Attention Modeling for Face Recognition via Deep Learning

Sheng-hua Zhong (csshzhong@comp.polyu.edu.hk)
Department of Computing, Hung Hom, Kowloon
Hong Kong, 999077 CHINA
Yan Liu (csyliu@comp.polyu.edu.hk)
Department of Computing, Hung Hom, Kowloon
Hong Kong, 99907 CHINA
Yao Zhang (csyaozhang@comp.polyu.edu.hk)
Department of Computing, Hung Hom, Kowloon
Hong Kong, 99907 CHINA
Fu-lai Chung (cskchung@comp.polyu.edu.hk)
Department of Computing, Hung Hom, Kowloon
Hong Kong, 99907 CHINA

Abstract

Face recognition is an important area of research in cognitive science and machine learning. This is the first paper utilizing deep learning techniques to model human's attention for face recognition. In our attention model based on bilinear deep belief network (DBDN), the discriminant information is maximized in a frame of simulating the human visual cortex and human's perception. Comparative experiments demonstrate that from recognition accuracy our deep learning model outperforms both representative benchmark models and existing bio-inspired models. Furthermore, our model is able to automatically abstract and emphasize the important facial features and patterns which are consistent with the human's attention map.

Keywords: face recognition; attention model; deep learning.

Introduction

Face recognition plays an important role in the social life and attracts interest from a very broad range of researchers and scientists (Anderson, 1998). In machine learning and computer vision areas, face recognition using computational models is a classical problem. The representative models include: Eigenface (Turk, et al., 1991), Fisherfaces (Belhumer, et al., 1997), support vector machine (SVM) (Müller, et al., 2001), and so on.

In cognitive science, face recognition is a vividly researched area (Gauthier, et al., 2000) (Afraz, et al., 2006) (Civile, et al. 2011). It is argued that face perception is involved in a unique cognitive process compared with non-face object or scene perception. Researchers in cognitive science seek to understand how the visual system transforms a face image from an initial, pixel-like representation, to a new powerful form of representation, and finally induce the selectively response of the neurons in inferior temporal cortex (Afraz, et al., 2006). Hence, the researchers have utilized some signal processing techniques to simulate the response of human visual system, such as human's attention allocation. Computational attention model is utilized to measure of the conspicuity and provide the predictions about which regions are likely to attract observers' attention (Koch et al., 1985) (Parkhurst et al., 2002). Many empirical validations have demonstrated that attention models have notable ability in various tasks, such as content aware resizing (Avidan et al., 2007), quality assessment (Zhong et al., 2010), and face recognition (Cappelli, et al., 2007) (Fang, et al., 2011).

This paper models human's attention for face recognition via deep learning technique. Deep learning models the learning tasks using deep architectures composed of multiple layers of parameterized nonlinear modules. Deep model is selected in this paper because of two considerations. First, the multiple layers deep architecture is consistent with the laminar structure of human's brain cortex and the information delivery in deep model simulates human's visual cortex. Second, deep learning has demonstrated distinguished ability of information abstraction and robust performance of data classification in various visual data analysis tasks (Hinton, et al., 2006).

This is the first paper utilizing deep learning techniques to model human's attention for face recognition. Compared with existing face recognition models, our proposed bilinear deep belief network (BDBN) has several attractive characters:

1) BDBN maximizes the discriminant information in a frame of simulating the human visual cortex and human's perception. As we known, nearly all existing machine learning models aims to find the discriminant solution to face recognition applications. Existing computational cognitive model emphasizes the identity between the model and the human visual system. Our model attempts to integrate the advantages of both techniques and provide a new thought of this problem.

2) Compared with existing computational face recognition models or representative attention models, BDBN has the ability to automatically extract and emphasize the important facial features and patterns which are consistent with the human attention map.

3) BDBN includes three learning stages: semiconducting bilinear discriminant initialization, greedy layer-wise reconstruction, and global fine-turning. The rational of

three-stage learning comes from the phenomenon of two peaks activation in visual cortex areas. With regard to object recognition, the early peak is related to the activation of an "initial guess" based on the acquired discriminative knowledge, while the late peak reflects the post-recognition activation of conceptual knowledge related to the recognized object.

Model

In this section, we design a deep learning algorithm with a deep architecture for the task of face recognition, includes bilinear discriminant initialization, greedy layer-wise reconstruction and global fine-tuning. The strategy of bilinear discriminant projection is utilized to construct a projection to map the original data into a discriminant preserving subspace. And it determines the initial parameters and sizes of the upper layer. To human, this strategy is consistent with the early peak related to the activation of "initial guess". In the stage of greedy layer-wise reconstruction, the parameter space is refined by the greedy layer-wise information reconstruction using Restricted Boltzmann Machines (RBMs) (Smolensky, 1986) as building blocks. In the stage of global fine-tuning, we refine the parameter space for better face recognition performance. And it is consistent with the late peak related to the activation of "post-recognition". After the deep learning model is constructed, the attention map is built based on the parameter space in the first RBM.

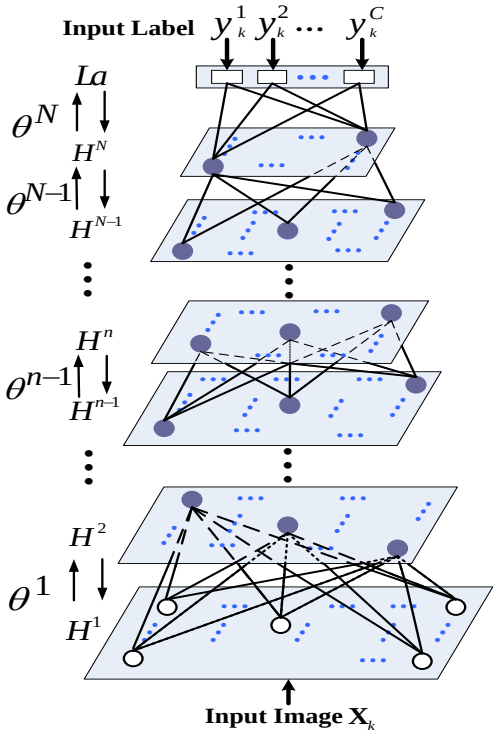


Figure 1: Architecture of the bilinear deep belief network.

Figure 1 shows the architecture of our bilinear deep belief network. A fully interconnected directed belief network includes input layer H^1, hidden layer H^2,…, H^N, and one label layer La at the top. The input layer H^1 has $I \times J$ units, and this size is equal to the dimension of the input features. In our model, we use the pixel values of sample datum $\mathbf{X}_k$ as the original input features. In the top, the label layer has C units, which is equal to the number of classes. The search of the mapping function from X to Y is transformed to the problem of finding the optimum parameter space θ^* for the deep architecture.

In our deep learning architecture, X is a set of data samples, $X=[\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_k,...,\mathbf{X}_K]$. $\mathbf{X}_k$ is a sample datum in the image space $\mathbb{R}^{I\times J}$ and K is the number of sample data. Y is a set of labels corresponding to X, $Y=[\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_k,...,\mathbf{y}_K]$. $\mathbf{y}_k$ is the label vector of $\mathbf{X}_k$ in $\mathbb{R}^C$, $y_k^c=\begin{cases}1 & \text{if } \mathbf{X}_k \in c\text{th class}\\ 0 & \text{if } \mathbf{X}_k \notin c\text{th class}\end{cases}$, where C is the number of classes. Based on the given training set, the aim in face recognition is to learn a mapping function from the image set X to the label set Y, and then recognize the new coming face images according to the learned mapping function.

Bilinear Discriminant Initialization

In order to preserve the discriminant information in the learning procedure, the objective function of bilinear discriminant initialization could be represented as follows:

$$\arg\max_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V})=\sum_{s,t=1}^{K}\|\mathbf{U}^T(\mathbf{X}_s-\mathbf{X}_t)\mathbf{V}\|^2(\alpha\mathbf{B}_{st}-(1-\alpha)\mathbf{W}_{st}) \quad (1)$$
$$s.t.\ \mathbf{U}^T\mathbf{U}=\mathbf{I}_P,\ \ \mathbf{V}^T\mathbf{V}=\mathbf{I}_Q$$

where balance weight $\alpha\in[0,1]$ is the parameter used to balance the between-class weights $\mathbf{B}_{st}$ and the within class weights $\mathbf{W}_{st}$, which are defined as follows (Yan, et al., 2007) (Sugiyama, 2007).

By simultaneously maximizing the distances between data points from different classes and minimizing the distances between data points from the same class, the discriminant information is preserved to the greatest extent in the projected feature space. Solving $\mathbf{U}$ (or $\mathbf{V}$) with fixed $\mathbf{V}$ (or $\mathbf{U}$) is a convex optimization problem. Let $\mathbf{E}_{st}=\alpha\mathbf{B}_{st}-(1-\alpha)\mathbf{W}_{st}$, with the fixed $\mathbf{V}$. The optimal $\mathbf{U}$ is composed of the first P eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_\mathbf{V}\mathbf{u}=\lambda\mathbf{u} \quad (2)$$

where $\mathbf{D}_\mathbf{V}=\sum_{st}\mathbf{E}_{st}(\mathbf{X}_s-\mathbf{X}_t)\mathbf{V}\mathbf{V}^T(\mathbf{X}_s-\mathbf{X}_t)^T$. Similarly, with the fixed $\mathbf{U}$, the optimal $\mathbf{V}$ is composed of the first Q eigenvectors of the following eigendecomposition problem:

$$\mathbf{D}_\mathbf{U}\mathbf{v}=\lambda\mathbf{v} \quad (3)$$

where $\mathbf{D}_\mathbf{U}=\sum_{st}\mathbf{E}_{st}(\mathbf{X}_s-\mathbf{X}_t)^T\mathbf{U}\mathbf{U}^T(\mathbf{X}_s-\mathbf{X}_t)$.

The above steps monotonically increase $J(\mathbf{U},\mathbf{V})$ and since the function is upper bounded, it will converge to a critical point with transformation matrices $\mathbf{U}$, $\mathbf{V}$.

By bilinear discriminant initialization, we obtain the discriminant initial connections in layer pair and utilize the optimal dimension to define the structure of the next layer.

$$A^{n}_{ij,pq}(0) = (\mathbf{U}^{n}_{ip})^{T}\mathbf{V}^{n}_{jq} \tag{4}$$

$$P^{n+1} = row(\mathbf{U}^{n}),\ Q^{n+1} = column(\mathbf{V}^{n}) \tag{5}$$

Greedy Layer-Wise Reconstruction

In this section, we describe how to construct the first RBM between the input layer H^1 and the first hidden layer H^2.

The energy of the state $(\mathbf{h}^1, \mathbf{h}^2)$ in the first RBM is:

$$E\left(\mathbf{h}^1, \mathbf{h}^2; \theta^1\right) = -(\mathbf{h}^1\mathbf{A}^1\mathbf{h}^2 + \mathbf{b}^1\mathbf{h}^1 + \mathbf{c}^1\mathbf{h}^2) \tag{6}$$

where $\theta^1 = \left(\mathbf{A}^1, \mathbf{b}^1, \mathbf{c}^1\right)$ are the model parameters between the input layer H^1 and first hidden layer H^2. Therefore, the log-likelihood probability of the model assigned to $\mathbf{h}^1$ in H^1 is:

$$\log P\left(\mathbf{h}^1\right) = \log \sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1, \mathbf{h}^2; \theta^1\right)} - \log \sum_{\mathbf{h}^1}\sum_{\mathbf{h}^2} e^{-E\left(\mathbf{h}^1, \mathbf{h}^2; \theta^1\right)} \tag{7}$$

By calculating the derivative of Equation (8), we could update the parameter space with respect to the parameter $\theta^1 = \left(\mathbf{A}^1, \mathbf{b}^1, \mathbf{c}^1\right)$.

$$\frac{\partial \log p(\mathbf{h}^1(0))}{\partial \theta^1} = -\sum_{\mathbf{h}^2(0)} p(\mathbf{h}^2(0) \mid \mathbf{h}^1(0)) \frac{\partial E(\mathbf{h}^2(0), \mathbf{h}^1(0))}{\partial \theta^1} + \sum_{\mathbf{h}^2(t)}\sum_{\mathbf{h}^1(t)} p(\mathbf{h}^2(t), \mathbf{h}^1(t)) \frac{\partial E(\mathbf{h}^2(t), \mathbf{h}^1(t))}{\partial \theta^1} \tag{8}$$

The above discussion is the greedy layer-wise abstraction for the first layer H^1 with its next adjacent layer H^2. Similar operations can be performed on the higher layer pairs.

Global Fine-Tuning

In this section, we use backpropagation to adjust the entire deep network to find good local optimum parameters $\theta = \mathbf{[A,b,c]}$ by minimizing the recognition error $[-\sum_{l} \mathbf{y}_l \log \hat{\mathbf{y}}_l]$, where $\mathbf{y}_l$ and $\hat{\mathbf{y}}_l$ are the correct recognition label and the output recognition label value of labeled sample datum $\mathbf{X}_l$ in X^L.

Above, we utilize the greedy layer-by-layer algorithm to learn a deep model with the help of discriminant information obtained from bilinear discriminant projection. Therefore, the convergence in our algorithm obtained from backpropagation is not slow. And the result generally converges to a good local minimum on the error surface.

Attention Modeling

The weights of first layer of BDBN are oriented, Gabor-like and resemble the receptive fields of V1 simple cell (Zhong, et al., 2011). Therefore, the first RBM is utilized to construct the attention model which is shown in Figure 2. To every neuron in the input layer, the weight value to the one in the first hidden layer is calculated as feature map. Then, the weight value of every neuron is normalized and combined into an attention map.

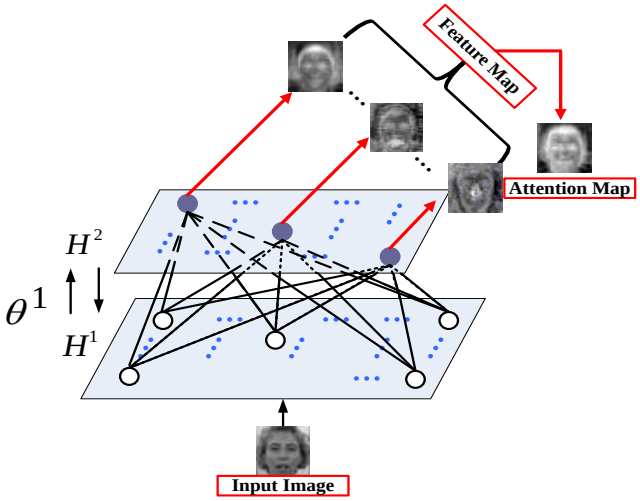


Figure 2: Construct attention model by first RBM in bilinear deep belief network.

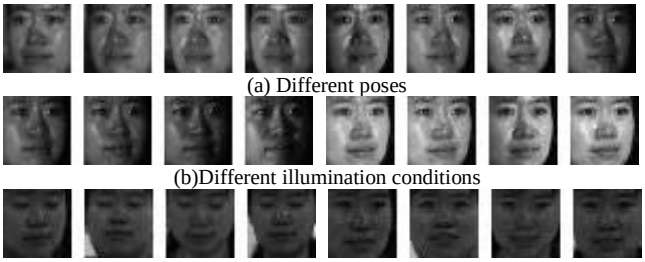
Experiment 1: Recognition Accuracy Analysis

Dataset

The CMU PIE face dataset collected between October and December 2000 contains 68 subjects with a total of 41,368 face images (Sim, et al., 2002). The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination and expression. In the first experiment, we use all the images under different illuminations and expressions with five near frontal poses (C05, C07, C09, C27, C29). Thus we obtain 170 images for each individual.

Procedure

For the CMU PIE face dataset, the preprocessing is applied following the general setting of experiment (He, et al., 2005). Original images are normalized (in scale and orientation) so that the two eyes are aligned at the same position. Then, the facial areas are cropped into the final images for matching. The size of each cropped image in all of the experiments is 32×32 pixels. Sample images after preprocessing are shown in Figure 3.



(a) Different poses

(b)Different illumination conditions

(c) Different facial expressions

Figure 3: Sample images after preprocessing from CMU PIE.

In our experiments, the balance weight of our model is set as 0.5 for simplicity. For parameters such as the learning rate and the momentum, we simply follow the general setting of previous work on computational deep networks (Bengio, et al., 2006), although more careful choice may lead to better performance. In the fine-tuning stage, the method of conjugate gradients is utilized and three line searches are performed in each epoch until convergence.

To adapt the real-world face recognition tasks, our computational neuroscience model BDBN is applied under a semi-supervised learning framework. It makes face recognition work well when labeled images are insufficient. 120 images are randomly selected for each person to form the training set and the rest to form the test set. Of the 120 images for each person, different numbers of images are randomly selected and labeled while the others remain unlabeled. The number of labeled data per subject is equal to 5, 10, 20 and 40, respectively. We perform 10 random splits and report the average results over the 10 trials.

Experimental Results

In the first experiment, we compare three representative face recognition models, including: Eigenface (Turk, et al., 1991), Fisherfaces (Belhumer, et al., 1997), SVM (Müller, et al., 2001), and existing bio-inspired Sparse Localized Features (SLF) model (Mutch and Lowe, 2008).

SVM, Eigenface, and Fisherfaces both are representative benchmark machine learning models for the task of face recognition. The bio-inspired Sparse Localized Features (SLF) (Mutch and Lowe, 2008) are an extensions of the C2 features from the Serre et al. HMAX model (Serre, et al., 2007). For this representation, we took advantage of the MATLAB code provided by the authors. Here, the SVM classification was based on a linear kernel with normalized training and testing data (zero-mean and unit-variance feature-wise).

The recognition accuracy rate with different numbers of labeled data is shown in Table 1. As shown in Table 1, the recognition accuracy rate of bio-inspired models SLF+SVM and BDBN is better than machine learning models Eigenfaces and SVM. And our proposed BDBN has the best performance than others.

Table 1: Recognition accuracy rate (%) on the test data with different numbers of labeled data per category on CMU PIE.

Num./Cat.	20	30	40	50
Eigenfaces	61.9±0.7	72.1±0.6	78.2±0.5	83.8±0.4
Fisherfaces	84.5±0.7	92.0±0.6	93.1±0.5	94.8±0.3
SVM	73.5±0.6	80.4±0.5	82.9±0.5	87.1±0.3
SLF+SVM	80.5±0.6	86.8±0.5	89.5±0.5	90.2±0.3
Semi_DBN	85.4±0.7	92.4±0.6	93.5±0.5	95.0±0.3
BDBN	88.4±0.7	93.9±0.6	94.3±0.5	96.6±0.3

Experiment 2: Face Feature Points Emphasis

Dataset

The BioID face dataset consists of 1521 gray level images collected contains 23 subjects (HumanScan, 2003). The face images in BioID are under a large variety of illumination, background.

In this dataset, the x and the y coordinate of the left eye and the right eye are provided. Furthermore, the 20 important facial feature points are manually placed, including: right eye pupil, left eye pupil, right mouth corner, left mouth corner, outer end of right eye brow, inner end of right eye brow, inner end of left eye brow, outer end of left eye brow, right temple, outer corner of right eye, inner corner of right eye, inner corner of left eye, outer corner of left eye, left temple, tip of nose, right nostril, left nostril, centre point on outer edge of upper lip, centre point on outer edge of lower lip, and tip of chin. These facial feature points are thought to be very useful for facial analysis and gesture recognition (Jesorsky, et al., 2001) (Wang, et al., 2002) (Cappelli, et al., 2007).

Procedure

As a deep learning model for face recognition, BDBN has demonstrated the impressive recognition performance in this first experiment. In this experiment, we intend to investigate the consistency between the emphasized regions in BDBN and the attention map of human being.

The number of images in every category of BioID is varied, from 35 to 118. Therefore, firstly, we choose the categories with more than 50 face images as the subset we work on. Then, just like the procedure on face datasets, the original images are normalized (in scale and orientation) so that the two eyes are aligned at the same position. Finally, the facial areas are cropped and downsampled into the final images. The size of each final image in all of the experiments is 32×32 pixels, with 256 gray levels per pixel. Some sample images after preprocessing are shown in Figure 4.

Figure 4: Sample images after preprocessing from BioID.

Then, to every image in BioID face dataset, we directly input the original pixel value to the BDBN model. After

bilinear discriminant initialization and layer wise reconstruction, we evaluate the consistency between the constructed attention model and human's attention map.

Experimental Results

Computational attention model was called saliency map first appeared in (Koch, et al., 1985). Typically, multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted at multiple scales. After a feature map is computed for each of the features, they are normalized and combined into a master saliency map that represents the saliency of each pixel.

To face images, some facial areas are assessed to be attracted more attention and helpful to face recognition, for example eye, ear, nose and mouth (Hickman, et al., 2010). Fortunately, in this dataset, 20 important facial feature points are manually selected out and placed. Therefore, with marked facial feature points, the attention model based on deep learning model could be evaluated without eye tracking recordings.

Different from representative attention map which utilizes various features such as intensity, color, orientation, only the gray level pixel values are input into our model. Our attention model automatically extracts and emphasizes important features and patterns to construct facial attention model.

To demonstrate the effectiveness of our model, firstly, the visualization of the parameter space of proposed model is observed. Figure 5 (a) shows a sample image, and Figure 5 (b) shows the sample image with the facial feature points. Figure 5 (c) visualizes the parameter spaces between the input layers and the first hidden layer in BDBN. Each picture shown below represents one neuron in the hidden layer and each pixel quantizes the weight value between that neuron and the one in the input layer. Obviously, the proposed BDBN can automatically extract and emphasize the important areas of human's face, such as the eyes, eyebrows, noses, cheeks, mouths and chins.

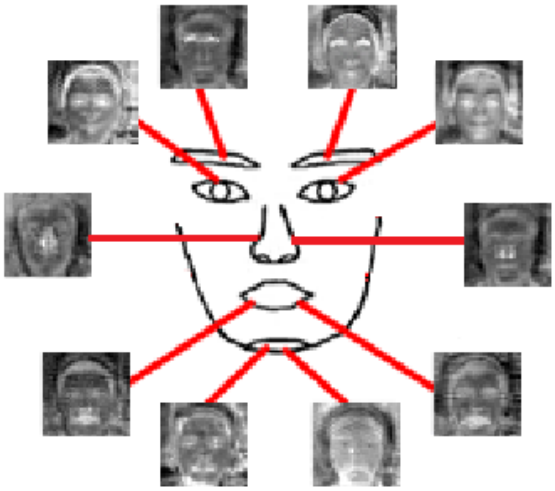
Then, we construct the saliency regions based on the emphasized regions of BDBN. Just like the Figure 5 (c), the weight value between each neuron in the input layer to the one in the first hidden layer is calculated at first. Then, the weight value of every neuron is normalized and combined into a saliency map. According to the x and the y coordinates of the 20 important facial feature points of every face image in the dataset, we statistically analyze the percentage of all facial feature points located in the saliency regions of the saliency map.

There are 63.71% facial feature points are located inside 30% most saliency regions and only about 1% facial feature points are located outside 80% most saliency regions. It is obviously that proposed BDBN covers most of important facial feature points. From Figure 4, some of other information and regions are useful to recognize people, such as the hairstyles and the face contour, although they are not belong to the 20 important facial feature points. And as shown in Figure 5 (c), these information and regions are also emphasized in the parameter space of proposed model. Therefore, if the importance from other important regions for face recognition is excluded, the facial feature points cover percentage in saliency map will be much better.

In Figure 6, the comparison of different computational attention maps are provided, including Graph Gabor attention map (Harel, et al., 2006), Itti classical attention map (Itti & Koch, 2000) and BDBN attention map. It is obviously that BDBN has better coverage than other models. It proves that BDBN provides a human-like judgment by referencing the human visual system.

(a)Sample face image

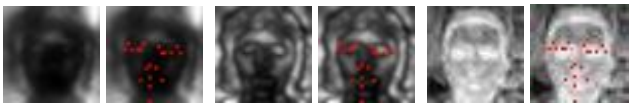
(b) Face image with facial feature points



(c) Emphasized regions of BDBN corresponding to facial feature regions

Figure 5: Samples of first layer weights learned by BDBN, and the consistency of these weights with facial feature points.

(a)Sample face image with facial feature points



(b) Graph Gabor map (c) Itti map (d) BDBN map

Figure 6: The comparison of different attention maps with facial feature points.

Conclusion and Future Work

In this paper, we make an attempt to construct an attention model for face recognition in a frame of simulating the human visual cortex and human's perception. To evaluate proposed face recognition models, we do experiments on two face images' datasets, CMU PIE and BioID. Experiments results not only show the distinguishing recognition ability of our deep model but also clearly demonstrate our intention of providing a human-like face image analysis by referencing the human visual cortex and perception procedure.

It is the general opinion that advances in cognitive science especially neuroscience will provide useful insights to computer scientists into how computer models construct, and vice versa. To a certain extent our attempt is an example to prove that the computational models are not only applied into the tasks of classification and recognition just as the optimal classifier, they also can provide human-like response by referencing the human visual system. In future, we will go on this direction to propose novel computational model by referring more characters of human visual system. And vice versa, in cognitive science, we will explore whether the human visual system possess the related mechanism which is consistent with the computational model from the viewpoint of mathematics.

References

Anderson, JR, (1998). Social stimuli and social rewards in primate learning and cognition. *Behavioural Processes* (pp. 159–175).

Afraz, SR, Kiani, R. and Esteky, H., (2006) Nature, 442, (pp. 692–695).

Avidan, S. and Shamir, A. (2007). Seam carving for content-aware image resizing", In *ACM Transactions on Graphics*.

Belhumer, P., Hespanha, P., and Kriegman, D., (1997). Eigenfaecs vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, (pp.711-720).

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., (2006). Greedy layer-wise training of deep networks, *Advances in Neural Information Processing Systems*.

Cappelli, R., Franco, A. Maio, D. (2007). Gabor Saliency Map for Face Recognition, In *Proceedings of the 14th International Conference on Image Analysis and Processing*, 443-447.

Civile, Ciro, McLaren, R.P., McLaren, L.P.L., (2011), Perceptual learning and face recognition: Disruption of second order relational information reduces the face inversion effect. In *33th annual meeting of the Cognitive Science Society*, 2083-2088.

Fang, F., Qing, L.Y., Wang, C.X., Miao J., Chen X.L., Gao, W.. (2011). Attention Driven Face Recognition, Learning from Human Vision System, In *International Journal of Computer Science Issues*.

Felleman, D. J., Van Essen, D. C., (1991). Distributed hierarchical processing in the primate cerebral cortex. In *Cereb. Cortex*.

Gauthier, I., Skudlarski, P., Gore, J.C., & Anderson, A.W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3 (2): 191–197.

Harel, J., Koch, C. and Perona, P.. (2006). Graph-Based Visual Saliency In *NIPS*.

He, XF, Cai, D., and Niyogi, P., (2005). Tensor subspace analysis, *Advances in Neural Information Processing Systems*.

Hickman, L. Firestone, AR, Beck, FM, and Speer, S., (2010). Eye fixations when viewing faces. *Journal of the american dental association jada electronic resource*, (pp. 40–46).

Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. In *Science*.

Hinton, G. E., (2007). Learning Multiple Layers of Representation. In *Trends. Cogn. Sci*.

HumanScan, (2003). BioID face database. https://www.bioid.com/download-center/software/bioid-face-database.html.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. In *Vision Res.*.

Jesorsky, O., Kirchberg, K., Frischholz, R.. (2001). Robust face detection using the hausdorff distance. In *Proceedings of the 3th International Conference on Audio- and Video-based Biometric Person Authentication*.

Koch, C. & Ullman, S.. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry, In *Human Neurobiology*. pp. 219-227.

Müller, KR, Mika, S., Rätsch, G., Tsuda, K., and BSchölkopf, (2001). An introduction to Kernel-based learning algorithms, *IEEE Transactions on Neural Networks*, vol. 12, no. 2, (pp 181-201).

Mutch, J. and Lowe, DG, (2008). Object class recognition and localization using sparse features with limited receptive fields, *International Journal of Computer Vision*.

Parkhurst, K. Law, and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. In *Vision Res.*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T., (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sim, T., Baker, S., and Bsat, M., (2002). The CMU Pose, Illumination, and Expression (PIE) Database, *Proceedings of IEEE International conference on Automatic Face and Gesture Recognition*.

Smolensky, P.. (1986). Information processing in dynamical systems: foundations of harmony theory. In *Parallel Distributed Processing: Explorations in The Microstructure of Cognition*, vol. 1: Foundations, MIT Press, (pp. 194-281).

Sugiyama, M., (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. In *JMLR*.

Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, Lahaina, Maui, Hawaii, (pp. 586–591).

Yan, S., Xu, D. Zhang, B., Zhang, H.J., Yang, Q., and Lin, S., (2007). Graph embedding and extension: a general framework for dimensionality reduction. In *PAMI*.

Yang, P., Shan, SG, Gao, W., Li. SZ, Zhang, D. (2004). Face recognition using Ada-Boosted Gabor features, *Automatic Face and Gesture Recognition*, (pp. 356-361).

Wang, Y. Chua, C., Ho,Y., (2002). Facial feature detection and face recognition from 2D and 3D images, In *Pattern Recognition Letters*. 1191–1120.

Zhong, S.H., Liu, Y., Liu, Y. and Chung, F.L.. (2010). A semantic no-reference image sharpness metric based on top-down and bottom-up saliency map modeling. In *IEEE International Conference on Image Processing*.

Zhong, S.H., Liu, Y., Liu, Y..(2011). Bilinear deep learning for image classification. In *Proceedings of the 19th ACM International Conference on Multimedia*.