Using CMU PIE Human Face Database to a Convolutional Neural Network - Neocognitron

José Hiroki Saito¹, Tiago Vieira de Carvalho¹, Marcelo Hirakuri¹, André Saunite¹, Alessandro Noriaki Ide² and Sandra Abib¹

1- Federal University of São Carlos - Computer Science Department - GAPIS Rodovia Washington Luis, Km 235, São Carlos – SP - Brazil {saito,hirakuri,sabib}@dc.ufscar.br, tiagocarvalho@uol.com.br, saunite@fai.com.br

2- University of Genoa - Department of Informatics, Systems and Telematics - Neurolab Via Opera Pia, 13 – I-16145 – Genoa - Italy noriaki@dist.unige.it

Abstract

This work presents the application of Neocognitron to the human face recognition. Using a large-scale human face database (CMU PIE), the optimal thresholds of the Neocognitron to human face recognition are verified. During the first experiment, increasing the activation thresholds of the Neocognitron, their best values to be used in the second experiment, increasing the number of training images per subjects, are obtained. As a result it is verified that a number of 25 training images per subjects is enough to obtain very high recognition rate (98%) to the frontal pose images from the database. 350 validation images, non-overlapping with the training images, were used.

1. Introduction

Convolutional neural networks, used in the field of speech and image analysis, as described in [1], use local weight sharing topology, different from fully connected feedforward networks. Their topology is more similar to biological networks. The present paper discusses the results obtained by Neocognitron, the first convolutional neural network, in the case of human face recognition, using the CMU PIE (Carneggie Mellon University, Pose, Illumination and Expressions) database [2].

Neocognitron, proposed by Fukushima [3] to handwritten character recognition, has been modified through the years and its performance has been improved. Its self-organized learning has obtained a recognition rate of 98.6%, for a blind test sample [5], using a large-scale database ETL-1[4].

Although Neocognitron has been tested intensively to handwritten character recognition, its performance to human face recognition has not significantly been verified, besides NEO [6]. NEO is a Neocognitron-like neural network implemented by Neubauer [6] that combines neurons from perceptron with localized network

structure of the Neocognitron. The following sections describe the Neocognitron neural network, the application of Neocognitron to the face recognition using the CMU PIE database, the results compared to NEO, conclusions, and future works.

2. Neocognitron

Neocognitron is a sequence of stages composed by two layers: S-layer, composed by S-cells, responsible for the feature extraction; and C-layer, composed by C-cells, responsible for the tolerance of shape and position. These cells are grouped in rectangular cell-planes and all cells at the same cell-plane are identical, regardless of their position. An S-cell at position **n**, in a plane k_l , $us_l(k_l, n)$, is calculated by:

$$u_{Sl}(k_{l},\mathbf{n}) = \frac{\theta}{1-\theta} \cdot \varphi \left| \frac{1 + \sum_{l=1}^{k_{l-1}} \sum_{l=l \in S_{l}} a_{l}(k_{l-1},i,k_{l}) \cdot u_{Cl-1}(k_{l-1},\mathbf{n}+i)}{1+\theta \cdot b_{l}(k_{l}) \cdot v_{Cl-1}(\mathbf{n})} - 1 \right|$$
(1)

where, θ is the threshold of the cell activation, and φ is a function defined as:

$$\varphi[x] = \begin{cases} x & if \quad x \ge 0 \\ 0 & if \quad x < 0 \end{cases}$$
(2)

The weight a_l (k_{l-1} , i, k_l) multiplies the input $uc_{l-1}(k_{l-1}, n+i)$, from the preceding stage k_{l-1} , where, i is the index of the inputs from the region S_l of the preceding cellplane. The summation is effected by cell-planes $k_{l-1} = l$ to K_{l-1} of the preceding layer. The weight $b_l(k_l)$ multiplies the inhibition value $v_{cl-1}(n)$, which is calculated by:

$$v_{Cl-1}(\mathbf{n}) = \sqrt{\sum_{k_{l-1}=1}^{K_{l-1}} \sum_{i \in S_l} c_{l-1}(i) \cdot u^2_{Cl-1}(k_{l-1}, \mathbf{n} + i)}$$
(3)

where the c_{l-1} (*i*) is a predefined constant.

A C-cell at position **n**, in a plane k_l , $uc_l(k_b \mathbf{n})$, is calculated by:

$$u_{Cl}(k_{l},\mathbf{n}) = \psi \left[\sum_{i \in S_{l}} d_{l}(i) \cdot u_{Sl}(k_{l},\mathbf{n}+i) \right]$$
(4)

where the function ψ is defined as:

$$\psi[x] = \begin{cases} \frac{x}{\alpha + x} & \text{if } x \ge 0\\ 0 & \text{if } x < 0 \end{cases}$$
(5).

3. Neocognitron Training

During the training phase, it is defined which feature will be recognized by each cellplane of a given S-layer. First, there are no cell-planes in any S-layer. A special cellplane named *seed-selection-plane*, which is able to recognize all features from the preceding layer is used. The training process starts from the first level, and goes through, until all the levels are trained. First, an input pattern is presented to the network. All the features of the input pattern are recognized by the *seed-selection-plane*, in this way many cells can be activated. At this moment, all the activated cells are verified in order to select the most strongly activated cell, which is considered as the *winner*. When the *winner* is selected, its weights are reinforced. After reinforcement the *winner-cell* can recognize the corresponding feature. If all cells in a cell-plane are identical, a cell-plane with all cells identical to the *winner-cell* (*seed-cell*) is created and it becomes a valid cell-plane, or a trained cell-plane. The procedure is repeated, but taking into account that the *winner-cell* cannot be activated coincidently with any previously trained cell-plane. When the coincidence occurs the next strongest cell is to be selected as the *winner*.

After presenting the input patterns many times, and detecting any new feature, the training of a given layer is completed, and the algorithm continues to the next stage, until the training of all the stages is completed.

4. CMU PIE Database

The CMU PIE Database [2] consists of a large number of subjects, each imaged for many times from several different poses, under significant illumination variation, and with a variety set of facial expressions. The database is organized as a collection of images for each subject. 13 cameras were used: 9 cameras in the horizontal sweep, each separated by 22.5⁰. The other 4 cameras are positioned: 2 above and below the central camera, and 2 at the corner of the room. To obtain illumination variation, the 3D room was arranged with a flash system with 21 flashes. Capturing images with the background lighting both on and off, 43 different illumination conditions were obtained. To acquire the variety set of facial expressions of the subjects they were asked to give neutral expressions as: smile, blink and talk. The database consists of 41368 images of 68 subjects, divided in two major partitions, the first with pose and expression variation only, the second with pose and illumination variation, i.e. there is no simultaneous variation in illumination and expression.

5. Simulation for Human Face Recognition

Figure 1 shows the Neocognitron structure used in face recognition. At the leftmost side we can see the input pattern U_0 that consists of a 57x57 pixels image, followed by a contrasted image with the same size. The contrasted image is obtained by a contrast-extracting cell of layer U_G [9]. The contrasted image layer is followed by the reduced image layer, of 20x20 pixels. The reducing of the image, or thinning-out, is obtained by applying a spatial blur, followed by a neighborhood elimination of the resulting cells [5].

The reduced input image is then applied as input to the stage 1, to the U_{SI} layer, whose output is reduced to $U_{S2'}$ 13x13 cell-planes, before the application to the U_{CI} layer. The outputs of the U_{CI} layer are connected to stage 2, as input to the U_{S2} cells. The U_{S2} cell-planes (13x13), are reduced to 7x7, obtaining the $U_{S2'}$ layer, which is used as input to the U_{C2} layer. The last stage starts with the U_{S3} layer of 7x7 cell-planes, followed by the $U_{S3'}$ layer of 3x3 cell-planes, and U_{C3} layer. The output layer is composed of a set of 1x1 (single neuron) cell-planes, each one corresponding to a different class of the input pattern to be classified by the network.



Fig. 1. Implemented Neocognitron Structure.

6. Results

10 subjects from the CMU PIE database (4002, 4014, 4036, 4047, 4048, 4052, 4057,4062, 4063, and 4067) [2] were randomly chosen. As the main goal of the work is the network training, and validation, the talking images of the subjects were selected, because of the existence of 60 images per pose per subject. So, during the experiments, the frontal images of the subjects (640x486) were used, after windowing and scaling of the face region (57x57), as showed in Figure 2.



640x486

Fig.2. CMU PIE database image sample, and the corresponding image sample used during the experiments, after windowing and scaling.

In the first experiment, the best structure was obtained by varying the network size and increasing the training thresholds θ 1, θ 2, and θ 3, from 0.70 to 0.77, 0.65 to 0.72, and 0.60 to 0.67, respectively, using 10 samples per subject during training. Figure 3a shows the plots K1, K2, and K3, which correspond to the number of S-cell planes

from stages 1, 2 and 3, respectively, which grow by increasing their activation thresholds. The last plot (ERROR) shows that as the structure increases the number of misclassifications (ERROR) decrease, and varies more smoothly. The ERROR number was taken coincidently with the number of non-classification, as shown in Figure 3b. The diagram shows the NREC (non-classification number) and ERROR (misclassification number) values which correspond to the variation of the threshold θ^2 during recognition, from 0.40 to 0.61, and fixing the thresholds $\theta 1 = 0.75$, and $\theta 3 = 0.30$, at the smooth region10 training patterns, and 50 validation images, non-overlapping with the training pattern, were used so that the misclassification or non-classification rate is approximately 2% of the 500 validation images at the stability. The classification rate for the 100 training images is 100%.



Fig. 3. Experiment1: Training for 10 images per subject, varying activation thresholds – (a) network structure size (K1, K2, K3) and ERROR, varying the training threshold θ 1, θ 2, and θ 3 (b) recognition response NREC (non-classification) and ERROR (misclassification) at the stable region, varying the recognition threshold θ 2.



Fig. 4. Experiment 2: Varying images per subject (10 to 25), and fixing the training threshold ($\theta 1 = 0.75$, $\theta 2 = 0.70$, and $\theta 3 = 0.65$) – (a) network structure K1, K2, and K3, number of cell-planes per stage, and (b) NREC, number of non-classification, ERROR, misclassification, and TOTAL (non-classification plus misclassification).

In the second experiment the verification of the classification rate was achieved by increasing the number of training patterns per subject. Figure 4a shows the resulting network size, expressed in number of cell-planes (K1, K2 and K3), by varying the number of images per subject used for training and fixing the training threshold to $\theta 1 = 0.75$, $\theta 2 = 0.70$, and $\theta 3 = 0.65$. Figure 4b, shows the network performance. The first plot corresponds to the non-classification number (NC), the second plot is the misclassification number (ERROR), and the third plot is TOTAL, taking into account the validation images, non-overlapping with the training images. It can be seen that the TOTAL (non-classification plus misclassification), which is the inverse of the correct classification, decreases with the increase of the number of training patterns. It can be seen that the best classification rate (98%), misclassification (1.42%), and non-classification (0.57%), were obtained with 25 training samples/subject, corresponding to a network structure of K1 = 129, K2= 186, and K3= 154, cell-planes. The classification rate for the training images was 100%.

7. Summary and Conclusions

In the previous sections, we have shown that the recognition rate obtained during validation is very high, and it can be increased using more training images This result is comparable to the constrained test sample obtained by NEO [6], despite of the different images and input resolution used, and also comparable to the Neocognitron handwritten character recognition [5]. For future works, we may consider other poses, illumination, and expressions, from CMU PIE database, for training and recognition experiments, creating an unconstrained situation described by Neubauer [6].

8. Acknowledgements

The authors would like to thank Prof. Simon Baker from Carneggie Mellon University for the kindness of sending the PIE database, used in this work.

References

- Yann LeCun and Yoshua Bengio. Convolutional Networks for Images, Speech, and Time Series. *The Handbook of Brain Theory and Neural Networks*, M.Arbib, Ed. Cambridge, MA: MIT Press, 2nd ed., pages 276-279, 2003.
- [2] Sim,T.; Baker, S.; Bsat, M. The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces, CMU-RI-TR-01-02, pages. 1-17, USA, 2002.
- [3] Fukushima, K. Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36, pages. 193-202,1 1980.
- [4] ETL1 database: http://www.etl.go.jp/~etlcdb/index.html
- [5] Fukushima, K. Neocognitron for Handwritten Digit Recognition, *Neurocomputing*, V.51, pages 161-180, 2003.
- [6] Neubauer, C. Evaluation of Convolutional Neural Networks for Visual Recognition, *IEEE Transactions on Neural Networks*, Vol. 9, n.4, pages 685-696, 1998.
- [7] Fukushima, K.; Wake, N. Improved Neocognitron with Bend Detecting Cells, International Joint Conference on Neural Networks, pages 190-195, Baltimore-Maryland, 1992.