as input for a DNN. All the local features $f$ from the DNNs are combined as a complementary representation.

In this paper, we present a different approach to solve the problem above. For the image of high-resolution, we divide it into non-overlapping local regions and pre-train a separate Deep Neural Network(DNN) for each region, as in Fig 1(c), which helps to reduce the input dimension of each network. All the local features from the modular DNNs are combined as a complementary representation for the image. Then we use side information constraints to form the optimization goal

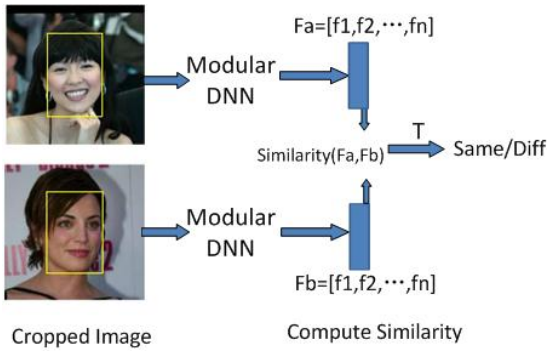


**Fig. 2**. Architecture of Face Verification System. $Fa/Fb$ represents the combined feature of a image from the Modular DNN. $T$ is the decision threshold.

of the deep networks and adopt gradient descent to update the connection weights. Finally, face verification is performed by comparing the similarity build on the learned networks with a threshold. Fig 2 shows the architecture of our model.

## 2. METHODS

### 2.1. Network Architecture

The RBM is a bipartite, undirected graphical model with a visible layer **v** and a hidden layer **h**, as in Fig 1(a). The model has three sets of parameters: (1) weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$ that defines a potential between visible variables and hidden variables; (2) the biases $c \in \mathbb{R}^{d_1}$ for visible units; (3) the biases $b \in \mathbb{R}^{d_2}$ for hidden units. $d_1$ and $d_2$ are the number of units in the visible layer and the hidden layer respectively. To deal with real-valued data, the model uses gaussian units for visible variables and binary units for hidden variables. The energy function can be defined as follows:

$$E = \sum_{i \in vis} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{j \in hid} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{i,j}$$
$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})) \quad (1)$$

where Z is a normalization constant and $\sigma_i$ is the standard deviation of the Gaussian noise for visible unit $i$. In our method, we first normalize each component of the data to have zero mean and unit variance $\bar{\sigma}$ and then use noise-free reconstructions for training. This operation will simplify the learning process largely. Under the energy function, the conditional distribution of the model can be computed as follows:

$$P(h_j|\mathbf{v}) = \sigma(\frac{1}{\bar{\sigma}} \sum_i W_{i,j} v_i + b_j)$$
$$P(v_i|\mathbf{h}) = \mathcal{N}(v_i\ ;\ \bar{\sigma} \sum_j W_{i,j} h_j + c_i, \bar{\sigma}^2) \quad (2)$$

where $\sigma(s) = \frac{1}{1+\exp(-s)}$ is the sigmoid function (applied in the whole article), and $\mathcal{N}(\cdot;\cdot,\cdot)$ is a Gaussian distribution. With Eq.(2), we can derive that $\sigma$ is also the activation function of the hidden layer, as in Fig 1(a).

With the conditional probabilities, we can pre-train the RBM using greedily contrastive divergence [8, 9]. This pre-training process helps to get a proper initial state for the network. After training a RBM, the hidden unit activations can be used as input to further train the next layer RBM similarly.

After pre-training all the RBMs, we could construct a deep neural network by stacking the RBMs hierarchically. In this paper, we build the DNN with up to two layers of RBMs, one stacked by another, as in Fig 1(b). One primary motivation for using deeper models with many layers is that they have the potential to be much more representationally effective than shallower models. Moreover, it is also capable of capturing hierarchical, high-level features.

In particular, the neural network becomes hard to train when the input is of high dimension. Instead of downsampling the image to a smaller scale, we divide it into several non-overlapping local regions, and connect each DNN to only one region, as in Fig 1(c). The modular DNN can obtain a more effective representation because features are learned only if they are useful for expressing the corresponding region.

### 2.2. Side Information Constraints

Unlike most classification algorithms where training examples are given class labels, the common supervised constraints for face verification are the labeled side information. The information is often available in the form of pairwise constraints, i.e. pairs of similar or dissimilar face images. A general theme is to learn a transformation such that the state of the similar pair should be close, while the dissimilar pair be far. In our experiment, we first transfer the distance of the pair to a probability $\rho$, and then model the theme in the form of log-likelihood maximization.

Let $S = \{x, y\}$ be the set of similar pairs, and $D = \{x, y\}$ be the set of dissimilar pairs. The distance between two images x, y is defined as : $d(x, y) = (x - y)^T (x - y)$. We model the probability $\rho_p$ with respect to $(x, y) \in S$, i.e. the pair label $l$ is 1, as:

$$\rho_p(x, y|f, b, l = 1) = \sigma(b - d(f(x), f(y))) \quad (3)$$

and the probability $\rho_n$ with respect to $(x, y) \in D$, i.e. the pair label $l$ is 0, as:

$$\rho_n(x, y|f, b, l = 0) = 1 - \sigma(b - d(f(x), f(y))) \quad (4)$$

Then the log-likelihood of all the pairwise training samples can be written as:

$$Lg(f, b) = \sum_{(x,y) \in S} \log(\rho_p(x, y|f, b)) + \sum_{(x,y) \in D} \log(\rho_n(x, y|f, b)) \quad (5)$$

where $b$ is a bias term, and the function $f$ indicates a mapping space $f : x \to f(x)$.

From the formation of the likelihood function, we could figure out that maximizing the likelihood means defining a mapping $f$ where the distances of similar pairs get small while that of dissimilar pairs get large. In this paper, we exploit deep network to formulate the mapping $f$ explicitly, and regard the log-likelihood function $Lg(f, b)$ as the optimization criterion of the deep learning algorithm following. So the learned network will satisfy the need of face verification.

### 2.3. Inference and Learning

After constructing a deep network, we could perform the approximate inference in a feed-forward (bottom-up) manner and the training procedure in a feedback (up-down) manner. In the pre-training phase, we train separate RBMs for each region in the image. To exploit the global structure, we introduce the side information constraints to combine all the local regions together in the fine-tuning phase. We find the global optimum solution by using patch-based gradient descend algorithm.

**Inference**: Generally, the modelling ability of deep networks gets stronger as the number of hidden layers increase within a certain range. To control the model complexity, we just construct a three-layer DNN as the basic model in the experiment. Performance could be further improved with a deeper model. We use $z_i^{(l)}$, $a_i^{(l)}$ to denote the weighted sum and the corresponding activation on unit $i$ in layer $l$ respectively. The active functions of the first and second layer are $f^{(1)}(x)$ and $f^{(2)}(x)$, where $f^{(l)}(x) = \sigma(x), (l = 1, 2)$. The model parameter is $W = \{W^{(1)}, W^{(2)}\}$, as in Fig 1(b). When the image is divided into $N$ regions, the model parameter of the DNN connected to the region $x^c$ in the image $x$ is $W^c = \{W^{c(1)}, W^{c(2)}\} \in W, (c = 1, ..., N)$. So the modular DNN for the image can be denoted as $w = \{W^1, ..., W^N\}$. A fast bottom-up inference algorithm conditioned on the input vector $x^c$ can be written as:

$$\begin{aligned} z_i^{(1)}(x^c) &= x^c * W_i^{c(1)}, \quad a_i^{(1)}(x^c) = f^{(1)}(z_i^{(1)}(x^c)) \\ z_k^{(2)}(x^c) &= a^{(1)}(x^c) * W_k^{c(2)}, \quad a_k^{(2)}(x^c) = f^{(2)}(z_k^{(2)}(x^c)) \end{aligned} \tag{6}$$

**Learning**: With the feed-forward inference, the network defines a transformation mapping $f(w) : x \to a^{(2)}(x) = [a^{(2)}(x^1), ..., a^{(2)}(x^N)]$. Then, we cast the problem in Eq.(5) into the following optimization form:

$$\begin{aligned} \min_{W,b} E &= -Lg(f(W^1, ..., W^N), b) \\ &= -\sum_{(x,y)\in S,D} \log(\rho_{p,n}(x, y|f(w), b)) \end{aligned} \tag{7}$$

We adopt gradient descend algorithm for optimization, in which the gradient $\partial E/\partial W^c, \partial E/\partial b$, are computed by back propagation.

$$\begin{aligned} \partial E/\partial W_{i,k}^{c(2)} &= \sum_{(x,y)\in S,D} \delta_{p,n}(x, y)(\delta_{i,k}(x^c) - \delta_{i,k}(y^c)) \\ \partial E/\partial W_{j,i}^{c(1)} &= \sum_{(x,y)\in S,D} \delta_{p,n}(x, y)(\delta_{i,j}^{(1)}(x^c) - \delta_{i,j}^{(1)}(y^c) \\ \partial E/\partial b &= \sum_{(x,y)\in S,D} \delta_{p,n}(x, y) \end{aligned} \tag{8}$$

where

$$\begin{aligned} \delta_k^{(2)}(t) &= 2(a_k^{(2)}(x^c) - a_k^{(2)}(y^c))f'^{(2)}(z_k^{(2)}(t)) \\ \delta_{i,j}^{(1)}(t) &= t_j \sum_{k=1}^{d_3} \delta_k^{(2)}(t) W_{i,k}^{(2)} f'^{(1)}(z_i^{(1)}(t)) \\ \delta_{i,k}(t) &= \delta_k^{(2)}(t) a_i^{(1)}(t), \ t \in \{x^c, y^c\} \\ \delta_{p,n}(x, y) &= l - \sigma(b - d(a^{(2)}(x), a^{(2)}(y))), l \in \{0, 1\} \end{aligned} \tag{9}$$

With the gradient, the parameter $w$ and $b$ can be updated by Eq.(10) until convergence. The convergence criterion we adopt is that the decrement of cost function E is less than a constant $\varepsilon$. To speed up the learning process, we subdivide the dataset into mini-batches and use the mini-batch to update the modular DNN. Each mini-batch contains 500 samples. $\alpha$ is the learning rate. We can adjust it according to the gradient, so that the increment of the variable $\alpha \times (\partial E/\partial)$ is comparative to the original value. We set $\alpha$ at 0.001 in our experiment.

$$\begin{aligned} W^c &= W^c - \alpha \times (\partial E/\partial W^c), \ (c = 1, ..., N) \\ b &= b - \alpha \times (\partial E/\partial b) \end{aligned} \tag{10}$$

## 3. EXPERIMENT

### 3.1. Dataset and Experiment setting

LFW contains 13,233 face images of 5749 persons, which are from images in the wild, taken from Yahoo! News [10]. The dataset is divided into ten folds where the subject identities are mutually exclusive. Under the unrestricted setting, the identity information of each image is available. We can generate image pairs with arbitrary two images in each fold.

All images are cropped to 80*150 by cutting the center of the aligned images, as in Fig 2. To verify the robustness for different kinds of features, our experiments are performed on four descriptors: Dense-Sift [11], Sift of Nine Key Points(9-Sift) [1], LBP [2] and HLBP [3]. We divide the image into $6 * 3$ non-overlapping blocks, and histogram features are extracted to form the block-wise descriptors. For 9-Sift, we use a key point as a block. Before training, we apply Principal Components Analysis (PCA) to the descriptors and normalize them to zero mean and unit variance. The target dimension of PCA is determined as 75 by experiment. We set the number of units for the three-layer DNN as 75, 128, and 128 respectively. Besides, the threshold $T$ compared with the similarity

| | 1000 | 2000 | 3000 |
|---|---|---|---|
| LBP | 85.32 ±0.32 | 88.23 ± 0.36 | 89.43 ±0.42 |
| HLBP | 86.43 ± 0.39 | 88.95± 0.40 | **89.70** ±0.35 |
| 9-SIFT | 85.97±0.31 | 87.24±0.34 | 87.98± 0.32 |
| DSIFT | 86.93±0.38 | 87.95±0.36 | 89.25±0.40 |
| Combined | 89.05±0.35 | 89.98±0.41 | **93.11**±0.34 |

**Table 1**. Accuracy of different features with increasing number of training samples on LFW (%).

| Method | û ±$S_E$ |
|---|---|
| LDML-MkNN [1] | 87.50±0.40 |
| combined PLDA [12] | 90.07 ± 0.51 |
| SLBP [13] | 90.00 ± 1.33 |
| CMD [13] | 91.70 ± 1.10 |
| CMD+SLBP [13] | 92.58 ± 1.36 |
| **MHFL**(This work) | **93.11** ± 0.34 |

**Table 2**. Comparison of our method with current state-of-the-art methods on LFW (%).

for face verification in Fig 2 is set to 0.5. The performance is measured by ten-fold cross validation.

### 3.2. Result and Analysis

Table 1 shows the performance of the modular hierarchical feature learning(MHFL) on LFW with an increasing number of training pairs:1000, 2000, 3000. The best result for a single descriptor (HLBP) is 89.7%. The last line of Table 1 gives the accuracy derived from fusing the similarity of different descriptors with a linear SVM. For the fusing, we use the similarity scores computed from the four descriptors to construct a four dimension feature vector for each sample. Then, we train SVM with the similarity-based features on the training pairs. As different descriptors capture complementary information, accuracy can be improved by combining these scores. The highest performance of our method is 93.11%, which outperforms the state-of-art method in Table 2. The best reported result is 92.58%. That system uses features of high dimension and considers all possible pairs generated from training samples which are much more than 3000 pairs per fold used in our method. Our performance can be further improved by exploring a better settings of hyper-parameter and using more training samples.

Besides, we give a detail analysis of the hyper-parameter: the PCA dimension for the original input. Training data are usually pixels or hand-crafted descriptors, which generally contain much redundant information. To solve this problem, we apply PCA to all the input data. This operation has been verified to be very effective and improve the performance by 2%-3%. We perform experiments on 2000 training pairs for

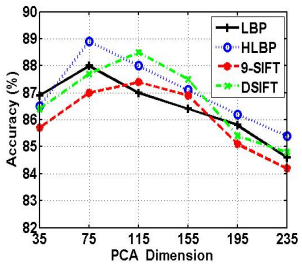


**Fig. 3**. Performance with the PCA dimension for the input

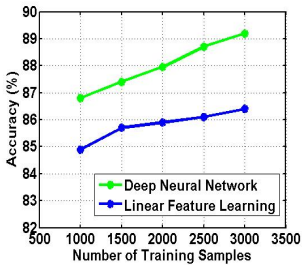


**Fig. 4**. Comparison of DNN and Linear Feature Learning

this section. Fig 3 shows that 75 to 115 is a good range for all the descriptors. In order to simplify the learning, We use 75 as a compromised choice for our model. With high dimension, the system may be disturbed by redundant noise, while with too low dimension, the performance may drop because of losing much discriminative information.

In addition, Fig 4 further reports the advantage of our deep nonlinear feature learning over the linear learning method. We take logistic discriminant metric learning(LDML) [1] as the linear example, and perform experiments on the DSIFT descriptor. With increasing number of training samples, deep nonlinear learning does better than linear learning, and the gap becomes larger. The result indicates that with hierarchical structure, the deep model captures high correlation between the variables more efficiently than the linear model.

## 4. CONCLUSION

In this paper, we propose a novel feature learning method by exploiting deep neural networks to face verification. We use stacked RBMs as the basic network and assemble all the local DNNs to form the modular DNN for the image, which is capable of learning complementary, hierarchical representations. We combine deep learning with side information constraints during optimization process to develop good discriminative ability. Experimental results show the effectiveness of our method for face verification on LFW dataset.

## 5. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under grants No.61172103, No.612714 29, and No.60933010. And this work was also supported by Tencent Research Institute of Beijing, and the project of Patent Analysis Working Platform for Technology Transfer under the No.PXM 2012-178214-000111.

## 6. REFERENCES

[1] Matthieu Guillaumin, Jakob J. Verbeek, and Cordelia Schmid, "Is that you? Metric learning approaches for

face identification," in *ICCV*, 2009, pp. 498–505.

[2] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen, "Face Recognition with Local Binary Patterns," in *ECCV*, 2004, pp. 469–481.

[3] Zhenhua Guo, Lei Zhang, David Zhang, and Xuanqin Mou, "Hierarchical multiscale LBP for face and palmprint recognition," in *ICIP*, 2010, pp. 4521–4524.

[4] Vinod Nair and Geoffrey E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *ICML*, 2010, pp. 807–814.

[5] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *ICML*, 2009, pp. 77–616.

[6] Gary B Huang, Honglak Lee, and Erik Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2518–2525.

[7] S. M. Ali Eslami, Nicolas Heess, and John M. Winn, "The shape boltzmann machine: A strong model of object shape.," in *CVPR*, 2012, pp. 406–413.

[8] Geoffrey E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504 – 507, 2006.

[10] Gary B. Huang, Manu Ramesh, Tamara Berg, Erik Learned-Miller, and Allen Hanson, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," Tech. Rep., 2007.

[11] David G. Lowe, "Distinctive Image Features from Scale Invariant Keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[12] Peng Li, Umar Mohammed, James Elder, and Simon Prince, "Probabilistic Models for Inference about Identity," *PAMI*, vol. 34, pp. 144–157, 2012.

[13] S.Zhu C.Huang and K.Yu, "Large Scale Strongly Supervised Ensemble Metric Learning, with Applications to Face Verification and Retrieval," Tech. Rep., 2011.