Fully Automatic Coding of Basic Expressions from Video.

Gwen Littlewort, Ian Fasel, Marian Stewart Bartlett, Javier R. Movellan
(INC MPLab Tech Report 2002.03)
Machine Perception Laboratory
Institute for Neural Computation
University of California, San Diego, CA 92093
gwen@inc.ucsd.edu

Abstract

We present results on a user independent fully automatic system for real time recognition of basic emotional expressions from video. The system automatically detects frontal faces in the video stream and codes them with respect to 7 dimensions: neutral, anger, disgust, fear, joy, sadness, surprise. The face finder is based on [18] with a more complex feature space and multiframe exclusion rules. The expression recognizer receives image patches located by the face detector. A Gabor representation [2] of the patch is formed and processed by bank of 63 SVMs [3]. The final coding into 7 expression categories is performed via multinomial ridge logistic regression, a natural generalization of SVMs to the multinomial case. Strategies for performing multiclass decisions using SVM's are compared. The effectiveness of Gabor magnitude filters is examined. Different methods for combining information from the upper and lower regions of the face are also discussed. Results on the Cohn-Kanade dataset of posed facial expressions are discussed [9]. The generalization performance to novel subjects on 7-way forced choice based on 614 frames was 91.5% correct. Most interestingly the outputs of the classifier change smoothly as a function of time, providing a potentially valuable representation to code facial expression dynamics in a fully automatic and unobtrusive manner.

1 Introduction

Charles Darwin was one of the first scientists to recognize that facial expression is one of the most powerful and immediate means for human beings to communicate their emotions, intentions, and opinions to each other. In addition to providing information about affective state, facial expressions also provide information about cognitive state, such as interest, boredom, confusion, and stress, and conversational signals with information about speech emphasis and syntax.

A number of ground breaking systems have appeared in the computer vision literature for facial expression recognition. These systems include measurement of facial motion through optic flow [19, 13, 15, 7] and through tracking of high-level features [17], methods for relating face images to physical models of the facial skin and musculature [13, 16, 12, 7], methods based on statistical learning of images [5, 14, 11, 2], and methods based on biologically inspired models of human vision [1]. Automated systems may have a tremendous impact on basic research by making facial expression measurement more accessible as a behavioral measure, and by providing data on the dynamics of facial behavior at a resolution that was previously unavailable. Computer systems with this capability have a wide range of applications in basic and applied research areas, including man-machine communication, security, law enforcement, psychiatry, education and telecommunications.

In this paper we present results on a user independent fully automatic system for real time recognition of basic emotional expressions from video. The system automatically detects frontal faces in the video stream and codes each frame with respect to 7 dimensions: neutral, anger, disgust, fear, joy, sadness, surprise. We analyze the effectiveness of different image representations, and methods for combining information from different regions of the face.

2 Preparing training data

2.1 Dataset

The system was trained and tested on Cohn and Kanade's DFAT-504 dataset [4]. This dataset consists of 100 university students enrolled in introductory psychology classes and ranging in age from 18 to 30 years. 65% were female, 15% were African-American, and 3% were Asian or Latino. Videos were recoded in analog S-video using a camera located directly in front of the subject. Subjects were instructed by an experimenter to perform a series of 23 facial expressions. Subjects began and ended each display from a neutral face. Before performing each display, an experimenter described and modeled the desired display. Image sequences from neutral to target display were digitized into 640 by 480 or 490 pixel arrays with 8-bit precision for grayscale values.

For our study, we selected 313 sequences from the dataset. The only selection criterion was that a sequence be labeled as one of the 6 basic emotions. The sequences came from 90 subjects, with 1 to 6 emotions per subject. The first and last frames (neutral and peak) were used as training images and for testing generalization to new subjects. The trained classifiers were later applied to the entire sequence.

2.2 Locating the faces

A fully automatic system searches for faces in each new image frame, by scanning over all locations at multiple scales. In the current system, each frame is treated independently. To deal rapidly with the huge amount of data, most of which is not part of a face, there is a hierarchical rejection of non-faces, starting with very simple criteria which are quick to calculate, and getting progressively more complex as the number of possible face candidates remaining gets smaller. The cascades are trained using adaboost. [18]

The performance for finding faces in various online face-labelled data bases was over 90%, with a false alarm rate of one per million. The performance was much better than this on the data set used for this study, because the faces were frontal, focussed and well lit, with simple background. [8]

2.3 Preprocessing

The face-finder returns the coordinates of a square box around the face. No further registration was performed. The contents of each face box was rescaled to 80x80 and symmetrically cropped to 48x48. After croping the distance between the centers of the eyes was 24 pixels on average. Figure 1 shows every other frame of a typical sequence.

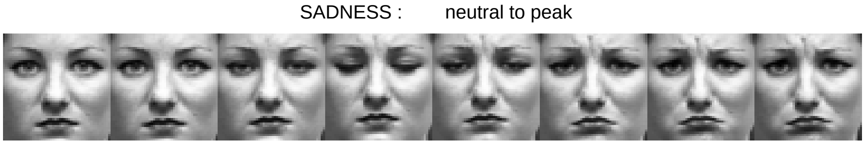

Figure 1: Typical sequence after locating, rescaling and cropping faces

These images were converted into a Gabor representation using a bank of 40 Gabor filters, one for each of 8 orientations, in increments of $\frac{\pi}{8}$, and 5 spatial frequencies, in half octave increments with wavelengths from 4 to 16 pixels. The convolutions with the image were implemented using fast Fourier transforms. The magnitudes of the complex-valued convolution provided the representation used. The outputs were normalization across the image, to unit vector length for each filter. [10], [2],[6]

3 Training with two stages of classifiers

In the training dataset, the number of subjects per emotion category is not balanced. This presents a subtle statistical problem, in that personal differences become predictive of the different expression categories. One approach to avoid this problem was to use only those subjects that have all expression categories. Unfortunately there were only 9 such subjects, resulting in a critical loss of data. We found that the most effective strategy to avoid this problem was to split the classification into two stages. The first stage classifiers were trained on every possible pair of emotions. With six emotions and neutral, there are 21 possible emotion pairs (anger-fear, joy-neutral and so on). The second stage used a multi-class classifier to yield a single reading for each emotion. The number of subjects who had both of any given pair of emotions ranged from 21 to 64. The average number of training examples per classifier was 75.

The output representation of stage 1 contained far less identity information than the input, so that the second stage training could ignore identity, and make use of all available examples. To test the reduction in identity information across stage 1, neutral faces were matched to their nearest neighbor in the emoting set. Comparing input and output representations, the percent correct fell from 98.4% to 19.8% for 90 subjects.

3.1 Stage 1 : Pairwise Emotion Classifiers

Support Vector Machines [3] were used for the pairwise classifiers in stage 1.

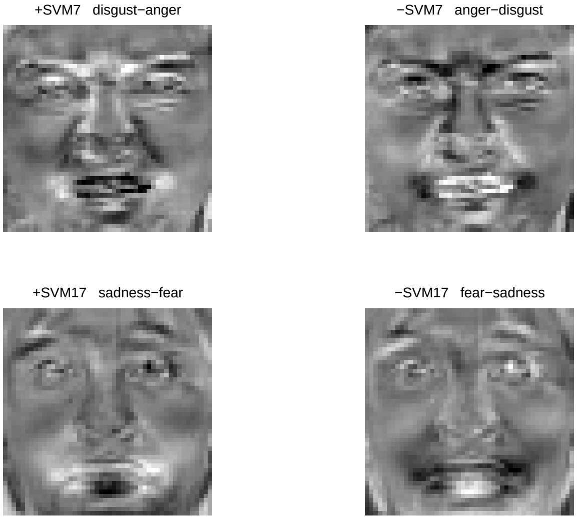

Figure 2: Receptive fields for two SVMs. Right column is simply the negative of left.

SVMs are suitable for this task because the high dimensionality of the Gabor representation does not affect the training time for kernel classifiers and because the number of training examples was not large. Each SVM is trained to distinguish 2 emotions. Training images came only from subjects who displayed both of these emotions. Only the highest intensity frame from each sequence was used. Figure 2 shows examples of the receptive fields learned by these SVMs when pixel-based representations are used.

3.2 Stage 2 classifiers

The goal of the second stage is to convert the representation produced by the first stage into a probability distribution over 7 expression categories. To this effect, we have implemented and evaluated several approaches: nearest neighbor, a simple voting scheme and MLR (multinomial logistic ridge regression).

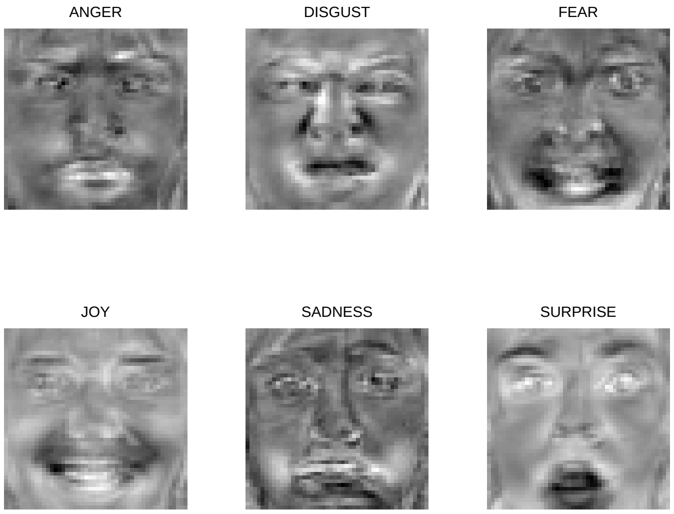

Figure 3: Receptive fields for each emotion.

K-Nearest neighbor proved unsuccessful, scoring 20 to 30% below other classifiers.

In the voting scheme, each SVM output contributes to the two relevant emotions, with one positive and one negative vote. The voting matrix applied to the thresholded outputs of Stage 1 is shown on the left of Figure 4.

MLR is a maximum likelihood approach, which closely resembles support vector machines when the number of classes is two, but which generalizes naturally to multiple classes. MLR is equivalent to a single layer perceptron with weight decay and with SoftMax competition between the outputs. The regression was implemented using the Newton-Raphson method and a ridge term coefficient of 0.001.

Figure 3 shows the combination of MLR weights with SVM receptive fields as in Figure 2.

4 Results

Classification performance was evaluated in terms of generalization to new subjects, using a leave-one-out paradigm.

4.1 Preprocessing and pairwise classifiers

We evaluated the performance of classifiers with and without Gabor preprocessing. Linear, polynomial, Laplacian and Gaussian SVM kernels were tried. Linear and unit-width Gaussian kernel functions worked the best. We evaluated training initial classifiers on whole faces, upper half faces or lower half faces. To compare the performances for these different cases, voting was used for the second stage. The percentage correct for novel subjects is shown, for various conditions, in Table 1.

Face Region	Gabor Lin	Gabor RBF	Pixel Lin	Pixel RBF
Upper	75.2	76.9		
Lower	80.9	83.2		
Whole	85.2	86.2	73.1	73.3
Lower+Upper	82.9	83.1	70.2	
Whole+Up+Lo	86.3	86.3	74.9	76.2

Table 1: Performance for Voting on SVMs

The best performance came from a combination of linear SVMs on upper Gabors, lower Gabors and whole face, which performed at 87.5%

An interesting comparison is early integration (training the SVMs on the whole face) versus late integration (training separate SVMs for upper and lower face, and then combining the outputs). Psychophysical experiments show that in humans the perception of facial expressions is well described by a late integration model in which the lower and upper region are independently analyzed [20]. Our results were consistent with this model in that we found that processing the upper and lower regions of the face independently results on loss of only a few percent points.

Table 1 shows that using Gabor representation adds more than ten percentage points to the performance, so it is a crucial part of the system. Non-linear kernel functions provide a small improvement over linear SVMs in some cases, however, linear SVMs have a speed advantage in real time applications.

4.2 Comparing Stage 2 classifiers

MLR is typically 4 percentage points better than voting. The best performance sofar was based on combining 3 sets of pairwise classifiers. Training Gaussian SVMs on upper face Gabors, lower face Gabors and on whole face images, followed by MLR, the performance rose from 87.9% (voting result) to 91.5%.

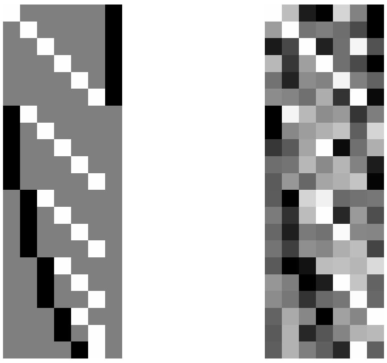

Figure 4: The Stage 2 voting weights are shown on the left and the optimal MLR weights are shown on the right. The 7 columns are emotions (Ang,Dis,Fear,Joy,Sad,Surp,Neut) and the 21 rows are the pairwise SVMs.

The weight matrix learned by MLR is compared with the voting matrix, for the case of linear SVMs on whole face Gabors, as shown in Figure 4. The standard deviation of these weights across 90 subjects was less than 5% of the weight value. At first glance, the learned weights look like a noisy version of the voting weights. For example, the first 6 rows represent the emotion:neutral SVMs, and the white diagonal that connects each pair to its emotion, and the black bar in the neutral column on the right, are clearly visible in both matrices. However, the learned weights have more cross-talk, that is, pairwise classifiers can influence the estimated level of a third emotion. For example, the 3rd row (fear:neutral) has positive links to fear and negative to neutral as expected, but it also has negative links to joy and anger and positive to disgust. These links are not reciprocated. Fear gets a negative link from anger:neutral, but not from joy:neutral or disgust:neutral. Furthermore the anger:fear SVM seems not to vote for fear. There is clearly some complex interference between the representations of these emotions.

5 Sequence Processing

Although each image is separately processed and classified, it is possible to string together the outputs from the frames of the original video sequences to obtain graphs of the time evolution of the expression as shown in Figure 5.

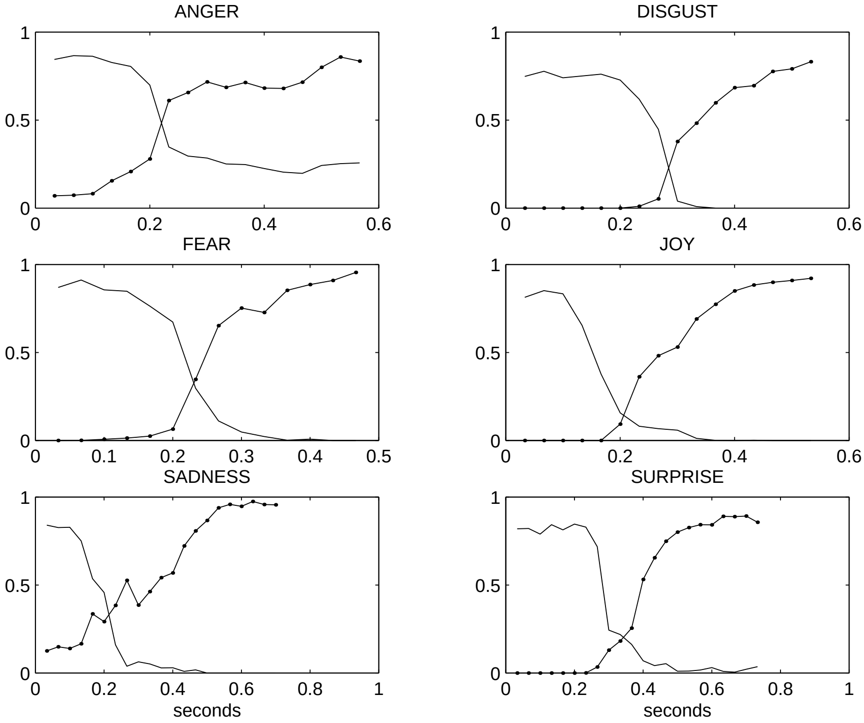

Figure 5: The graphs show the neutral output decreasing and the output for the relevant emotion increasing as a function of time, for 6 sequences from subject 32.

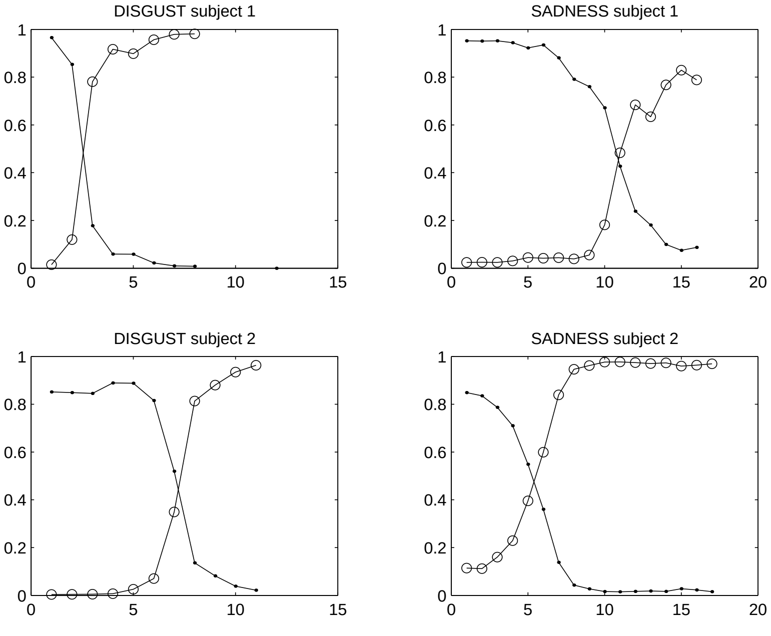

Figure 6: Emotion and neutral levels in disgust and sadness sequences for 2 subjects.

6 The Emotion Mirror

The goal of emotion mirroring is to render a character in real time that mimics the emotional expression of a person. The goal is not necessarily to mimic each facial muscle movement but the overall expression of the face. Here we described the approach we are experimenting with, and that will be submitted for a real time demonstration in NIPS2002. Figure 7 shows the prototype system at work. The images of the person on the left side are a natural recording. The images of the person on the right side are a reconstruction based on the emotion codes of the first subject, but expressed by the second subject. Given two subjects who have posed sequences for the same set of emotions, we match frames from the two subjects by the shortest distance between their 7-D emotion codes, that is, the output of each of the 7 emotion classifiers for each frame. Note that head movement and blinking are not coded and thus mirrored.

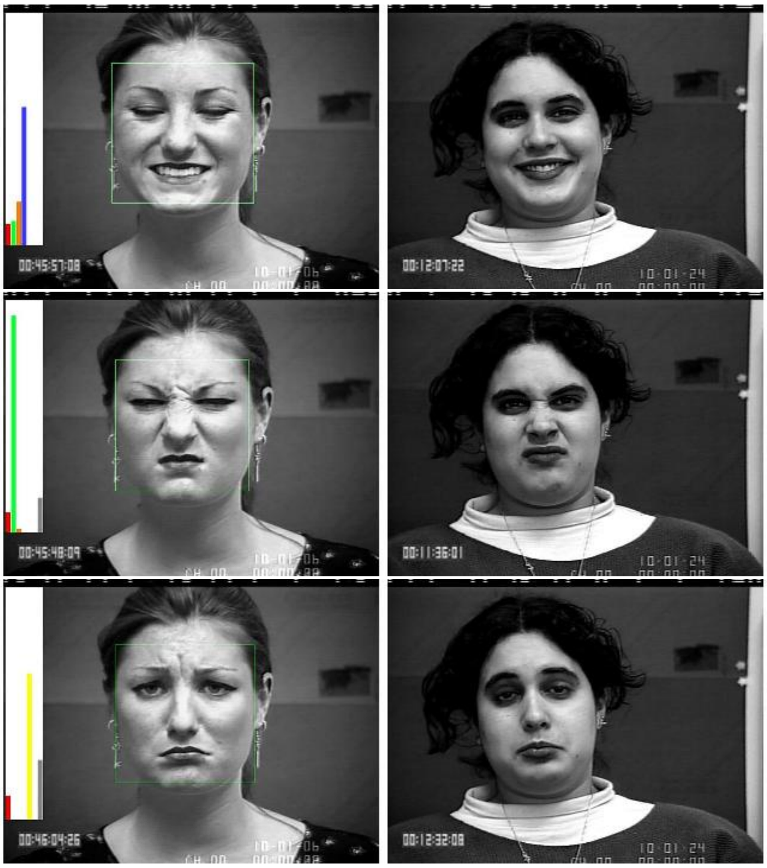

Figure 7: Examples of the emotion mirror. Bar code on left measures all 7 expressions

7 Conclusions

Our results shows that user independent fully automatic real time coding of basic expressions is an achievable goal with present computer power, at least for applications in which frontal views can be assumed. The problem of classification into 6 basic expressions can be solved with high accuracy by a simple linear system, after the images are preprocessed by a bank of Gabor filters. These results are consistent with those reported by on a smaller dataset [14]. We showed a two-stage method to train a system efficiently using unbalanced databases, in which the same subjects do not show examples of each expression category.

It is interesting to note that good performance results are obtained when directly processing the output of an automatic face detector without the need for explicit detection and registration of facial features.

References

[1] Marian S. Bartlett. *Face Image Analysis by Unsupervised Learning*, volume 612 of *The Kluwer International Series on Engineering and Computer Science.* Kluwer Academic Publishers, Boston, 2001.

[2] M.S. Bartlett, G.L. Donato, J.R. Movellan, J.C. Hager, P. Ekman, and T.J. Sejnowski. Image representations for facial expression coding. In S.A. Solla, T.K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[3] B.E. Boser., I.M. Guyon, and V.N.Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

[4] J.F. Cohn, A.J. Zlochower, J.J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, 36:35–43, 1999.

[5] G. Cottrell and J. Metcalfe. Face, gender and emotion recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, San Mateo, CA, 1991. Morgan Kaufmann.

[6] GW Cottrell, MN Dailey, C Padgett, and Adolphs R. *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, chapter Is all face processing holistic? The view from UCSD. Erlbaum, 2000.

[7] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–63, 1997.

[8] I. Fasel and J. R. Movellan. Comparison of neurally inspired face detection algorithms. In *Proceedings of the international conference on artificial neural networks (ICANN 2002).* UAM, 2002.

[9] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of the fourth IEEE International conference on automatic face and gesture recognition (FG'00)*, pages 46–53, Grenoble, France, 2000.

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

[11] A. Lanitis, C. Taylor, and T. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[12] H. Li, P. Roivainen, and R. Forchheimer. 3-d motion estimation in model-based facial image coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):545–555, 1993.

[13] K. Mase. Recognition of facial expression from optical flow. *IEICE Transactions E*, 74(10):3474–3483, 1991.

[14] C. Padgett and G. Cottrell. Representing face images for emotion classification. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, Cambridge, MA, 1997. MIT Press.

[15] M. Rosenblum, Y. Yacoob, and L. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138, 1996.

[16] D. Terzopoulus and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[17] Y.L. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:97–116, 2001.

[18] Paul Viola and Michael Jones. Robust real-time object detection. Technical Report CRL 20001/01, Cambridge ResearchLaboratory, 2001.

[19] Y. Yacoob and L. Davis. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):636–642, 1996.

[20] Ellison, J.W., Massaro, D.W. (1997). ”Featural evaluation,integration, and judgement of facial affect,” Journal of Experimental Psychology: Human Perception and Performance, 23(1), 213-226.