Fully Automatic Facial Action Recognition in Spontaneous Behavior

Marian Stewart Bartlett[1], Gwen Littlewort[1], Mark Frank[2], Claudia Lainscsek[1],
Ian Fasel[1], Javier Movellan[1]

[1]Institute for Neural Computation, University of California, San Diego
[2] SUNY Buffalo, New York
mbartlett@ucsd.edu

Abstract

We present results on a user independent fully automatic system for real time recognition of facial actions from the Facial Action Coding System (FACS). The system automatically detects frontal faces in the video stream and codes each frame with respect to 20 Action units. We present preliminary results on a task of facial action detection in spontaneous expressions during discourse. Support vector machines and AdaBoost classifiers are compared. For both classifiers, the output margin predicts action unit intensity.

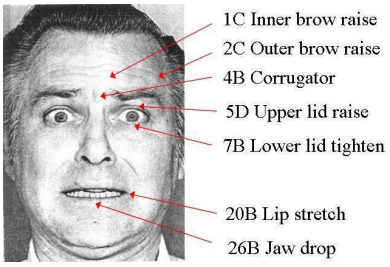

Figure 1. Example FACS codes for a prototypical expression of fear. Spontaneous expressions may contain only a subset of these Action Units.

1 Introduction

In order to objectively capture the richness and complexity of facial expressions, behavioral scientists have found it necessary to develop objective coding standards. The facial action coding system (FACS) [2] is the most objective and comprehensive coding system in the behavioral sciences. A human coder decomposes facial expressions in terms of 46 component movements, which roughly correspond to the individual facial muscles. An example is shown in Figure 1. Several research groups have recognized the importance of automatically recognizing FACS [1, 9, 8, 5]. Here we describe progress on a system for fully automated facial action coding.

We present results on a user independent fully automatic system for real time recognition of facial actions from the Facial Action Coding System (FACS). The system automatically detects frontal faces in the video stream and codes each frame with respect to 20 Action units. In previous work, we conducted empirical investigations of machine learning methods applied to the related problem of classifying expressions of basic emotions [6]. We compared AdaBoost, support vector machines, and linear discriminant analysis, as well as feature selection methods techniques. Best results were obtained by selecting a subset of Gabor filters using AdaBoost and then training Support Vector Machines on the outputs of the filters selected by AdaBoost. The combination of AdaBoost and SVM's enhanced both speed and accuracy of the system. An overview of the system is shown in Figure 2. Here we apply this system to the problem of detecting facial actions in spontaneous expressions. The system presented here detects 20 action units, is fully automatic, and operates in real-time.

2 Automated System

2.1 Real-time Face Detection

We developed a real-time face detection system that employs boosting techniques in a generative framework [3] and extends work by [10]. Enhancements to [10] include employing Gentleboost instead of AdaBoost, smart feature search, and a novel cascade training procedure, combined in a generative framework. Source code for the face detector is freely available at http://kolmogorov.sourceforge.net. Accuracy on the CMU-MIT dataset, a standard public data set for benchmarking frontal face detection systems, is 90% detections and 1/million false alarms, which is state-of-the-art accuracy. The CMU test set has unconstrained lighting and background. With controlled lighting and background, such as the facial expression data employed here, detection

Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06)
0-7695-2503-2/06 $20.00 © 2006 IEEE

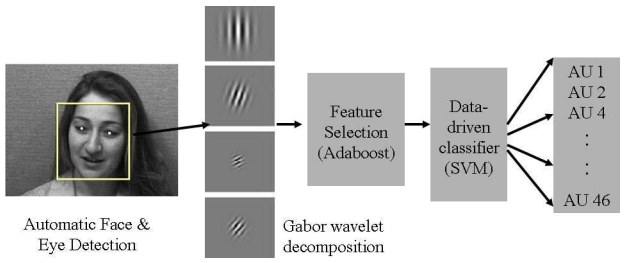

Figure 2. Overview of fully automated facial action coding system.

accuracy is much higher. The system presently operates at 24 frames/second on a 3 GHz Pentium IV for 320x240 images.

All faces in the training datasets were successfully detected. The automatically located faces were rescaled to 96x96 pixels. The typical distance between the centers of the eyes was roughly 48 pixels. Automatic eye detection [3] was employed to align the eyes in each image. The images were then passed through a bank of Gabor filters 8 orientations and 9 spatial frequencies (2:32 pixels per cycle at 1/2 octave steps) (See [6]). Output magnitudes were then passed to the classifiers. No feature selection was performed for the results presented here, although it is ongoing work that will be presented in another paper.

2.2 Facial Action Classification

Facial action classification was assessed for two classifiers: Support vector machines (SVM's) and AdaBoost.
SVM's. SVM's are well suited to this task because the high dimensionality of the Gabor representation O(10^5) does not affect training time, which depends only on the number of training examples O(10^2). In our previous work, linear, polynomial, and radial basis function (RBF) kernels with Laplacian, and Gaussian basis functions were explored[6]. Linear and RBF kernels employing a unit-width Gaussian performed best on that task. Linear SVMs are evaluated here on the task of facial action recognition.

AdaBoost. The features employed for the AdaBoost AU classifier were the individual Gabor filters. This gave 9x8x48x48= 165,888 possible features. A subset of these features was chosen using AdaBoost. On each training round, the Gabor feature with the best expression classification performance for the current boosting distribution was chosen. The performance measure was a weighted sum of errors on a binary classification task, where the weighting distribution (boosting) was updated at every step to reflect how well each training vector was classified. AdaBoost training continued until 200 features were selected per action unit classifier. The union of all features selected for each of the 20 action unit detectors resulted in a total of 4000 features.

3 Facial expression data

3.1 The RU-FACS Spontaneous Expression Database

Our collaborators at Rutgers University have collected a dataset of spontaneous facial behavior with rigorous FACS coding. The dataset consists of 100 subjects participating in a 'false opinion' paradigm. In this paradigm, subjects first fill out a questionnaire regarding their opinions about a social or political issue. Subjects are then asked to either tell the truth or take the opposite opinion on an issue where they rated strong feelings, and convince an interviewer they are telling the truth. This paradigm has been shown to elicit a wide range of emotional expressions as well as speech-related facial expressions. This dataset is particularly challenging both because of speech-related mouth movements, and also because of out-of-plane head rotations which tend to be present during discourse.

Two minutes of each subject's behavior is being FACS coded by two certified FACS coders. FACS codes include the apex frame as well as the onset and offset frame for each action unit (AU). Here we present preliminary results for a system trained on two large datasets of FACS-coded posed expressions, and tested on the spontaneous expression database.

3.2 Posed expression databases

The system was trained on FACS-coded images from 2 datasets. The first dataset was Cohn and Kanade's DFAT-504 dataset [4]. This dataset consists of 100 university students 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 displays. Subjects began each display with 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 pixel arrays with 8-bit precision for grayscale values. The facial expressions in this dataset were FACS coded by two certified FACS coders.

The second dataset consisted of directed facial actions from 24 subjects collected by Hager and Ekman. (See [1].) Subjects were instructed by a FACS expert on the display of individual facial actions and action combinations, and they practiced with a mirror. The resulting video was verified for AU content by two certified FACS coders.

4 Training

The combined dataset contained 2568 training examples from 119 subjects. Separate binary classifiers, one for each

AU, were trained to detect the presence of the AU regardless of the co-occurring AU's. We refer to this as context-independent recognition. Positive examples consisted of the last frame of each sequence which contained the expression apex. Negative examples consisted of all apex frames that did not contain the target AU plus neutral images obtained from the first frame of each sequence, for a total of 2568-N negative examples for each AU.

5 Results

5.1 Generalization Performance *Within* Dataset

We first report performance for generalization to novel subjects *within* the Cohn-Kanade and Ekman-Hager databases. Generalization to new subjects was tested using leave-one-subject-out cross-validation in which all images of the test subject were excluded from training. Results for the AdaBoost classifier are shown in Table 1. System outputs were the output of the AdaBoost discriminant function for each AU. All system outputs above threshold were treated as detections.

The system obtained a mean of 91% agreement with human FACS labels. Overall percent correct can be an unreliable measure of performance, however, since it depends on the proportion of targets to nontargets, and also on the decision threshold. A more reliable performance measure is area under the ROC (receiver-operator characteristic curve.) This curve is obtained by plotting hit rate (true positives) against false alarm rate (false positives) as the decision threshold varies. The area under this curve is denoted A′. A′ is equivalent to percent correct in a 2-alternative forced choice task, in which the system must choose which of two options contains the target on each trial. Mean A′ for the posed expressions was 92.6. Inspection of the ROC curves in Figure 3 shows the dependence of system performance on the number of training examples.

System outputs for full image sequences of test subjects are shown in Figure 4. Although each individual image is separately processed and classified, the outputs change smoothly as a function of expression magnitude in the successive frames of each sequence, enabling applications for measuring the magnitude and dynamics of facial expressions.

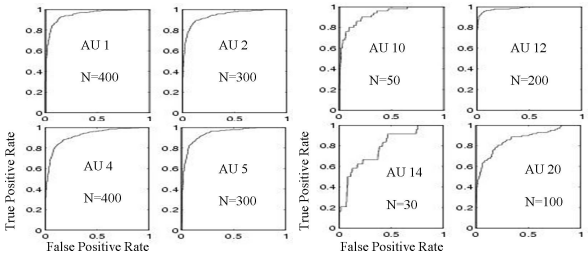

Figure 3. ROC curves for 8 AU detectors, tested on posed expressions.

Table 1. Performance for posed expressions. Shown is fully automatic recognition of 20 facial actions, generalization to novel subjects in the Cohn-Kanade and Ekman-Hager databases. N: Total number of positive examples. P: Percent agreement with Human FACS codes (positive and negative examples classed correctly). Hit, FA: Hit and false alarm rates. A′: Area under the ROC. The classifier was AdaBoost.

AU	Name	N	P	Hit	FA	A′
1	Inn. brow raise	409	92	86	7	95
2	Out. brow raise	315	88	85	12	92
4	Brow lower	412	89	76	9	91
5	Upper lid raise	286	92	88	7	96
6	Cheek raise	278	93	86	6	96
7	Lower lid tight	403	88	89	12	95
9	Nose wrinkle	68	100	88	0	100
10	Lip Raise	50	97	29	2	90
11	Nasolabial	39	94	33	4	74
12	Lip crnr. pull	196	95	93	5	98
14	Dimpler	32	99	20	0	85
15	Lip crnr. depr.	100	85	85	14	91
16	Lower Lip depr.	47	98	29	1	92
17	Chin raise	203	89	86	10	93
20	Lip stretch	99	92	57	6	84
23	Lip tighten	57	91	42	8	70
24	Lip press	49	92	64	7	88
25	Lips part	376	89	83	9	93
26	Jaw drop	86	93	58	5	85
27	Mouth stretch	81	99	100	1	100
	Mean		90.9	80.1	8.2	92.6

For many applications of automatic facial expression analysis, image compression is desirable in order to make an inexpensive, flexible system. The image analysis methods employed in this system, such as Gabor filters, may not be as affected by lossy compression as other image analysis methods such as optic flow. We therefore investigated the relationship between AU recognition performance and image compression. Detectors for three action units (AU 1, AU2, and AU4) were compared when tested at five levels of compression: No loss (original bmp images), and 4 levels of jpeg compression quality: 100%, 75in Figure 5. Performance remained consistent across substantial quantities of lossy compression. This finding is of practical importance for system design.

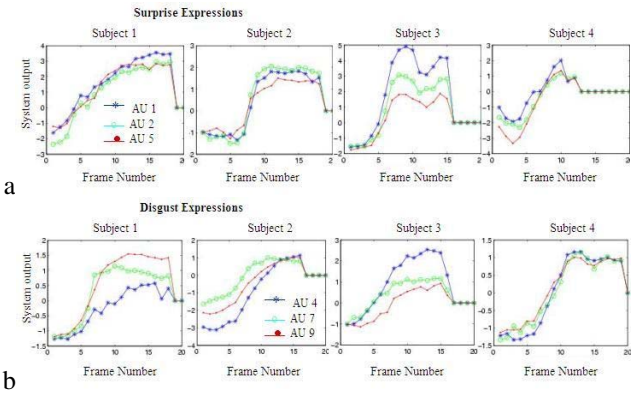

Figure 4. Automated FACS measurements for full image sequences. a. Surprise expression sequences from 4 subjects containing AU's 1,2 and 5. b. Disgust expression sequences from 4 subjects containing AU's 4,7 and 9.

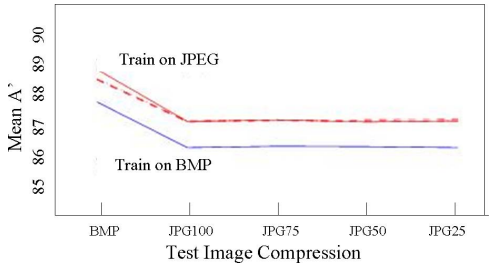

Figure 5. Effects of compression on AU recognition performance.

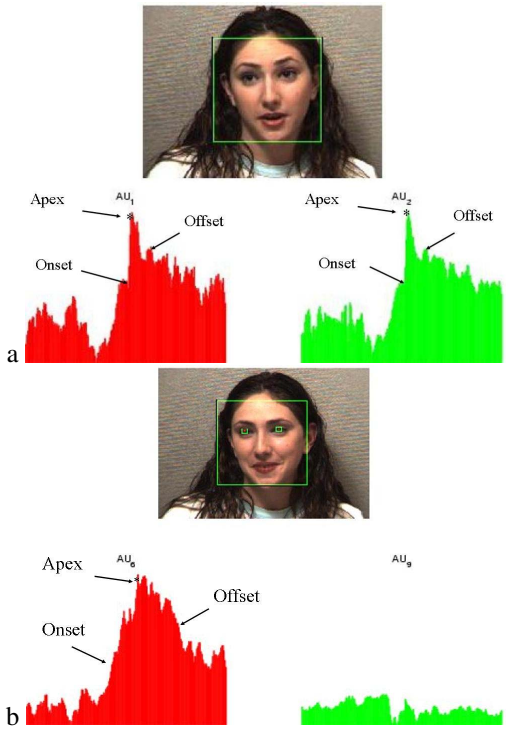

Figure 6. Sample system outputs for a 10-second segment containing a brow-raise (FACS code 1+2). System output is shown for AU 1 (left) and AU 2 (right). Human codes are overlayed for comparison (onset, apex, offset).

5.2 Generalization to Spontaneous Expressions

The system described in Section 4 was then tested on the spontaneous expression database. Hence this test included generalization to a new database, as well as handling speech and head movement, with both in-plane and out-of-plane rotations. Preliminary results are presented for 12 subjects. This data contained a total of 1689 labeled events, consisting of 33 distinct action units, 19 of which were AU's for which we had trained classifiers. Face detections were accepted if the face box was greater than 150 pixels width, both eyes were detected with positive position, and the distance between the eyes was > 40 pixels. This resulted in faces found for 95% of the video frames. Most non-detects occurred when there was head rotations beyond± 10 deg or partial occlusion. All detected faces were passed to the AU recognition system.

Here we present benchmark performance of the basic frame-by-frame system on the video data. Figure 6 shows sample system outputs for one subject, and performance is shown in Table 2. Performance was assessed several ways. First, we assessed overall percent correct for each action unit on a frame-by-frame basis, where system outputs that were above threshold inside the onset and offset interval indicated by the human FACS codes, and below threshold outside that interval were considered correct. This gave an overall accuracy of 93% correct across AU's for the AdaBoost classifier. Mean area under the ROC was .71.

Next an interval analysis was performed, which is intended to serve as a baseline for future analysis of output dynamics. The interval analysis measured detections on intervals of length I. Here we present performance for intervals of length 21 (10 on either side of the apex), but performance was stable for a range of choices of I. A target AU was treated as present if at least 6/21 frames were above threshold, where the threshold was set to 1 standard deviation above the mean. Negative examples consisted of the remaining 2 minute video stream for each subject, outside the FACS coded onset and offset intervals for the target AU, parsed into intervals of 21 frames. This simple interval analysis raised the area under the ROC to .75.

Table 2. Recognition of spontaneous facial actions. The classifier was AdaBoost. AU: Action unit number. N: Total number of testing examples. P: Percent correct over all frames. Hit, FA: Hit and false alarm rates. A': Area under the ROC. A'_{Δ}: Area under the ROC for interval analysis (see text). The classifier was AdaBoost.

AU	N	P	Hit	FA	A'	A'_{Δ}
1	169	87	35	9	78	83
2	153	84	29	13	62	68
4	32	97	15	2	74	84
5	36	97	7	1	71	76
6	50	92	32	4	90	92
7	46	91	12	7	64	66
9	2	99	0	0	88	93
10	38	95	0	0	62	65
11	3	99	0	0	73	83
12	119	86	45	7	86	88
14	87	94	0	0	70	77
15	77	94	23	4	69	73
16	5	99	0	0	63	57
17	121	93	15	2	74	76
20	12	99	0	0	66	69
23	24	98	0	0	69	75
24	68	95	7	3	64	63
25	200	54	68	50	70	73
26	144	91	2	1	63	64
Mean		93	15	5	71	75

5.2.1 AdaBoost v. SVM performance

Table 3 compares AU recognition performance with AdaBoost to a linear SVM. In previous work with posed expressions of basic emotions, AdaBoost performed similarly to SVM's, conferring a marginal advantage over the linear SVM [6]. Here we support this finding for recognition of action units in spontaneous expressions. AdaBoost had a small advantage over the linear SVM which was statistically significant on a paired t-test (t(7)=3.1, p=.018). A substantial performance increase was incurred for the SVM by employing an interval analysis. The SVM output y was converted to z-scores for each subject $z = (y - \mu/\sigma)$, and then z was integrated over a window of 11 frames. The temporal information in the classifier outputs contain considerable information that we intend to exploit in future work.

5.2.2 The margin predicts AU intensity

Figure 7 shows a sample of system outputs for a 2 minute 20 second continuous video stream. The output margin, meaning the distance to the separating hyperplane, contained information about action unit intensity. Correlations were computed between the margin of the linear SVM and the AU intensity as coded by the human coders for each subject for the 8 AU's shown in Table 3. Mean correlation across subjects was r=0.26. The AdaBoost margin, which is the output of the AdaBoost discriminant function, was also associated with action unit intensity. Mean correlation of the AdaBoost margin with human-coded intensity was 0.30. The difference between SVM's and AdaBoost was not statistically significant on a paired t-test. These correlations are not large, but there is indeed a signal on this challenging dataset. The system that was trained on posed facial expressions under highly controlled conditions was able to obtain relevant information when applied to real behavior, including the presence of an AU and the intensity of that AU.

Table 3. Comparison of AdaBoost to linear SVM's for AU classification in the spontaneous expression database. A'_{Δ}: Area under the ROC for interval analysis (see text).

		AdaBoost		SVM	
AU	N	A'	A'_{Δ}	A'	A'_{Δ}
1	169	78	83	73	83
2	153	62	68	63	76
4	32	74	84	74	86
5	36	71	76	63	73
10	38	62	65	60	71
12	119	86	88	84	90
14	87	70	77	65	73
20	12	66	69	60	74
	Mean	71.1	76.3	67.8	78.3

In order to examine the relation between the margin and the AU intensity in conditions with less noise, the correlation analysis was repeated for the posed data. The system was retrained on the even-numbered subjects of the Cohn-Kanade and Ekman-Hager datasets, and then tested on the odd-numbered subjects of the Ekman-Hager set. The 8 AU's shown in Table 3 were tested, and the correlation between the margin of the linear SVM and the AU intensity was computed for each test subject, and then collapsed across subjects. Mean correlation between the margin of the linear SVM and the AU intensity was 0.63 for the posed expression data.

6 Conclusions

Our results suggest that user independent, fully automatic real time coding of facial actions in the continuous video stream is an achievable goal with present computer power. The full system operates in real time. Face detection runs at 24 frames/second in 320x240 images on a 3 GHz Pentium IV. The AU recognition step operates in less than 10 msec.

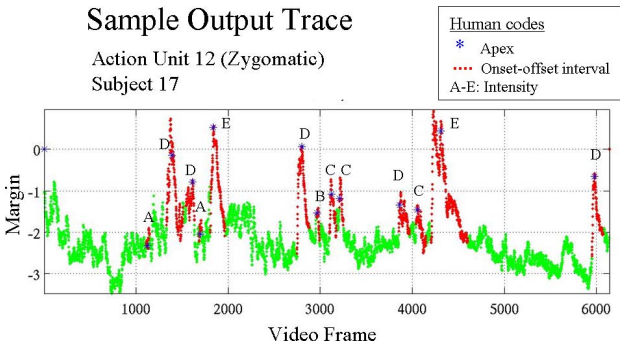

Figure 7. Output trajectory for a 2 minute 20 sec. video (6000 frames), for one subject and one action unit. Shown is the margin (the distance to the separating hyperplane). The human FACS labels are overlaid for comparison. Blue stars indicate the frame at which the AU apex was coded. The frames within the onset and offset of the AU are shown in red. Letters A-E indicate AU intensity, with E highest.

Here we presented preliminary results for the performance of the system on spontaneous expressions. The system was able to detect facial actions in this database despite the presence of speech, out-of-plane head movements that occur during discourse, and the fact that many of the action units occurred in combination. Moreover, the output margin predicted AU intensity. These results provide a benchmark for future work on spontaneous expression video.

The output sequence for both the AdaBoost and SVM classifiers contains information about dynamics that can be exploited for deciding the presence of a facial action. Future work will explore these dynamics, and compare improvement to the benchmark provided here. This system has the potential to provide information about expression dynamics that was previously intractable by hand coding. The accuracy of automated facial expression measurement may also be considerably improved by 3D alignment of faces. Moreover, information about head movement dynamics is an important component of nonverbal behavior, and is measured in FACS. Members of this group have developed techniques for automatically estimating 3D head pose in a generative model and for aligning face images in 3D [7].

The system presented here is fully automated, and performance rates for posed expressions compare favorably with other systems tested on the Cohn-Kanade dataset that employed varying levels of manual registration. The approach to automatic FACS coding presented here, in addition to being fully automated, also differs from approaches such as [8] and [9] in that instead of designing special purpose image features for each facial action, we explore general purpose learning mechanisms for data-driven facial expression classification. The approach detects not only changes in position of facial features, but also changes in image texture such as those created by wrinkles, bulges, and changes in feature shapes. Certainly the best way to advance the field is for multiple laboratories to develop different approaches, and make comparisons on standard databases such as the Cohn-Kanade database. A standard database of FACS coded spontaneous expressions would be of great benefit to the field and we are preparing to make the RU-FACS spontaneous expression database available to the research community.

Acknowledgments Support for this work was provided by NSF-ITR IIS-0220141 and NRL/HSARPA Advanced Information Technology 55-05-03.

References

[1] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[2] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[3] Ian R Fasel, Bret Fortenberry, and Javier R Movellan. A generative framework for real-time object detection and classification. *Computer Vision and Image Understanding*, 98, 2005.

[4] 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.

[5] A. Kapoor, Y. Qi, and R.W.Picard. Fully automatic upper facial action recognition. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures.*, 2003.

[6] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, and J.R. Movellan. An automatic system for measuring facial expression in video. *Image and Vision Computing*, in press.

[7] T. K. Marks, J. Hershey, J. Cooper Roddey, and J. R. Movellan. 3d tracking of morphable objects using conditionally gaussian nonlinear filters. In *CVPR Workshop on Generative-Model Based Vision*.

[8] M. Pantic and J.M. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man and Cybernetics*, 34(3):1449–1461, 2004.

[9] 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.

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