Real-time inference of mental states from facial expressions and upper body gestures

Tadas Baltrušaitis, Daniel McDuff, Ntombikayise Banda, Marwa Mahmoud,
Rana el Kaliouby, Peter Robinson and Rosalind Picard

***Abstract*— We present a real-time system for detecting facial action units and inferring emotional states from head and shoulder gestures and facial expressions. The dynamic system uses three levels of inference on progressively longer time scales. Firstly, facial action units and head orientation are identified from 22 feature points and Gabor filters. Secondly, Hidden Markov Models are used to classify sequences of actions into head and shoulder gestures. Finally, a multi level Dynamic Bayesian Network is used to model the unfolding emotional state based on probabilities of different gestures. The most probable state over a given video clip is chosen as the label for that clip. The average F1 score for 12 action units (AUs 1, 2, 4, 6, 7, 10, 12, 15, 17, 18, 25, 26), labelled on a frame by frame basis, was 0.461. The average classification rate for five emotional states (anger, fear, joy, relief, sadness) was 0.440. Sadness had the greatest rate, 0.64, anger the smallest, 0.11.**

I. INTRODUCTION

The automated recognition of emotional states is an important part of the development of affect sensitive AI systems [1]. Non-verbal cues such as facial expressions and gestures contain a significant amount of affective information. We present a dynamic model to infer facial actions, upper body gestures and mental states from video.

This work is based on the mind-reading model presented by el Kaliouby and Robinson [2]. This model infers complex mental states from head and facial expressions. The initial evaluation considered six mental states: agreeing, concentrating, disagreeing, interest, thinking, and uncertainty. The mind-reader model was shown to perform comparably to human labellers in labelling videos of these six mental states. We extend the system to incorporate body gestures and an expanded set of action units as well as training it for alternative mental states: anger, fear, joy, relief and sadness. Of the action units [3] detected in the original model; AU1, AU2, AU12, AU18, AU25 and AU26 are common with those for which we report performances in this study. We maintain the ability for the system to label video sequences continuously and in real-time.

II. APPROACH

A. Overview

The model presented performs inference on three progressively longer time scales and higher levels of abstraction. First, head orientation (and head action units) are detected from the configuration of 22 feature points using the FaceTracker, part of NevenVision's[1] facial feature tracking SDK. Then six action units are detected from the displacement of these features and six action units are detected using appearance-based features (excluding head orientation AUs). Secondly, Hidden Markov Models (HMMs) are used to classify sequences of action units into 23 (9 in the original system) head and facial gestures. Finally, a multi-level Dynamic Bayesian Network (DBN) is used to model the unfolding mental state based on the quantized probabilities of gestures from the HMMs. The emotional label for a given video sequence is chosen as the most probable state over the entire sequence. A hierarchical approach was chosen for this model as it limits the influence of each level on the one below it, reducing the potential parameter space and allowing for greater generalization in the case of limited training data [4].

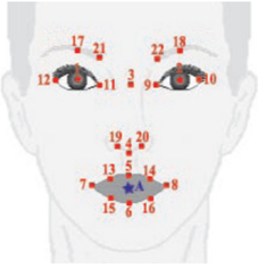

Fig. 1. The 22 feature points tracked by the NevenVision face tracker

B. AU Detection

Action Unit detection is done using two distinct approaches. The first one relies directly on tracked facial feature points to detect AU1, AU2, AU10, AU12, AU25, and AU26. The second one uses Gabor features, principal component analysis (PCA) and Support Vector Machines (SVMs) on regions estimated from the tracked feature points to detect AU4, AU6, AU7, AU15, AU17, and AU18.

This extension of the original approach was necessary because some AUs do not manifest themselves straightforwardly in movements of the feature points. Our approach combines the benefits of a geometric analysis of a face together with appearance based analysis, and results in more robust emotion detection. This is because the geometric approach is less sensitive to head orientation and lighting conditions, while the appearance based approach provides more accuracy in frontal poses.

T. Baltrušaitis, N. Banda, M. Mahmoud and P. Robinson are with the Computer Laboratory, University of Cambridge, Cambridge, UK, tadas.baltrusaitis, ntombikayise.banda, marwa.mahmoud, peter.robinson@cl.cam.ac.uk

D. McDuff, R. el Kaliouby and R. Picard are with the MIT Media Laboratory, Cambridge, USA, djmcduff, kaliouby, picard@media.mit.edu

[1] Licensed from Google Inc.

TABLE I

AU DETECTION RULES.

AU	Rule
1	$\angle(P_{12}-P_{11},P_{12}-P_{21})+\angle(P_{10}-P_9,P_{10}-P_{22})>\tau$
2	$\angle(P_{12}-P_{11},P_{12}-P_{21})+\angle(P_{10}-P_9,P_{10}-P_{22})>\tau$
10	$\lvert P_3-P_5\rvert>\epsilon$
12	$(\angle(P_7-A,P_8-A)+\angle(P_7-A,P_8-A)>\tau$ AND $\lvert A-P_7\rvert+\lvert A-P_8\rvert>\epsilon)$ OR $\lvert P_7-P_{11}\rvert+\lvert P_9-P_8\rvert<\epsilon$
25	$\lvert P_6-P_5\rvert>\epsilon$
26	$\lvert P_3-P_6\rvert>\epsilon$

1) Geometry: Geometry based detection relies on the use of 22 tracked feature points on the face (see Fig 1). Tracking is done using the NevenVision tracker, which in addition to the feature points provides the head orientation (tilt, pitch, and yaw). The tracker uses Gabor wavelet image transformations and neural networks for the tracking of subsequent images in the video, it is fully automatic and requires no manual labelling. The tracker is robust to a certain amount of out-of-plane head motion, and is good at detecting head pose. When evaluated on the Boston University dataset [5] the NevenVision tracker absolute errors of orientation estimation were as follows: roll $\mu = 3.09°, \sigma = 2.58°$, pitch $\mu = 5.73°, \sigma = 7.94°$, and yaw $\mu = 5.19°, \sigma = 4.74°$.

The original system relied on the initial frame of a video to provide the estimate of a neutral expression. Because the training and test sets used did not have the neutral expression available, we decided to create a model for neutral face estimation based on static facial features. This considerably increased the AU detection rates on the training dataset and made our system more practical for real-life applications where the neutral face is usually not available.

Since facial morphologies vary a lot across people we did not want to simply adopt an average neutral expression across a large sample of faces. To estimate a neutral face, we constructed a face model from 886 successfully tracked neutral face images from the MultiPie dataset [6]. The model was constructed by running our feature tracker on the images and creating a vector $\mathbf{x} = \{x_1, x_2, ..x_{22}, y_1, y_2, ...y_{22}\}$ (where x_i, and y_i are the automatically located feature point coordinates) for each sample face. The samples were then normalised for translation, rotation, and scaling using Procrustes analysis [7]. PCA was performed on the samples to determine the main modes of variation in the face shape, retaining 95% of variability explained by the model. The neutral face model could be expressed using $\mathbf{x} = \bar{\mathbf{x}} + \Phi\mathbf{p}$, where $\bar{\mathbf{x}}$ is the mean shape, Φ a matrix of the modes of variation and $\mathbf{p}$ the parameters, controlling the shape.

As an approximation we wanted to determine the neutral expression of a person from a single frame of a non-neutral expression, so we had to use features that remain stable under varying facial expressions. For this we used the corners of the eyes and the nose tip. We can model these parameters as linear combinations of the static features of the face:

$$\mathbf{p} = \mathbf{p_0} + a\mathbf{p_1} + b\mathbf{p_2} + c\mathbf{p_3} + d\mathbf{p_4} \tag{1}$$

where a, b, c, and d are respectively the scale normalised distance between eyes, distance from nose tip to nose root, average distance from eye corners to nose tip, and eye widths, and $\mathbf{p_0}, \mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}$ are vectors estimated from training data using linear regression between $\{\mathbf{p_i}\}$ and the vectors $\{(a, b, c, d)'\}$.

Features detected in the first frame are scale normalised, and rotation corrected using the head orientation and an approximation that facial points lie in the same plane. They are then used with Equation 1 and the PCA model to estimate the neutral expression. In all the subsequent frames this neutral expression approximation is used to detect AUs.

For AU detection we use the hand coded rules listed in Table I, they are taken from the original system [2] with several modifications taken from Pantic et al.[8]. The rules compare the angles and distances between the points when compared to the neutral expression (for distances between points the current and neutral expression ratio was used, for angles the angle differences were used). Several other rules have been tested and the ones that reached the highest F1 score on the training dataset were chosen for the final system. The thresholds used have been determined from the training dataset as well. Feature point tracking of eyebrows was not accurate enough to reliably disambiguate between AU1 and AU2, thus the rules for their detection are the same.

2) Gabors: AU4, AU6, AU7, AU15, AU17 and AU18 were implemented using Gabor features, PCA and SVMs. For each action unit a similar process was carried out. A 10 x 10 pixel region, located using the relevant feature points was taken. Gabor features on three scales and six orientations were calculated giving a feature vector of length 1800. The mean was subtracted from the data and PCA used to reduce the dimensionality. 45 features were extracted; in all cases this accounted for at least 95% of the energy. These features were used to train a binary SVM (one per AU). Radial Basis Function (RBF) kernels were used in all cases. Validation was performed using five-fold cross validation. The dataset was randomly split into five parts and the cross validation accuracy calculated for each. During validation the penalty parameter C, and the RBF kernel parameter, γ, were each varied from 10^k with k = -3, -2,...,3.

C. Emotion Detection

Emotion recognition is achieved by employing the hierarchical framework shown in Figure 2. The framework consists of three levels: actions, gestures and emotions. In the first level, video input is spatially abstracted into action units. This is followed by the recognition of gestures from the sequence of action units detected. The third level combines a set of gestures through a feature-selection algorithm to infer emotion. The gesture recognition and emotion classification levels are discussed in detail below.

1) Gestures: Gesture recognition is based on three modalities: head, face, and body.

Facial and head action units are quantized and input into left-to-right HMM classifiers to identify face and head gestures. Each gesture is modeled as a temporal sequence of action units (e.g. a head nod is a series of alternating up

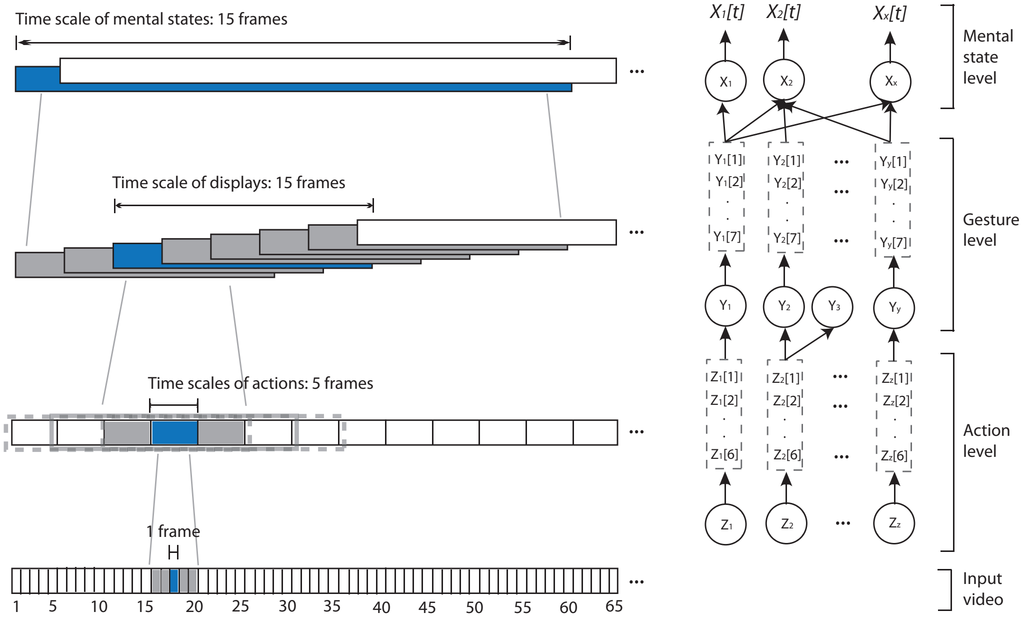

Fig. 2. The multi-level overview of the system

and down head movements). The advantage of using HMMs is to model the system spatio-temporally, and deal with the time warping problem. Several HMM topologies are used for gesture recognition. For example, the head nod HMM is a 4-state, 3-symbol HMM, where the symbols correspond to head up, head down, and no action. We use a similar topology for head shakes and supported mouth displays. For tilt and turn displays we use a 2-state HMM with 3 observable symbols. The symbols encode the intensity of the tilt and turn motions. Maximum likelihood training is used to determine the parameters of each HMM model $\lambda = \{\Lambda,\ \beta,\ \pi\}$ offline, described by transition probabilities, the probability distributions of the states, and priors. For each model λ and a sequence of observations $O = \{o_1,\ o_2,\ ...,o_T\}$ the forward-backward algorithm determines the probability that the observations are generated by the model.

Since we are interested in upper body videos, body information is incorporated into gesture recognition through shoulder information. A Hough line transform is used to detect the right and left shoulder angles. First, a Canny edge detector is applied on each frame followed by the standard Hough transform to detect all the lines in the image. Empirically derived, we assume that a shoulder angle ranges from the horizontal axis down to 40 degrees. Therefore, from all the detected lines, we extract the lines whose angles fall within this range, as they are likely to represent shoulder lines. After those lines are extracted, we compute the average angle for right and left shoulders. The angle difference between each frame and the initial frame is calculated and compared to a threshold to determine three shoulder gestures: shoulder up, shoulder down, and shoulder shake.

2) Emotions: Our system uses DBNs to represent the interaction between hidden emotions and the set of observable gestures. The dynamic structure of DBNs makes them able to characterize the emotion-gesture interaction more efficiently and able to capture the temporal dynamics of gestures. This requires the videos analysed to have temporal information that unfolds over time. Such information is inherent in complex mental states but not necessarily in basic emotions (whereupon the system is evaluated).

Emotion inference is carried out in real-time with each emotion modelled as a separate DBN classifier. The inference engine employs a sliding window technique which allows six gestures to be observed at every time instance t. The gestures are then used to compute a likelihood indication of how much they resemble each emotion. The inference decision is obtained by integrating the probabilities of each emotion over the entire video timeline and selecting the emotion with the greatest probability. A detailed description of the emotion classification can be found in [9].

III. EVALUATION

A. Dataset

The training and test data sets are the GEMEP-FERA datasets, which are the subsets of the GEMEP corpus [10]. The training dataset consists of recordings of 10 actors displaying a range of expressions, while uttering a meaningless phrase, or the word ‘Aaah’. The test dataset contains of six subjects: three of them are present in the training data (person-specific data), while the other three are new (person-independent data). Videos in the dataset are short videos of the upper body that do not start from a neutral expression. Average video length $\approx$ 2.67 seconds.

B. AU

The FERA2011 AU recognition challenge involves the classification of 12 AUs: 1, 2, 4, 6, 7, 10, 12, 15, 17, 18, 25 and 26. There were 87 videos in the training set and 71 in

TABLE II

F1 scores for AU detection. Baseline used is the one proposed by challenge organisers [11].

AU	Training Data	Person Independent	Person Specific	Overall (excluding training)	Baseline (person independent)	Baseline (person specific)	Baseline (overall)
AU1	0.572	0.681	0.445	0.615	0.633	0.362	0.567
AU2	0.501	0.635	0.466	0.580	0.675	0.400	0.589
AU4	0.974	0.446	0.393	0.427	0.133	0.298	0.192
AU6	0.988	0.739	0.458	0.671	0.536	0.255	0.463
AU7	0.987	0.323	0.433	0.371	0.493	0.481	0.489
AU10	0.520	0.327	0.383	0.349	0.445	0.526	0.479
AU12	0.648	0.657	0.556	0.625	0.769	0.688	0.742
AU15	0.969	0.113	0.173	0.144	0.082	0.199	0.133
AU17	0.851	0.300	0.189	0.275	0.378	0.349	0.369
AU18	0.930	0.127	0.178	0.148	0.126	0.240	0.176
AU25	0.762	0.815	0.823	0.818	0.796	0.809	0.802
AU26	0.597	0.475	0.565	0.513	0.371	0.474	0.415
Average	**0.775**	**0.470**	**0.422**	**0.461**	**0.453**	**0.423**	**0.451**

the test set. AUs were labelled frame-by-frame for presence (without intensity values). The frames including speech were not used in the evaluation of AU25, and AU26. No matching of people in training and testing datasets was performed.

For the feature point based AU detection approach the training data was used to adjust thresholds, for the appearance based approach it was used to extract examples for classifier training. The results of running our AU detection system on the test and training datasets are listed in the Table II. The first column shows the results of our system evaluated on the training data. The results are presented as F1 scores which combine the precision and recall of the detection.

From the results it can be seen that in general lower face AUs had an average score of 0.304 without AU25 and AU26, and 0.411 with AU25 and AU26, and performed worse than upper face AUs with an average score of 0.532. AU25 and AU26 performed much better (average score 0.665) than other lower face AUs. This is possibly because they were scored only for the frames not including speech, showing our systems weakness in AU detection when speech is present.

The results for the training data show that the feature point based system generalised much better than the Gabor based one, which performed much worse (possibly due to overfitting). This shows that the feature based approach would be more suitable for systems that are expected to perform well under different conditions.

Our system outperformed the baseline system of that proposed by the challenge organisers [11].

C. Emotion

The FERA2011 emotion recognition challenge involves the classification of the following five emotions: anger, fear, joy, relief and sadness. The emotion recognition system was trained offline on the GEMEP-FERA training dataset which consisted of 155 videos portraying five emotions using seven actors. Each emotion DBN model was trained by assigning videos depicting the emotion of interest as positive samples, and passing a sample of the other emotions as negative samples. This increases the ability of the DBN to discriminate by assigning penalties to gestures that best describe other emotions. The gesture compositions of the resulting models are shown Figure 3. The figure compares the weights of the seven gestures that contributed the most to the selection of emotions. Negative weights serve as penalties that strengthen the discriminative ability of a model.

The emotion recognition system was then evaluated on the FERA2011 emotion test dataset which consists of 134 videos depicting six actors, three of which were not in the training dataset. Table III lists the classification rates of the system.

TABLE III

Emotion Classification Rates

Emotion	Person Independent	Person Specific	Overall
anger	0.214	0.000	0.111
fear	0.467	0.200	0.360
joy	0.650	0.364	0.548
relief	0.375	0.800	0.538
sadness	0.533	0.800	0.640
Average	**0.448**	**0.433**	**0.440**

The corresponding confusion matrix for the overall test set is shown in Table IV.

The overall classification rates indicate that the system had trouble tracking emotions that exhibit extreme intensity such as anger and fear, with anger bearing the lowest rate at 11.1%. The system boasts classification rates of 80% for

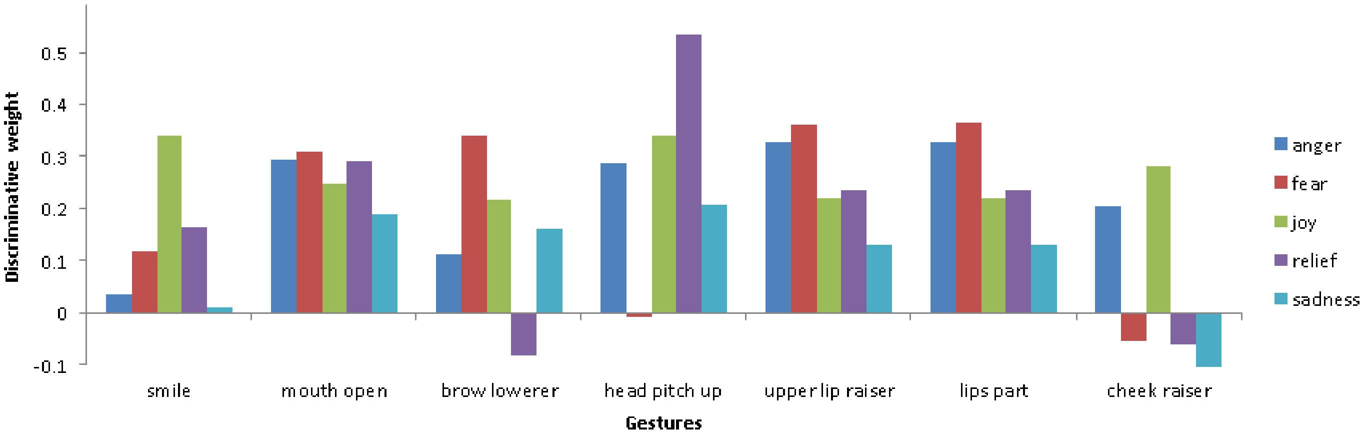

Fig. 3. Composition and discriminative weights of gestures as modelled by the DBNs for emotion recognition

TABLE IV

CONFUSION MATRIX FOR EMOTION RECOGNITION TEST

pred\truth	anger	fear	joy	relief	sadness
anger	3	3	5	2	2
fear	7	9	3	2	0
joy	4	4	17	1	1
relief	7	3	5	14	6
sadness	6	6	1	7	16

relief and sadness in the person-specific partition test, and fares well overall for joy, relief and sadness. On average the system fared better on the person independent partition confirming the notion that probabilistic models are able to generalize well.

When viewing these classification rates in relation to the distribution factor graphs, it can be noted that all emotions shared *mouth open* as a common gesture which carried a significant weight in the classification of emotions. Since most videos portrayed a speaking actor, this led to all emotions being awarded high probabilities and making the system rely heavily on gestures that were not as common. This is also evident in gestures such as *lips part* and *upper lip raiser* which are derivations of mouth motions.

Another reason for poor classification results of anger might be the intensity of the expression that leads to more false positives in AU detection confusing the overall classification. This might be confirmed by the good performance of less intense emotions such as relief and sadness.

The system outperformed the Uniform Local Binary Patterns baseline for the overall classification of fear, relief and sadness with our system yielding 1.25, 3.67 and 2.2 factor increases for the respective emotions. It also outperformed the random system baseline on four of the five emotions.

TABLE V

PERFORMANCE COMPARISON OF OUR SYSTEM TO THE BASELINE

Emotion	LBP Baseline	Naive (Random) Baseline	MindReader
anger	0.890	0.222	0.111
fear	0.200	0.160	0.360
joy	0.710	0.161	0.548
relief	0.460	0.115	0.538
sadness	0.520	0.20	0.640
Average	**0.560**	**0.172**	**0.440**

From these results, one can therefore conclude that the training set for the DBN models was insufficient, and that mouth gestures were the largest contributors to the misclassification of results where the gesture-based decision is concerned. Another contributing factor is the short duration of the videos as it is commonly agreed upon that for reliable recognition of emotion, video time lengths need to exceed 2 seconds [9]. Close to half of the test set comprised videos of less than 2 seconds in duration.

In addition, the original design of our system was for detecting complex mental states which unfold over time and have complex temporal dynamics. Basic emotions lack such dynamics, and therefore the system is not as good for their recognition. This is supported by the fact that our system performed better on subtle expressions of relief and sadness.

IV. DISCUSSION

We have presented a real-time system that can automatically infer human mental states from facial expressions and head and shoulder gestures.

Our system allows video sequences to be labelled continuously and in real-time. The presence of each action unit and the likelihood of each gesture and emotional state are reported on a frame by frame basis. Also, the use of probabilistic models, HMMs and DBNs, improves the systems robustness to missing data. In addition to this the modular architecture of the system makes it easy to modify and improve specific modules or add new action units, features, gestures, or mental states without changing the main architecture of the system.

One limitation of our AU detection system for the six geometry based AUs is the lack of availability of a neutral expression. We used an approximation of a neutral expression to alleviate this. This created a problem where the system would not perform as well on people whose static features did not predict the neutral expression well under our model. But even with the lack of a verified neutral expression from the video our system still managed to recognise AUs from feature points through the use of our neutral face model.

The system was evaluated on the GEMEP-FERA dataset to verify the recognition of 12 AUs (AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU15, AU17, AU18, AU25, AU26) and five emotions (anger, fear, joy, relief, sadness), as outlined in the challenge guidelines [11]. Our system outperformed the proposed LBP baseline for AU detection. It also outperformed the LBP baseline for three out of five emotions.

The test dataset does not contain complex mental states (except for relief) or neutral expressions. Recent studies showed that the real world is dominated by neutral expressions [12] and complex mental states [13]. Thus, this dataset might not be a good model for the real world, where affect recognition systems would be ultimately used. It is also hard to tell how the performance on recognizing acted emotions would generalise to naturally occurring data. There is growing evidence [14] about the differences between the acted and naturally expressed emotions.

In addition, there is evidence that even AU amplitude and timings differ in spontaneous and acted expressions [15]. Ideally these AU and emotion recognition systems are to be used in natural and spontaneous environments, but it is not entirely obvious how the recognition rates on the challenge dataset would transfer to such situations.

There are several ways to improve our system. Currently, our emotion classification system takes into account the spatio-temporal relationship between AUs, but their intensity, offset, and onset properties are not considered. Since some emotions share the same action units but with different properties, the addition of such features could improve the overall performance of the system. Moreover, the HMMs for gesture recognition that we used in our system were the ones used in the work of el Kaliouby and Robinson [2] without any retraining. This emphasizes the fact that the system generalizes well to new datasets. On the other hand, retraining the HMMs could improve the recognition results.

The modular architecture of the system makes it easy to modify and improve any specific modules or add new action units, features, gestures, or mental states without changing the main topology of the system.

V. ACKNOWLEDGMENTS

We acknowledge funding support from EPSRC, Thales Research and Technology (UK), Bradlow Foundation Trust, Yousef Jameel Scholarships, the Media Laboratory Consortium and Procter and Gamble. We would like to thank Abdelrahman N. Mahmoud and Youssef Kashef for their help in the programming.

REFERENCES

[1] M. Pantic and L. Rothkrantz, “Toward an affect-sensitive multimodal human-computer interaction,” *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[2] R. Kaliouby and P. Robinson, “Generalization of a vision-based computational model of mind-reading,” *ACII*, 2005.

[3] P. Ekman and W. Friesen, *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press, 1977.

[4] N. Oliver, E. Horvitz, and A. Garg, “Hierarchical representations for learning and inferring office activity from multimodal information,” in *ICMI*, 2002.

[5] M. L. Cascia, S. Sclaroff, and V. Athitsos, “Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models,” *IEEE Trans. PAMI.*, 2000.

[6] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Active Shape Models-Their Training and Application,” *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, January 1995.

[8] M. Pantic and L. Rothkrantz, “Facial Action Recognition for Facial Expression Analysis from Static Face Images,” *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 34, 2004.

[9] R. E. Kaliouby, P. Robinson, S. Keates, and E. D. Centre, “Temporal context and the recognition of emotion from facial expression,” in *Proceedings of the HCI International Conference*. APA, 2003.

[10] T. Bnzinger and K. R. Scherer, “Introducing the geneva multimodal emotion portrayal(gemep) corpus,” in *Blueprint for Affective Computing: A Sourcebook*, K. R. Scherer, B. T., and R. E. B., Eds. Oxford University Press, Oxford, 2010, pp. 271–294.

[11] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, “The first facial expression recognition and analysis challenge,” in *IEEE Intl Conf. Automatic Face and Gesture Recognition*, 2011.

[12] S. Afzal and P. Robinson, “Natural affect data - collection & annotation in a learning context,” 2009.

[13] P. Rozin and A. B. Cohen, “High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans,” *Emotion*, vol. 3, no. 1, pp. 68 – 75, 2003.

[14] R. Cowie, “Building the databases needed to understand rich, spontaneous human behaviour,” in *Automatic Face & Gesture Recognition, 2008. FG ’08. 8th IEEE International Conference on*, 2008.

[15] J. Cohn and K. Schmidt, “The timing of facial motion in posed and spontaneous smiles,” *IJWMIP*, 2004.