Facial Expression Recognition Using a Neural Network

Christine L. Lisetti
Department of Computer Science
Stanford University
Stanford, CA 94305 USA
lisetti@psych.stanford.edu

David E. Rumelhart
Department of Psychology
Stanford University
Stanford, CA 94305 USA
der@psych.stanford.edu

Abstract

We discuss the development of a neural network for facial expression recognition. It aims at recognizing and interpreting facial expressions in terms of signaled emotions and level of expressiveness. We use the backpropagation algorithm to train the system to differentiate between facial expressions. We show how the network generalizes to new faces and we analyze the results. In our approach, we acknowledge that facial expressions can be very subtle, and propose strategies to deal with the complexity of various levels of expressiveness. Our database includes a variety of different faces, including individuals of different gender, race, and including different features such as glasses, mustache, and beard. Even given the variety of the database, the network learns fairly succesfully to distinguish various levels of expressiveness, and generalizes on new faces as well.

Introduction

Within the field of computer vision, there has recently been an increasing interest in the field of computer vision to recognize facial expressions. Psychologists have established correlations between various affective states and facial expressions, which humans can recognize with some level of accuracy. As it is characterized as a problem of pattern recognition in human cognition, performance by computers with pattern recognition abilities could potentially perform as well than some humans for this task. Futhermore, it is expected that three to ten years from now, the price of cameras will have dropped considerably. This will make visual awareness research for artificially intelligent systems a very interesting alley for developing computer environments.

Indeed, there exists a number of applications which can benefit from automatic facial expression recognition. Face recognition in real-life environments, for example, such as airports, and banks, often involves various different viewpoint and expressions of an individual. Being able to recognize facial expressions can assist facial recognition algorithms (Yacoob, Lam, and Davis 1995). Furthermore, body language is an important part of human-human communication (Birdwhistle 1970). Developing methods for a computer to automatically recognize human expressions, could enhance the quality of human-computer interaction and enable the construction of more natural adaptive interfaces and environments (Hayes-Roth et al. 1998), (Lisetti 1998). Facial expression recognition is useful for adapting interactive feedback in a tutoring system based on the student's level of interest, or for monitoring pilots and drivers alertness state. Automatic recognition could also be used on video recording of group interactions, to trace and document changes in the expressions of the participants, or to retreive pieces of a video based upon a particular facial expression of a subject (Picard 1997). Yet another application is found in the development of psychological emotion theories by facilitating the experimental data collection of facial expressions associated with particular emotions.

A number of systems have already dealt with facial expressions using different technical approaches such as the memory-based rule system, JANUS (Kearney and McKenzie 1993), spatio-temporal templates (Essa, Darrell and Pentland 1994), image motion (Black and Yacoob 1995), among others. One of our motivations is to explore the potential that neural networks offer for this type of pattern recognition problem. An early neural network which dealt with facial expressions was the single perceptron which could classify smiles from frowns, and which was tested on one person only (Stonham 1986). Since then, other connectionist systems have been implemented to classify and recognize facial expressions with good results (Cottrell and Metcalfe 1991), (Rosenblum, Yacoob, and Davis 1994).

Our research project continues to explore the potential of neural networks to recognize facial expressions. In the present paper, we work with two expressions, namely *neutral* and *smiling*. In order to address some questions from psychology about the discreteness of facial expressions, we study how variations in expressiveness can affect the performance of the system. We discuss what approaches are most promising given that some expressions may be ambiguous, even for human recognition. We also mention the future direction of

our ongoing project.

Facial Expressions as Emotional and Communicative Signals

Facial expressions can be viewed as communicative signals (Chovil 1991), associated with syntactic displays, speaker displays, or with listener comment displays in a conversation. This approach has been used to improve human-computer interaction with speech dialog (Nagao and Takeuchi 1994).

Facial expressions can also be considered as expressions of emotion (Ekman and Friesen 1975), raising ongoing debates about their discretness and universality. One of the most documented research effort led by Ekman has permitted to identified six basic universal emotions: *fear, anger, surprise, disgust, happiness, and sadness* (Ekman and Friesen 1975). Others like Russel prefer to think that facial expressions and labels are probably associated, but that the association may vary with culture (Russel 1994).

Whether or not there exist universal facial expressions, Wierzbicka points at the difficulty to talk about emotions, and warns that what we *refer to* as basic emotions with *labels* such as "anger", may have concepts which may very well be culturally determined (Wierzbicka 1992). Studying these concerns is beyond the scope of this paper, and need to be addressed in further details when dealing with particular applications. Relevant expressions and their interpretations may indeed vary depending upon the chosen type of application (Lisetti, 1998).

In this present paper, however, we address some of the issues above by focusing on getting a neural network to be able to recognize *differences* in levels of expressiveness from two emotions: *happy* and *neutral*.

Facial Expression Interpretation Using a Backpropagation

By contrast with non-connectionist approaches which usually use geometrical face codings, connectionist approaches have typically used image-based representation of faces in the form of 2D pixel intensity array. While this model has the advantage of preserving the relationship between features and texture, it has a very high sensitivity to variations in lighting conditions, head orientation, and size of the picture (Valentin et al. 1994). These problems typically justify a large amount of work in preprocessing the images. Normalization for size and position is necessary and can be performed automatically with algorithms for locating the face in the image and rescaling it (Turk and Pentland 1991). In our particular data set, there was no need for rescaling, as all the images were consistent with each other along that dimension.

The Data Base of Images

We used images from a variety of sources. The results described below have been derived mostly from using the FERET data base of face images which included smiling faces and neutral faces. [2] The FERET database includes pictures of faces with various poses (such as full face, profile, and half profiles) for each person. These pictures are useful to build face recognition algorithms in terms of person identification from different angles.

Since we are presently exclusively interested in facial expressions, however, we built a sub-set of the FERET data base to include only two different poses per person: namely one full face with a neutral expression, and the other full face with a smile. Not every one of the pictures had the same degree of neutrality, or the same degree of "smilingness". We have designed various approaches to test this scalability among images. As stated above, one of the advantages of the FERET images was that all the images were consistent enough in terms of size and position.

Interpreting facial expressions of an individual in terms of signaled emotions requires us to work with minute changes of some features with highly expressional value. Some examples of those are found in the mouth such as the orientation of the lips (up or down), in the eyes such as the openess of the eyes, etc.

There are *three areas of the face capable of independent muscular movement*: the brow/forehead; the eyes/lids and root of the nose; and the lower face including the cheeks, mouth, most of the nose and the chin (Ekman 1975). Furthermore, not every expression is shown in the same area of the face. For example *surprise* is often shown in the upper part of the face with wrinkles in the forehead and raised eyebrows, while *smile* is mostly shown in the lower face.

The Network Architecture

Our network is designed to deal separately with the three areas of the face capable of independent muscle movement mentioned above. It is illustrated in figure 1. Each portions of the face is pre-processed by cropping the initial full-face/background images (manually at this stage) to smaller sizes, and normalized in terms of intensity by histogram.

The network includes one layer of input units, one layer of hidden units and one layer of output units. The input units are 2D pixel intensity arrays of the cropped images. The output units express the value of expressiveness of a particular expression ranging from 1 to 8, or from 1 to 6, depending upon the experiment. The hidden layer is connected to each portion of the face and to each output unit in order to simulate "holistic" face perception. Alternatively, the connections can be loosen, to model more local pattern recognition algorithm.

Finally, because, in the future, we also want to be able to refer to the recognized expression with a label such as "happy", or "angry", we provide an ordered

[2] Portions of this research in this paper use the FERET database of facial images collected under the ARPA/ARL FERET program.

input mask of binary values for each of the emotions to be recognized. These can be the 6 basic emotions, plus neutral: bit 1: fear, bit 2: angry, bit 3: surprise, bit 4: disgust, bit 5: happiness, bit 6: sadness, bit 7 neutral. These binary output values are to be turned on to '1' if the expression is identified, while the others remain set to '0'. An example of output will be in addition to the level of expressiveness from the regular output units, the left set of output units will point to the label "3" by turning on the third bit of the mask as in $[0.0, 1, 0, 0, 0, 0]$.

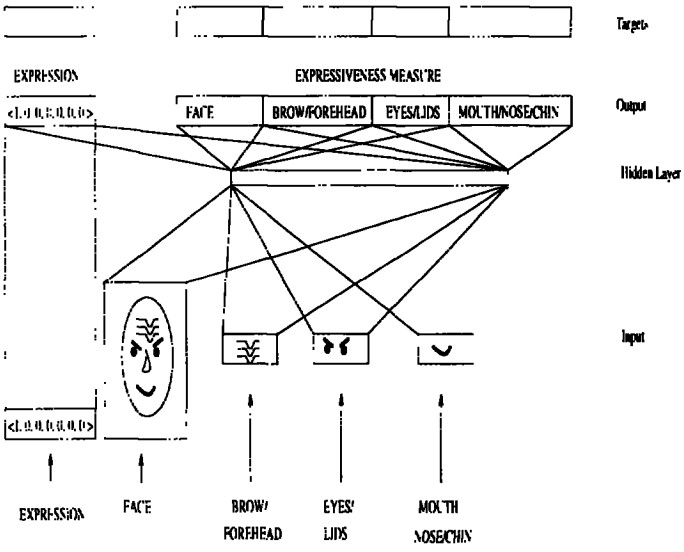

Figure 1: Network Design

In order to test each portion of the network, however, we built subnetworks each dealing with different portions of the face. We discuss here the results for the full face subnetwork, and the mouth/part-of-nose/part-of-chin subnetwork. We designed different strategies to test which approach would be better fit to recognize facial expressions in terms of degree of expressiveness: in this case, the degree of smilingness.

We reduced our images to 68X68 pixel images, for our full-face experiments, and we reduced them to 74X73 for our lower face experiments.

Full Face Network

Test 1: Graded Answers In our initial stage, we preprocessed the images manually, cropped the outer edges of the face leaving the full face as shown in figures 2 and 3. We performed histogram equalization for normalizing intensity across the different images.

Data: The network had 40 hidden units, one input unit per pixel, and one single output unit. We trained the network on 40 input images. We included images of 20 different persons, two images per person. We selected randomly 30 images in our training set and 10

Figure 2: Full Face Neutral

Figure 3: Full Face Smiling

images in our testing set for generalization. Target values ranged from 1 to 8 (similarly to psychological tests used in emotion research), to indicate the level of expressiveness of the image.

Results: The results of the training shown in figure 4 indicate that the network learned accurately. The plots compare: (1) the given target expressed in terms of the degree of expressiveness of a particular image ('x') with (2) the actual output of the network generalization and training ('o'). Generalization was then tested on new faces of persons that the network had never "seen" before . The generalization results are shown in figure 5, indicating that the network did not generalize completly, as can be observed by the differences between targets and outputs.

Lower Face Network

Happiness and satisfaction are typically shown in the third area of the face capable of independent movement. It includes the mouth, most of the nose and the chin (Ekman 1975). We therefore isolated this area of the face and generated cropped images including the lower face only. The input images were 74x73 pixel images.

We designed two procedures to test how well the network could generalize on new images that it had not been trained with: (1) testing whether the network could generalize if it had been trained on the person with one expression but not on the other, (2) testing whether the network could generalize on people which it had never been exposed to (i.e. with neither expression).

Test 2: With prior exposure to the person Data: In this case, we selected intermittently neutral faces, and smiling faces. That is, instead of training the network on both expressions for each individual, we trained the network on each individual, sometimes including both expressions but sometimes witholding either one of the two expressions. In that manner, we

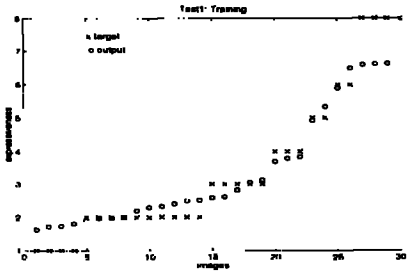

Figure 4: Full Face Processing

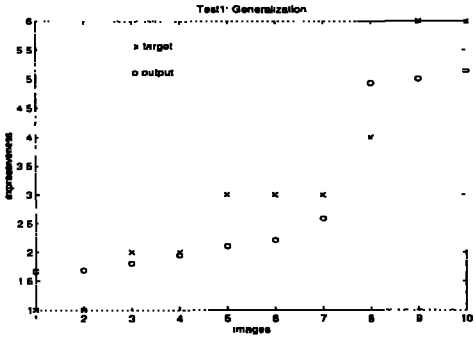

Figure 5: Full Face Processing

could test how well it generalized, without having been exposed to both expressions of the same person during the training stage.

Results: The results are shown in figures 7 and 8.

Once again the training was very successful as the network approximated its output to be very close to the values we had given it as targets. The network generalized very accurately for each of the test cases as can be observed from figure 8. We also wanted to know if the network could generalize on smiles of people that it had never been exposed to.

Test 3: Without prior exposure to the person
Data: This time, we trained the network on 116 input images of size (74X73).

Results: We included images of 58 different persons, two images per person. We selected 94 images for our training set and 22 images for our testing set for generalization. Plots in figures 9 and 10 compare: (1) the given target expressed in terms of the degree of expressiveness of a particular image ('x') with (2) the actual output of the network generalization and training ('o'). As can be observed from the graphs, the outputs match closely the given targets. A summary of the error averages for each test is given below in absolute values (see table 1). It lists the difference between target values and output values, averaged over the number of images.

Figure 6: Full Face Processing

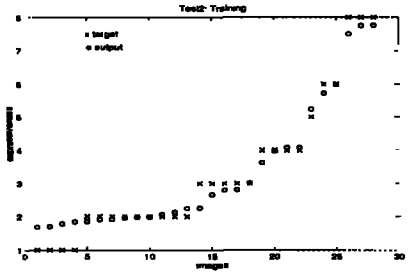

Figure 7: Lower Face Processing

Conclusion

Our results indicate that zooming in particular areas of the face for expression detection offers better results than processing the full face as a whole. Our next approach will be to isolate each area of the face, and combine their inputs into a single network to increase precision of our recognition algorithm as described earlier. One extension of the facial expression system will be the integration of the recognition scheme with a real-time tracker. This coupling is planned to enable the system to perform real-time recognition of facial expressions.

Acknowledgement

We would like to acknowledge *Intel Corporation* for partial funding for this research.

References

Birdwhistle. 1970. *Kinesics and Context: Essays on Body Motion and Communication.* University of Pennsylvania Press.

Black, & Yacoob, Y. 1995. Recognizing Faces Showing Expressions. Proceedings of the International Workshop on Automatic Face and Gesture Recognition, IEEE Press.

Black, M., and Yacoob, Y. 1995. Tracking and Recognizing Rigid and Non-Rigid Facial Motions using Local Parametric Models of Image Motion. Proceedings Int'l Conference Computer Vision ICCV'95, 374-381.

Chovil, N. 1991. Discourse-Oriented Facial Displays in Conversation. *Research on Language and Social Interaction* 25:163-194.

Cottrell, G. and Metcalfe 1991. EMPATH: Face, Emotion, and Gender Recognition using Holons. *Ad-*

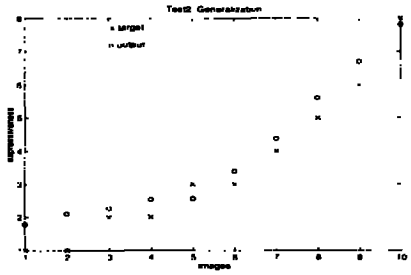

Figure 8: Lower Face Processing

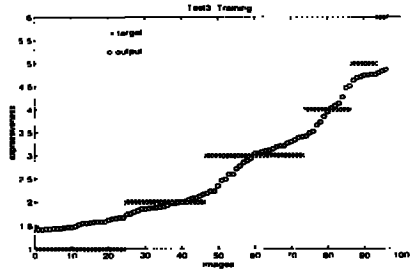

Figure 9: Lower Face without Prior Exposure

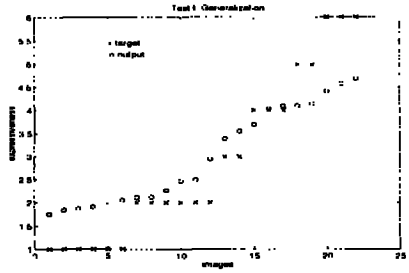

Figure 10: Lower Face Without Prior Exposure

Table 1: Error Averages (absolute values)

	TRAINING	GENERALIZATION
Test 1	0.0496	0.1072
Test 2	0.0315	0.0722
Test 3	0.0428	0.0756

vances in Neural Information Processing, Morgan Kaufmann Publishers.

Ekman, P., and Friesen, W. V. 1975 *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*, Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Essa, I.; Darrell, T.; and Pentland, A. 1994 Tracking Facial Motion, Proceedings of IEEE Workshop on Nonrigid and Articulate Motion. IEEE Computer Society Press.

Hayes-Roth, B.; Ball, G.; Lisetti, C.; Picard, R.; and Stern, A. 1998. Panel on Affect and Emotion in the User Interface. In Proceedings of the 1998 International Conference on Intelligent User Interfaces, 91-94. New York, NY: ACM Press.

Kearney, G. D. 1993. Machine Interpretation of Emotion: Design of a Memory-Based Expert System for Interpreting Facial Expressions in Terms of Signaled Emotions. *Cognitive Science* 17, 589-622.

Lisetti, C. L. 1998. An Environment to Acknowledge the Interface between Affect and Cognition. In Working Notes of the AAAI Spring Symposium on Intelligent Environment, AAAI Press.

Nagao, K., and Takeuchi, A. 1994. Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics.

Picard, R. 1997. *Affective Computing*. Cambridge, MA: MIT Press book.

Rosenblum, M.; Yacoob, Y.; and Davis, L. 1994. Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture. IEEE Workshop on Motion of Non-Rigid and Articulated Objects.

Rowley, H. A.; Baluja, S.; and Kanade, T. 1996. Neural Network-Based Face Detection. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 6:285-319.

Russel, J. 1994. Is There Universal Recognition of Emotion From Facial Expression? A Review of Cross-Cultural Studies *Psychological Bulletin* 115(1)102-141.

Stonham, T. J. 1986. Practical face recognition and verification with Wisard. In H. Ellis and M.A. Jeeves (Eds.), *Aspects of face processing*. Lancaster, England: Martinus Nijhoff.

Turk, M., and Pentland, A. 1991. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*. 3, 71-86.

Valentin, T.; Abdi, H.; O'Toole, A.; and Cottrell, G. 1994. Connectionist Models of Face Processing: A Survey. *Pattern Recognition* 27: 1290-1230.

Valentin, T. (Ed.) 1995 *Cognitive and Computational Aspects of Face Recognition: Explorations in Face Space*. New York, NY: Routledge.

Wierzbicka, A. 1992. Defining Emotion Concepts. *Cognitive Science* 16:539-581.

Yacoob, Y.; Lam, H.; and Davis, L. 1995. Recognizing Faces Showing Expressions. In Proceedings of the International Workshop on Automatic Face and Gesture Recognition, IEEE Press.