A Neural Network Facial Expression Recognition System using Unsupervised Local Processing

Leonardo Franco
Alessandro Treves
Cognitive Neuroscience Sector - SISSA
2-4 Via Beirut, Trieste, 34014 Italy
lfranco@sissa.it, ale@sissa.it

Abstract

A local unsupervised processing stage is inserted within a neural network constructed to recognize facial expressions. The stage is applied in order to reduce the dimensionality of the input data while preserving some topological structure. The receptive fields of the neurons in the first hidden layer self-organize according to a local energy function, taking into account the variance of the input pixels. There is just one synapse going out from every input pixel and these weights, connecting the first two layers, are trained with a hebbian algorithm. The structure of the network is completed with specialized modules, trained with backpropagation, that classify the data into the different expression categories. Thus, the neural net architecture includes 4 layers of neurons, that we train and test with images from the Yale Faces Database. We obtain a generalization rate of 84.5% *on unseen faces, similar to the* 83.2% *rate obtained when using a similar system but implementing PCA processing at the initial stage.*

1. Introduction

Face perception is a very important component of human cognition. Faces are rich in information about individual identity, but also about mood and mental state, being accessible windows into the mechanisms governing our emotions. Facial expression interactions are relevant in social life, teacher-student interaction, credibility in different contexts, medicine, etc. Face expression recognition is also useful for designing new interactive devices offering the possibility of new ways for humans to interact with computer systems.

From a neurophysiological point of view face recognition appears to be very important. Experiments both in monkeys and humans show the existence of dedicated areas in the brain where neurons respond selectively to faces ([11, 17, 6]). Also it has been shown that complex visual processing related to discrimination of faces is a very rapid task that can be completed in approximately 100 msec suggesting the involvement of a feed-forward neural mechanism [13].

In this work we construct a modular neural network including two parts, trained in different ways: first, a hidden layer of neurons having the task of reducing the dimensionality of the data to a more suitable form, to be classified by specialized modules, sometimes called "experts", of the second part, trained with backpropagation.

Unsupervised algorithms have been shown to be very effective in the task of reducing the high dimensionality of the input data, for example PCA or ICA algorithms [15, 2]. In our work an unsupervised procedure is applied locally, preserving some topology of the original images. Our choice is motivated by biological considerations based on the idea that a network will operate better if the variance in the input representation is distributed across many input neurons, and not just to a few, as a PCA algorithm tends to do [19].

On the other hand, modularity appears to be a very effective solution to complicated tasks allowing better generalization properties, reducing the longer training times, and being also adaptive [7, 8]. Modular networks have been used successfully in several tasks such as speaker verification, face recognition, time series prediction, etc. [3, 4, 18], being also very useful tools for exploring hypothesis about brain function [5, 10].

Different systems have been constructed to deal with facial expressions, see for instance [14] and references therein, but few of them use a neural network approach. For example, in [16] a feedforward network with PCA input encoding of some facial features (eyes and mouth) is trained to classify emotions, obtaining an 84 % generalization rate; Lisetti & Rumelhart [14] have constructed a backpropagation network to classify the degree of expressive-

ness of faces. Our work continues to explore the potential of neural networks to perform this kind of task, trying to respect some biological constraints, using the capabilities of modular systems, gand reducing to a minimum the preprocessing stage.

2. The Database of Images

The Yale Face Database [1] contains 165 gray scale images in GIF format of 15 male individuals of different races and features (glasses, beard, moustache, etc.). There are 11 images per subject including different expressions, views, illumination condition, etc.

We use a subset of the database that consists of 14 subjects displaying 4 facial expressions: neutral face, happy, sad and surprised. The images were cropped to obtain input images 8 pixels width by 24 pixels height covering a portion of the face located on the left side. (See Figure 1). The images were centered taking the tip of the nose as reference and some light illumination correction was applied to a couple of images; both operations were carried out by a human observer. The resulting images were transformed to pgm 8-bit gray scale format, ready to be fed into the network after a linearly scaled transformation of pixel intensity to the interval $[0, 1]$.

Figure 1 shows a sample of the different expressions displayed by one of the subjects. In the leftmost image the area of the face cropped and used as input is shown.

Figure 1. Sample subject showing the four full face expressions (neutral, happy, sad and surprised). The white rectangle inside the rightmost figure corresponds to the area cropped and used as input for the neural network.

3. Network Structure

The network consists of a 4 layer modular neural structure composed of sigmoidal units. The input layer has 192 units corresponding to the 24x8 pixels of the area cropped from the original images. Every input neuron transmits information through a single hebbian weight, projecting to a specific neuron in the first hidden layer, selected according to a self-organized process. Thus one has at this level a new reduced representation of the images, 48 neurons, that preserves some topological aspects of the original input. The whole network architecture is shown in figure 2, where at the top we show a sample input image followed by the structure of the receptive fields corresponding to the first hidden layer neurons.

After this unsupervised compression the network splits into three modules corresponding to the expressions different from the neutral face: happy, sad and surprised. The structure of the modules could depend on the emotion they specialize in; in the case we consider here they have all the same type of architecture: one hidden layer fully connected with one output unit. There is a difference in the number of hidden neurons belonging to each modules since the recognition of happy and surprised faces is much easier than the recognition of sad ones, a fact that was previously known from experiments both with humans and computers [5]. It was necessary to put 4 hidden neurons for sad faces while 3 neurons were enough for happy and surprised ones. In this way the output of the whole network has 3 neurons that should be all OFF when a neutral face is presented, while when a face displaying an emotion is shown, the corresponding module unit should be ON.

Table 1. Generalization error rates for the modules, specialized in happy, sad and surprise faces, and for the whole net using first layer self-organizing receptive fields-hebbian (SORF-Hebbian), PCA, and random processing. The generalization error is measured after the training procedure succeeds, when the training error per example turns out to be around 0.02.

Expression Module	Error (SORF-Hebbian)	Error (PCA)	Error (Random)
Happy	0.057	0.082	0.089
Sad	0.044	0.032	0.154
Surprise	0.053	0.053	0.071
Total	0.154	0.167	0.314

3.1. Self-organization of receptive fields

As we mentioned before, the first layer of weights self-organizes according to the variance of the input pixels, with the aim of obtaining a distributed activity in the first hidden layer of neurons. Initially, the receptive fields of first hidden layer neurons are square blocks of 2×2 pixels of the input images.

The receptive fields evolve according to the following dynamics: a neuron from the first hidden layer is selected,

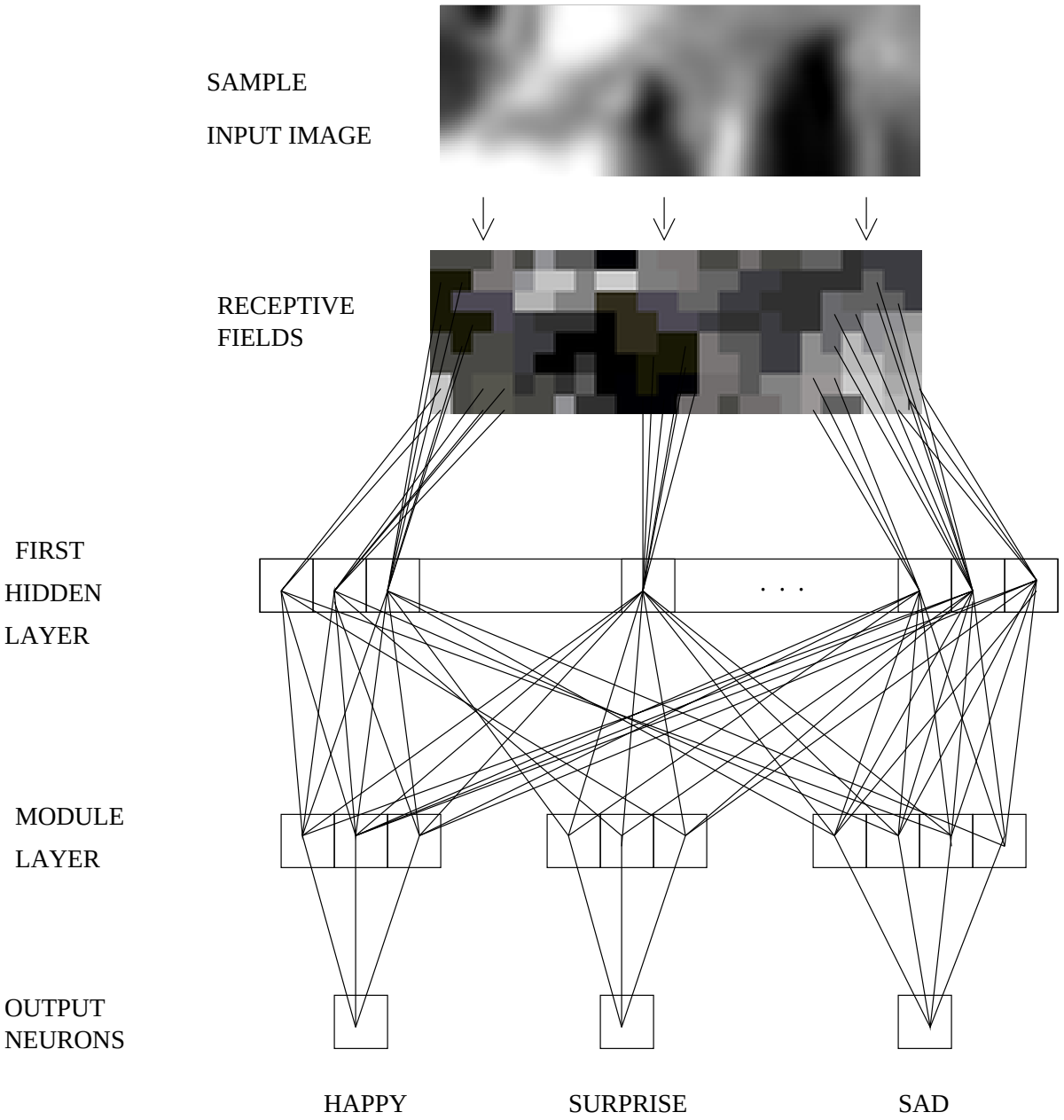

Figure 2. Schematic structure of the network architecture used to perform facial expression recognition. The network has 192 inputs corresponding to the part of the face considered and being projected via hebbian weights to 48 neurons in the first hidden layer, with self-organized receptive fields. The modules, specialized in the different expressions: happy, sad and surprise, have a one hidden layer structure with an output that should be activated when a face displaying its corresponding expression is presented at the input. At the top, one of the input sample images is shown.

then the convenience of increasing its receptive field size, by absorbing a weight from a contiguous pixel at the input level, is computed. A possible modification of the receptive field is evaluated through a simulated annealing procedure [12] involving a local energy function calculated for the two receptive fields involved, one that may increase, that we call RF^+ and the other that could decrease its size, RF^-. The change in the configuration is always accepted if the energy difference between initial and modified states is negative but could also be accepted in case the difference is positive, depending on a time decreasing probability. We define the energy of the receptive fields, $E(RF^i)$, as:

$$E(RF^i) = \left[\left(\sum_{j \in RF^i} \mathrm{V}ar_j \right) - N_i \overline{\mathrm{V}ar} \right]^2 ,$$

where $\mathrm{V}ar_j$ is the variance of a pixel belonging to the receptive field i under consideration, $\overline{\mathrm{V}ar}$ is a constant that we set to the mean variance of all the pixels, and N_i is the number of pixels forming the receptive field.

Through this procedure we obtain new reorganized receptive fields with a variable dimension that we constrain between 1 to 10 pixels. The dynamics tends to form small receptive fields containing pixels with high variance and larger receptive fields including many low variance pixels. In this way more importance is given to pixels with a higher variance but at the same time it is possible to obtain a distributed response, at the first hidden layer units, by clustering many low variance pixels in some receptive fields.

In figure 3 are shown three states corresponding to initial, intermediate and final stages of the self-organizing process. Different colour mean a different receptive field, the mean variance of the constituting pixels being indicated by the brightness: the darker tones correspond to a lower mean variance. Note that the size of the darker clusters is larger compared to the clearer ones.

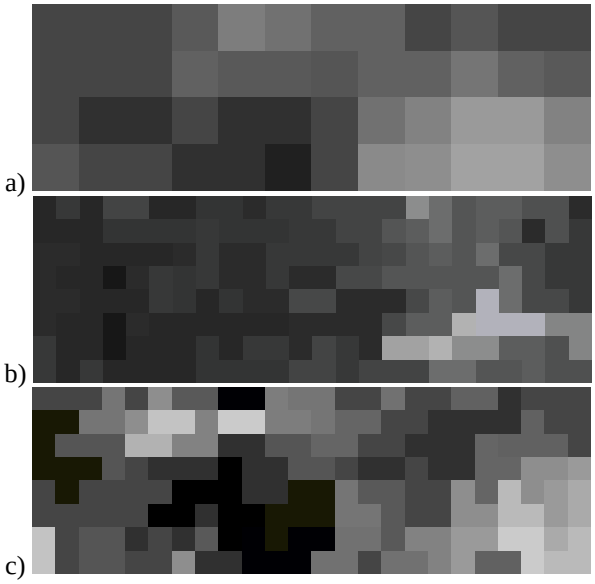

Figure 3. Receptive field evolution at different stages of the self-organizing process a) Initial state b) Intermediate state c) Final state. The mean variance of the input pixels corresponding to the receptive fields is indicated by the level of brightness, with darkest pixels being the low variance ones.

4. Training procedure

As the amount of data available for training and testing is limited (14 subjects, 56 images), we decided to use a cross-validation test, normally used in similar situations. In this procedure 13 out of the 14 available subjects are chosen to train the network and the 4 unseen faces of the remaining subject are used to test the generalization ability of the system. This procedure is repeated 14 times, one for each subject being kept out of the training set.

The first layer of 192 weights, one for each input pixel, is trained with an unsupervised hebbian algorithm.

The Hebbian rule used is Oja's rule, known to tend to a principal component analysis of the input vectors, converging to the largest eigenvector, while normalization is ensured [15, 9]. The change in the weight values can be written as,

$$\Delta w_i = \eta V (\xi_i - V w_i), \tag{1}$$

where w_i is one of the weights connecting an input neuron with value ξ_i to a first hidden layer neuron, with net input V,

$$V = \sum_{i=1}^{4} \xi_i w_i, \tag{2}$$

and η is a learning constant that was kept fixed to 0.05.

The rest of the weights, those belonging to the modules, were trained with the standard backpropagation algorithm (Hertz, Krogh & Palmer, 1993; Haykin, 1995). To prevent overtraining and to permit a better generalization capacity we monitor the training error on each input image, to stop the training on such image when this error is lower than 0.10. Since the backpropagation training is an on-line procedure, at the end of the training phase the average error per example is decreased to 0.02, approximately.

All layers of weights were trained at the same time upon the presentation of an input image.

5. Results and Discussion

We explore the generalization ability of a modular neural system to classify facial expressions. Using a mixed learning scheme composed by unsupervised-supervised training, we obtain a generalization ability on novel faces of 84.5%, compared to a 83.2% when replacing the unsupervised procedure by a principal component (PCA) one. The generalization error rates produced by the three modules specialized in different expressions are shown in table 1. We compared the results from our mixed "Self Organized Receptive Fields - Hebbian" (SORF-Hebbian) scheme to those obtained replacing the unsupervised process with PCA preprocessing, and also with a set of random weights and receptive fields in the initial configuration (see 3a). For the case of using the PCA analysis we project the input data onto the N principal component, and train the backpropagation modules with the resulting data. We experiment with different values of N, obtaining the best results with $N = 36$. The random processing case corresponds to setting the weights of the first layer to random values uniformly within the range $[0 - 0.7]$, while no learning is applied. This procedure has shown to perform better than the simple average of weights, and it is shown here just for comparison.

Self organization and Hebbian learning, in our case applied locally, confirm to be interesting procedures for compressing data in a more biological way, compared to a PCA approach.

The advantage of the modular approach is that it permits the addition of new modules to recognize different expressions that could be trained separately.

We are currently considering many possible extensions of this work, trying to implement the unsupervised processing stage through a competitive learning scheme and also to test the system with a more extensive database. It would also be desirable that the network itself should be capable of performing the identification of a face in a complex input image, permiting the use of the system in a more realistic way, for example to mount it on a robot.

6. Acknowledgments

We acknowledge partial support from the Human Frontiers Program Grant RG0110/1998-B.

References

[1] P. N. Belhumeur and D. J. Kriegman. The yale face database. *URL: http://cvc.yale.edu/projects/yalefaces/yalefaces.html*, 1997.

[2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1997.

[3] Y. Bennani. Multi-expert and hybrid connectionist approach for pattern recognition: speaker identification task. *Intl. Journal of Neural Systems*, 5(3):207–216, 1994.

[4] M. N. Dailey and G. W. Cottrell. Organization of face and object recognition in modular neural network models. *Neural Networks*, 12(7-8):053–1074, 1999.

[5] M. N. Dailey, G. W. Cottrell, and R. Adolphs. A six-unit network is all you need to discover happiness. In *Twenty-Second Annual Conference of the Cognitive Science Society*, 2000.

[6] R. Desimone. Face selective cells in the temporal cortex of monkey. *Journal of Cognitive Neuroscience*, 3:1–8, 1991.

[7] L. Franco and S. A. Cannas. Generalization properties of modular networks implementing the parity function. *IEEE Transactions in Neural Networks (In Press)*, 2001.

[8] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan/IEEE Press, 1994.

[9] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*. Addison-Wesley, 1991.

[10] R. A. Jacobs. Nature, nurture, and the development of functional specializations: a computational approach. *Psychonomic Bulletin & Review*, 4(3):299–309, 1997.

[11] N. Kanwisher, J. McDermott, and M. M. Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17:4302–4311, 1997.

[12] S. Kirpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, pages 671–680, 1983.

[13] S. R. Lehky. Fine discrimination of faces can be performed rapidly. *Journal of Cognitive Neuroscience*, 12(5):848–855, 2000.

[14] C. L. Lisetti and D. E. Rumelhart. Facial expression recognition using a neural network. In *Proceedings of the 11th International Florida Artificial Intelligence Research Society Conference (FLAIRS'98)*, Menlo Park, CA, 1998.

[15] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1(1):61–68, 1989.

[16] C. Padgett and G. W. Cottrell. A simple neural network models categorical perception of facial expressions. In *Proceedings of the Twentieth Annual Cognitive Science Conference*, Madison, WI, Mahwah: Lawrence Erlbaum, 1998.

[17] D. I. Perret, E. T. Rolls, and W. Caan. Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Reserch*, 47:329–342, 1982.

[18] V. Petridis and A. Kehagias. *Predictive Modular Neural Networks: Applications to Time Series*. Kluwer Academic Publishers, Boston, 1998.

[19] E. T. Rolls and A. Treves. *Neural Networks and Brain Function*. Oxford, 1998.