AN INTELLIGENT SYSTEM FOR FACIAL EMOTION RECOGNITION

R. Cowie, E. Douglas-Cowie [1], J.G Taylor [2], S. Ioannou, M. Wallace, S. Kollias [3]

[1] Department of Psychology, Queen's University of Belfast
Northern Ireland, United Kingdom
email: r.cowie@qub.ac.uk
[2] Mathematics Department, King's College,
University of London, Strand, London WC2R 2LS, UK
email: john.g.taylor@kcl.ac.uk
[3] Department of Electrical and Computer Engineering
National Technical University of Athens,
Heroon Polytechniou 9, 157 80 Zographou, Greece
email: {sivann, wallace, stefanos}@image.ntua.gr

ABSTRACT

An intelligent emotion recognition system, interweaving psychological findings about emotion representation with analysis and evaluation of facial expressions has been generated and its performance has been investigated with experimental real data. Additionally, a fuzzy rule based system has been created for classifying facial expressions to the six archetypal emotion categories. The continuous 2-D emotion space was then examined and a pool of known and novel classification and clustering techniques have been applied to our data obtaining high rates in classification and clustering into quadrants of the emotion representation space.

1. INTRODUCTION

Despite common belief, social psychology research has shown that conversations are usually dominated by facial expressions, and not spoken words, indicating the speaker's predisposition towards the listener. Mehrabian indicated that the linguistic part of a message, that is the actual wording, contributes for seven percent to the effect of the message as a whole, the paralinguistic part, that is how text is vocalized, contributes for thirty eight percent, while a speaker's facial expression contributes for fifty five percent to the effect of the spoken message [1]. This implies that facial expressions form a major modality in human communication, and need to be considered by HCI/MMI systems.

Several steps have been made towards an recognition of facial expression [9],[4] either by considering and modeling facial deformations globally (holistic methods) or by measuring specific facial feature deformations (e.g. eyebrows, eyes, mouth) and creating appropriate descriptive expression models (analytic approach). We have chosen the latter approach and have created a system capable of analyzing image frames from a video stream of a speaker into MPEG-4 compliant Facial Definition Parameters (FDPs). FDPs are in turn used to calculate the Facial Animation Parameters (FAPs). The FAPs can correlate strongly with emotionality and can be used to classify a face with respect to the emotional state it expresses.

2. A RULE-BASED SYSTEM

In our research, a rule-based system for emotion recognition was created, characterising a user's emotional state in terms of the six universal, or archetypal, expressions (joy, surprise, fear, anger, disgust, sadness). A set of rules has been created in terms of the MPEG-4 FAPs for each of these expressions, by analysing the FAPS extracted from the facial expressions of the Ekman dataset. This dataset contains several images for every one of the six archetypal expressions, which, however, are rather exaggerated. A result of this fact is that the rules extracted from this dataset if used in real data, cannot have accurate results, especially if the subject is not very expressive. Table 1 illustrates the confusion matrix of the mean degree of beliefs for each of the archetypal emotions anger, joy, disgust, surprise and the neutral condition, computed over the EKMAN dataset [10], while Table 2 shows the more often activated rule for each of the above-mentioned expressions.

0-7803-9332-5/05/$20.00 ©2005 IEEE

Table 1: Results in images of different expressions

	Anger	Joy	Disgust	Surprise	Neutral
Anger	0.611	0.01	0.068	0	0
Joy	0.006	0.757	0.009	0	0.024
Disgust	0.061	0.007	0.635	0	0
Surprise	0	0.004	0	0.605	0.001
Neutral	0	0.123	0	0	0.83

Table 2: Activated rules

Expressions	Rule more often activated (% of examined frames)
Anger	47%
Joy	39%
Disgust	33%
Surprise	71%

3. FEATURE EXTRACTION

Automatic recognition of facial parameters is a difficult problem, and relatively little work has been reported [13]. Most expression evaluation systems either require facial markers [4] or manual initialization [3]. Automatic detection of the exact border of facial features is a much more difficult problem than detecting the presense of a feature in an image area especially in real-life applications. In our approach extraction is performed, resulting in a set of binary maps, indicating the position and extent of each facial feature (i.e. eyebrows, eyes, mouth and nose). The left, right, top and bottom–most coordinates of the eye and mouth masks, the left right and top coordinates of the eyebrow masks as well as the nose coordinates, are the facial feature points (FPs) which are used for defining the FAP values to be used as inputs to the emotion recognition system.

In most real-life applications nearly all video media have reduced vertical and horizontal color resolutions; moreover, the face occupies only a small percentage of the whole frame and illumination is far from perfect. While it is feasible to detect the face and all facial features, it is very difficult to find the exact boundary of each one (eye, eyebrow, mouth) in order to estimate its deformation from the neutral-expression frame. To overcome this limitation we have created a novel system that combines the result of multiple feature extractors into a final result, based on the evaluation of their performance on each frame; the fusion method is based on the observation that having multiple masks for each feature lowers the probability that all of them are invalid since each of the feature extractors usually produces different error patterns. An example showing the result of the eye mask extractors along with the fusion result in a single frame is depicted in Figure 1.

The feature masks have to be calculated in near-real time; the feature extractors which have been used for extraction of these masks, include:

1. A feed-forward back propagation neural network trained to identify eye and non-eye facial area. The network has thirteen inputs; for each pixel on the facial region the NN inputs are luminance Y, chrominance values Cr & Cb and the ten most important DCT coefficients (with zigzag selection) of the neighboring 8x8 pixel area.
2. A second neural network, with similar architecture to the first one, trained to identify mouth regions.
3. Luminance based masks, which identify eyelid and sclera regions.
4. Edge-based masks.
5. A region growing approach to detect regions of high texture based on standard deviation

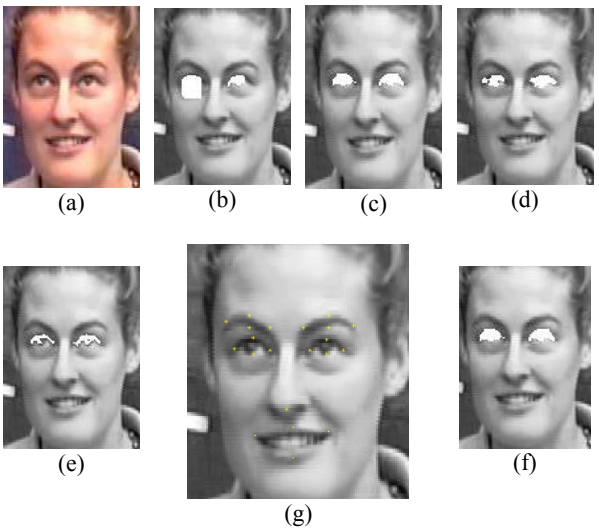

Figure 1: (a):original frame, (b),(c),(d),(e): the four detected masks, related to extractors based on (4),(5),(3),(1) correspondingly (f):fusion result mask for the eyes, (g):all detected feature points from the final masks

The mask fusion approach described in the following is not bound to specific feature extractors; more and different extractors than those described above can be developed for each feature, as long as they provide better results in difficult situations where other extractors fail. The fusion algorithm is based on a Dynamic Committee Machine (DCM) structure that combines the masks based on their validity confidence, producing a final mask together with the corresponding estimated confidence for each facial feature. Each of those masks represents the best-effort result of the corresponding mask-extraction method used. The most common problems, especially encountered in low quality input images, are connection with other feature boundaries or mask dislocation due to noise. If y_{comb} is the combined machine output and t the desired output it

has been proven in the committee machine (CM) theory [1] that the combination error $y_{comb} - t$ from different machines f_i is guaranteed to be lower than the average error:

$$(y_{comb} - t)^2 = \frac{1}{M}\sum_i (y_i - t)^2 - \frac{1}{M}\sum_i (y_i - y_{comb})^2 \quad (1)$$

In a Static CM, the voting weight for a component is proportional to its error on a validation set. In DCMs, (Figure 2) input is directly involved in the combining mechanism through a Gating Network (GN), which is used to modify those weights dynamically.

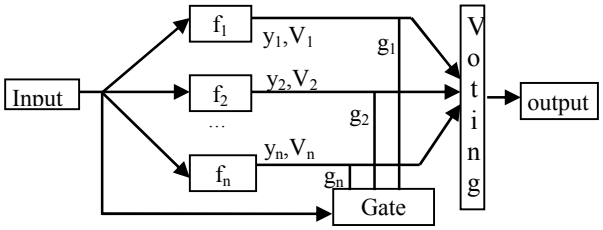

Figure 2: Dynamic Committee Machine Architecture

In our case, the final masks for the left eye, right eye and mouth, $\mathbf{M}_f^{e_L}, \mathbf{M}_f^{e_R}, \mathbf{M}_f^{m}$ are considered as the machine output and the final confidence values of each mask for feature x $M_x^{c_f}$ are considered as the confidence of each machine. Therefore, for feature *x*, each element m_f^x of the final mask $\mathbf{M}_f^x$ is calculated from the n masks as:

$$m_f^x = \frac{1}{n}\sum_{i=1}^{n} m_i^x M_f^{cx_i} h^i g^i \quad (2)$$

$$h^k = \begin{cases} 1, M_f^{c,x_k} \geq \left(t_{vd} \cdot \left\langle M_q^{c,x_k} \right\rangle_q \right) \\ 0, M_f^{c,x_k} < \left(t_{vd} \cdot \left\langle M_q^{c,x_k} \right\rangle_q \right) \end{cases} \quad (3)$$

where m_i^x is the element of mask M_i^x, M_f^{c,x_i} the validation value of mask *i* and h^i is used to prevent the masks with $M_f^{c,x_k} < \left(t_{vd} \cdot \left\langle M_q^{c,x_k} \right\rangle_q \right)$ to contribute to the final mask. A sufficient value for t_{vd} is 0.8.

The role of the gating variable g^i is to favor the color-based feature extraction methods ($\mathbf{M}_1^e, \mathbf{M}_1^m$) in images of high color and resolution. In this stage, two variables are taken into account: image resolution and color quality. More information about the used expression profiles can be found in [2].

The final feature masks are used to extract the Feature Points (FPs) considered in the definition of the FAPs. The latter are used as input features to the recognition system. Each FP inherits the confidence level of the final mask from which it derives; for example, the four FPs (top, bottom, left and right) of the left eye share the same confidence as the left eye final mask. FAPs can be estimated via the comparison of the FPs of the examined frame to the FPs of a frame that is known to be neutral, i.e. a frame which displays no facial deformations.

4. EMOTION RECOGNITION

Data showing facial expressions in normal (non extreme) interactions were generated and annotated (feeltraced in the continuous 2-D activation/evaluation space [12]) by Queens University of Belfast, using the Sensitive Artificial Listener framework (SAL), an environment where people can engage in genuine emotional expression.

The aim of the emotion recognition is to identify the quadrant [12] to which each analyzed data belongs, as an indicator of the emotional state of the person involved in the interaction (positive/negative, active/passive). At this point, we should notice that the feeltrace ratings produced by QUB were based on all three different modalities, e.g. linguistic and paralinguistic speech and facial, aiming at multimodal emotion recognition. In this paper, we describe the creation of an emotion recognition system, which is able to operate based on analysis of the recorded facial expressions. We consider three primary classes corresponding to the active quadrants of the 2-D space; the positive-passive quadrant was excluded, since no examples were generated in it.

The data examined in this research were 2 x 50.000 frames extracted from two subjects from the available datasets [6]. All frames were processed and the FAP values for each one were calculated and stored. Confidence values were created for all frames and those with low confidence were not considered further in the analysis. Next, the frames were separated in about 280 time intervals (per data set) corresponding to tunes [5] (i.e. segments of the pitch contour bounded at either end, by a pause of 180 ms or more) identified by the audio analysis of the respective speech recordings. The frame which was the most 'facially expressive' (with large FAP values) in each tune was then selected, thus generating a set of 280 frames per dataset, which was used for training purposes. A second data set of about 930 frames was also created and used for testing.

5. CLUSTERING RESULTS

A variety of techniques were used to recognize the underlying emotional states, based on FAP feature analysis. These included neural network classifiers, clustering techniques and neurofuzzy networks.

Of significant interest is usage of unsupervised hierarchical clustering (we developed a methodology for clustering with high–dimensional extensions and probabilistic refinement [11]), since this can form a basis for future merging of different emotional representations (i.e. different hierarchical levels), and categorization in either coarser or more detailed classes (half-plane, quadrants, discrete emotions).

Hierarchical Agglomerative Clustering (HAC) is a clustering approach that is ideal for cases in which the count of clusters in the data is not known before hand. On the other hand, drawbacks for HAC algorithms include high complexity and susceptibility to errors in the initial steps.

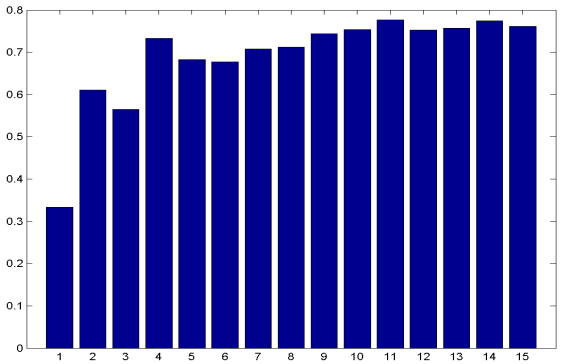

Figure 3: FAP vector clustering results; x denotes the number of clusters and y its performance on the dataset

Hierarchical clustering was used first on the generated data set to produce clusters of similar data samples. An aggregating distance function, such as the Euclidean, or Mahalanobis, distance, was used to identify the underlying patterns and produce clusters of similar data samples. Figure 3 indicates the results of the clustering procedure, over the datasets of FAP vectors (consisting of 16 values each); x denotes the number of clusters and y its performance on the dataset after the above procedure is completed. It can be seen that the obtained results, can include about 10 clusters with very good performance, i.e.; when comparing the matching of data included in the extracted clusters with their original ratings, they are in the range of 78%, which is quite high for the non-extreme emotion recognition problem we are facing. By training a neural network with the 280 FAP sets and testing generalisation on the rest of the 930 FAP sets respectively, the success rate was increased to 84,7%. Similar results have been obtained when using the Falcon-Art neurofuzzy network [8].

6. CONCLUSIONS

An emotion recognition system, combining psychological findings about emotion representation, with automatic analysis and evaluation of facial expressions, has been generated and actual performance has been investigated with experimental real data. FAP extraction based on a novel confidence-based feature extraction system was used to feed a fuzzy rule based system to classify facial expressions to six archetypal emotion categories. The continuous 2-D emotion space was then examined and a pool of known and novel classification and clustering techniques have been applied to the SAL data obtaining high rates in classification and clustering of data to quadrants of the emotion representation space.

7. REFERENCES

[1] A. Krog, J. Vedelsby, Neural network ensembles, cross validation and active learning, in Tesauro G., Touretzky D., Leen T. (Eds) Advances in neural information processing systems 7, pp. 231-238, Cambridge, MA. MIT Press, 1995.

[1] A. Mehrabian, Communication without Words, Psychology Today, vol. 2, no. 4, pp. 53-56, 1968.

[2] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis and S. Kollias, "Parameterized facial expression synthesis based on MPEG-4", EURASIP Journal on Applied Signal Processing, Vol. 2002, No. 10, pp. 1021-1038, Hindawi Publishing Corporation, October 2002.

[3] Authentic Facial Expression Analysis, N. Sebe, M.S. Lew, I. Cohen, Y. Sun, T. Gevers, T.S. Huang, International Conference on Automatic Face and Gesture Recognition (FG'04), pp. 517-522, Seoul, Korea, May 2004.

[4] B. Fasel, and J. Luettin, "Automatic Facial Expression Analysis: a survey," Pattern Recognition, Vol. 36, pp. 259-275, 2003.

[5] Cowie, R., Sawey, M. and Douglas-Cowie, E. A new speech .analysis system: ASSESS Proc International Conference of Phonetic Sciences, 3, 278-281, Stockholm, 1995

[6] ERMIS, Emotionally Rich Man-machine Intelligent System IST-2000-29319 (http://www.image.ntua.gr/ermis)

[7] J.W. Young, Head and face anthropometry of adult U.S. civilians, FAA Civil Aeromedical Institute, 1993.

[8] Lin, C.J., Lin, C.T.: An ART-Based Fuzzy Adaptive Learning Control Network. IEEE Trans. Fuzzy Systems 5(4): 477-496, 1997

[9] M. Pantic, L.J.M Rothkrantz, Automatic Analysis of Facial Expressions: The State of the Art, IEEE Transactions on PAMI, Vol.22, No.12, December 2000

[10] P. Ekman, Facial expression and Emotion. Am. Psychologist, Vol. 48, 1993.

[11] P. Mylonas , M. Wallace and S. Kollias, "Using k-nearest neighbour and feature selection as an improvement to hierarchical clustering", Methods and Applications of Artificial Intelligence, Vouros G.A., Panayiotopoulos T. (Eds.), Lecture Notes in Computer Science 3025, Springer, 2004.

[12] R. Plutchik, Emotion: A psychoevolutionary synthesis, Harper and Row, NY, USA, 1980.

[13] Ying-li Tian, Takeo Kanade and Jeffrey F. Cohn, "Recognizing Action Units for Facial Expression Analysis" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 23, No. 2, February 2001