Computer Recognition of Facial Actions: A study of co-articulation effects

Evan Smith[1], Marian Stewart Bartlett[2], Javier Movellan[1,2]
[1]Cognitive Science Dept., University of California, San Diego
[2]Institute for Neural Computation, Salk Institute, La Jolla, CA

Abstract

Interpolation methods have previously been found to be effective for handling coarticulation effects in speech recognition when there is insufficient data to reliably train models for all combinations of phonemes. These interpolation models employed Hidden Markov Models (HMM's), trained on one output class at a time. Here, a neural network analog of the HMM interpolation methods is discussed and applied to the problem of analyzing facial expressions. The task was to recognize facial actions defined in the Facial Action Coding System (Ekman & Friesen, 1978). This system defines 46 component actions that comprise facial expressions, and are the analog of phonemes in facial expression. As in speech, there are thousands of "words" that the face can express (Scherer & Ekman, 1982). The network demonstrated robust recognition for the six upper facial action units, whether they occurred individually or in combination.

1. Introduction

Facial expressions contain much information beyond what is conveyed by basic emotion categories such as happy and sad. The Facial Action Coding System (FACS) (Ekman & Friesen, 1978) is a method for describing facial behavior more comprehensively by decomposing facial expressions into component actions (the phonemes of facial expressions). The system defines 46 action units (AU's), which roughly correspond to the movement of each of the individual facial muscles. Over 7000 combinations of facial actions have been observed in spontaneous facial behavior (Scherer & Ekman, 1982). FACS is the leading method for measuring facial movement in behavioral science. Currently the coding is performed manually by human experts but computer systems are being developed to recognize facial actions directly from video. Such systems could be used to develop new human-computer interaction tools, low bandwidth facial animation coding such as MPEG-4, and would make facial measurement more widely accessible as a research tool in psychology and neuroscience.

An important problem that needs to be addressed for realistic applications of such systems is robustness to co-articulation effects. Co-articulation effects refer to the fact that a facial action can look different when produced in combination with other actions.

(See Figure 1.) In this paper, we evaluate the performance of an automatic FACS recognition system with respect to this issue. In the past, the system demonstrated good recognition performance on a set of 12 individual facial actions (Donato et al. 1999; Bartlett et al. 2000), but the training and testing did not investigate robustness to co-articulation effects.

Co-articulation effects is a term borrowed from the also present in the speech recognition literature (Rabiner & Juang, 1993). Phonemes can have very different waveforms when produced in context of other phonemes. A standard solution to the co-articulation problem in speech is to extend the units of analysis to include context. For example, instead of developing phoneme models one may develop triphone models, which include previous and posterior context. In our case, we could develop models for combinations of 2 and 3 FACS units. While the number of possible combinations grows exponentially, in speech only a small percentage of such combinations appears in practice. The problem with this approach is that the amount of data available to teach combination models decreases dramatically with the number of combinations and thus the models become less reliable. This is known in the statistics literature as the bias/variance dilemma (Geman, Bienenstock, Doursat, 1992). Simple models that do not take into account context tend to be more robust, while context-dependent models tend to be more precise.

An approach used in the speech recognition community to address this problem is to combine context independent models and context dependent models. We will illustrate the approach with an example. Suppose we have 10 examples of Action unit 1 in isolation (AU1), 10 examples of AU2 in isolation and 10 examples of AU1+AU2 (a combination of AU1 and AU2). First, we create two ‘‘context independent” models, one trained with the 20 examples of AU1 (alone and in AU1+AU2) and another one trained with the 20 examples of AU2. Second, we create three ‘‘context dependent” models, each of which use 10 examples. The context independent models will in general be more robust but less precise, while the context dependent models will be more precise but less robust. Finally the two sets of models are combined by interpolation: if λ_i and λ_d are the parameters of the context independent and context dependent models for AU1, the interpolated model would have parameters $\varepsilon\lambda_i+(1-\varepsilon)(1-\lambda_d)$. The value of ε is set using a validation set to maximize generalization performance.

This technique has been successfully applied to several problems in speech recognition using HMM’s (Rabiner & Juang, 1993). Interpolation models were found to be effective when there is insufficient data to reliably train models for all possible combinations of phonemes. A similar situation occurs in facial expressions, where there is insufficient data to train all possible combinations of the 46 facial actions in defined in FACS.

Here we apply the concepts of context dependent and context independent training to neural networks. One way to perform context dependent training with a neural network is to employ one output unit for each possible combination of action units, effectively treating each combination as a separate class. Referring to the AU1+AU2 example, there would be three output units corresponding to the three possibilities: AU1 alone, AU2 alone, and AU1+AU2. There would be no hidden units, and no competition in the output

layer. To train on all observed combinations of the 46 Action Units, over 7000 output units would be required, along with an enormous amount of training data.

One way to perform context independent training with a neural network is to use one output unit per facial action, with no hidden units, and no competition in the output layer. This architecture treats each action unit as a class, regardless of other co-occurring actions. Using the AU1+AU2 example above, there would be two output units, one for each AU. Output unit number 1 would be trained to detect AU1 regardless of whether AU1 occurred alone, or in conjunction with AU2. Again, there would be no hidden units, and no competition in the output layer.

In the architecture with no hidden units, each of the n output units performs context-independent detection of the associated AU. The architecture is effectively a set of n context-independent models with no cross-communication between them. Here we employ a neural network with one output unit per facial action, as in the context independent architecture, plus we add a hidden layer. By adding hidden units, we implicitly obtain an interpolation model. The hidden units can learn AU combinations, as they received learning signals from all of the output units during training. The more hidden units, the more context dependent classes the network can learn. This is a neural network analog of the interpolation models discussed in the speech recognition community. The ε parameter that weights the contribution of the context-dependent and context-independent models is incorporated into the weights of the neural network.

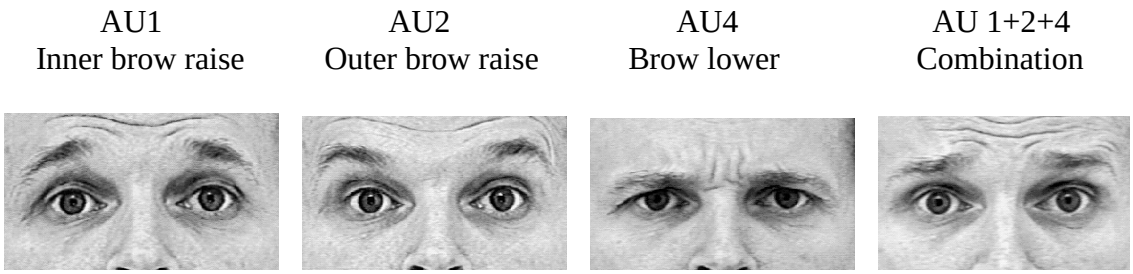

Figure 1. Three action units demonstrated individually and in combination. The combination of AU 1+2+4 occurs in fear.

2. Methods

Two image databases were employed, the Ekman-Hager database of directed facial actions (Bartlett et al., 1999), and the Cohn-Kanade database of FACS-coded posed expressions (Cohn, Kanade, & Tian, 2000). The databases consist of image sequences of subjects performing specified facial actions (Ekman-Hager) or directed facial expressions (Cohn-Kanade). Each sequence began with a neutral expression and ended with a high magnitude muscle contraction. We used 111 Ekman-Hager sequences from 20 subjects and 340 Cohn-Kanade sequences from 48 subjects. We restricted our system to an attempted to classify six upper face actions.

The face of each subject was located in the first frame in each sequence using the centers of the eyes and mouth. These coordinates were obtained manually by a mouse click. The coordinates from Frame 1 were used to register the subsequent frames in the sequence. The aspect ratios of the faces were warped so that the eye and mouth centers coincided across all images. The three coordinates were then used to rotate the eyes to horizontal, scale, and finally crop a window of 66x96 pixels containing the upper face. To control for variations in lighting, logistic thresholding and luminance scaling was performed. Difference images (δ-images) were obtained by subtracting the neutral expression in the first image of each sequence from the subsequent images in the sequence.

Entire δ-images were convolved with a bank of Gabor filters, wavelets made up of 2-D sine waves modulated by a Gaussian. Our filter banks consisted of Gabors at 8 orientations and 5 spatial scales, using the Gabor representation defined by Donato et al. (1999) (based on Lades et al., 1993). A vector representation of the Gabor filter outputs comprised the input to a neural network classifier.

The three layer neural network was made with 253440 input units, consisting of the outputs of the 40 Gabor filters at each of the 66x96 pixel locations, 15 hidden units and 6 output units one for each AU. The choice of 15 hidden units was based on pilot data that showed that performance began to drop at about 18 hidden units. Network weights were trained using back-propagation with weight decay to output a 1 in the corresponding output unit for each AU that was present in the input image. Hence the desired output for AU1+AU2 was {1 1 0 0 0 0}. There was no competition in the output layer, and output activities greater than a threshold of 0.5 were considered active. Training and testing was preformed using leave-one-out cross validation. The Cohn-Kanade data was split into 4 approximately equal sized sets selected such that data from a given subject appeared in only one set. Three of these sets were combined with the entire Ekman-Hager set and presented to the network for batch training. Following the training period the remaining Cohn-Kanade set was presented to the network and its outputs were recorded. The network weights were reinitialized and another Cohn-Kanade set was held out for testing. At the end of this process, data corresponding to generalization performance on all Cohn-Kanade data had been collected.

3. Results & Discussion

Network performance was tested for generalization to novel subjects. The performance can be evaluated in two ways. Recall that there are six output units corresponding to each of the six upper face action units. Accuracy of each output unit can be analyzed separately, providing a recognition rate for each facial action regardless of context. For example, we can measure the proportion of times Output 1 correctly indicates the presence or absence of AU1 regardless of whether AU1 occurs alone or in combination with other AU's. Table 1 gives recognition accuracies for the 6 action units. Mean recognition rate for the six action units was 98.0%. A second way to evaluate performance of the network is to measure the joint accuracy of all six output units. Here

all six outputs must be correct in order for the network output to be scored as correct. For example, if the test image is AU1+AU2, the output must be [1 1 0 0 0 0]. Table 1 shows a joint accuracy of 93.0%. Similar results were obtained by Tian et al. (2001) using a different form of input representation. Performance was also evaluated while varying the numbers of hidden units through 0 (perceptron), 3, 6, 9, 12, 15, 18 and 21. Best performance was obtained with 12 and 15 hidden units. Performance dropped slightly for 18 and 21 units.

AU Recognition Accuracy

AU1	AU2	AU4	AU5	AU6	AU7	Mean	Joint
0.9823	0.9866	0.9794	0.9806	0.9791	0.9734	0.9802	0.9298

Table 1. The first six entries indicate the probability that the network produced the correct output for the associated AU. "Mean" indicates the mean recognition accuracy across the six AU's. The final entry (Joint) indicates the probability that the network produced the entire desired output vector for all given AU combinations. (There is no AU3 in FACS).

In this paper we presented a neural network analog of the interpolation models discussed in the speech recognition community, and applied it to the problem of recognizing facial actions. The network demonstrated robust recognition for the six upper facial action units, whether they occur individually or in combination. Our results show that neural networks may provide a reasonable approach to handling coarticulation effects in automatic recognition of facial actions.

References

Bartlett, M., Donato, G., Movellan, J., Hager, J., Ekman, P., and Sejnowski, T. (2000). *Image representations for facial expression coding. In S. Solla*, T. Leen, & K. Mueller, Eds. Advances in Neural Information Processing Systems 12: 886-892.

Bartlett, M., Hager, J., Ekman, P., & Sejnowski, T. (1999). *Measuring facial expressions by computer image analysis*. Psychophysiology, 36:253-264.

Donato, G., Bartlett, M.S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying *Facial Actions*. IEEE Transaction on Pattern Analysis and Machine Intelligence, 21 (10): 976-989.

Ekman, P. & Friesen, W. V. (1978). Facial action coding system. Palo Alto: Consulting Psychologist Press.

Geman, S., Bienenstock, E., & Doursat, R. (1992). *Neural Networks and the Bias/Variance Dilemma*. Neural Computation, 4(1).

Kanade, T., Cohn, J., & Tian, Y. (2000). *Comprehensive database for facial expression analysis*. In Proceedings of International Conference on Face and Gesture Recognition. Pages 46-53.

Lades, M. and Vorbruggen, J. and Buhmann, J. and Lange, J. and Konen, W. and von der Malsburg, C. and Wurtz, R. (1993). *Distortion invariant object recognition in the dynamic link architecture*, IEEE Transactions on Computers, 42(3):300-311.

Rabiner, L.R., & Juang, B.-H. (1993). Fundamentals of speech recognition. Englewood Cliffs, NJ: Prentice-Hall.

Scherer, K., and Ekman, P. (1982). Handbook of Methods in Nonverbal Behavior Research. Cambridge University Press, Cambridge, UK.

Tian, Y., Kanade, T., & Cohn, J. (2001). *Recognizing action units for facial expression analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (2).