A Robust Scheme for Facial Analysis and Expression Recognition

S. Ioannou, M. Wallace, K. Karpouzis and S. Kollias
Image, Video and Multimedia Systems Laboratory,
School of Electrical and Computer Engineering, National Technical University of Athens, GREECE

Abstract
Since facial expressions are a key modality in human communication, the automated analysis of facial images and video for the estimation of the displayed expression is central in the design of intuitive and human friendly human computer interaction systems. In this paper we present a robust integrated system able to consider issues such as uncertainty and lack of confidence in the process of feature extraction from image and video in the process of facial expression analysis and recognition. The proposed approach has been implemented in the framework of an EU funded R&D project

Keywords
Facial analysis, expression recognition, feature points, anthropometric validation.

1 Introduction
In recent years there has been a growing interest in improving all aspects of the interaction between humans and computers, providing a realization of the term "affective computing" [1]. Humans interact with each other in a multimodal manner to convey general messages; emphasis on certain parts of a message is given via speech and display of emotions by visual, vocal, and other physiological means, even instinctively (e.g. sweating) [2]. Everyday face-to-face communication utilizes many and diverse channels and modalities, increasing the flexibility of a communication scheme. In these situations, failure of one channel is usually recovered by another channel; this kind of behavior should actually be considered as a blueprint of the requirements of robust, natural and efficient multimodal HCI [3].

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 only for seven percent to the effect of the message as a whole; the paralinguistic part, that is how the specific passage is vocalized, contributes for thirty eight percent, while facial expression of the speaker contributes for fifty five percent to the effect of the spoken message [4]. This implies that the facial expressions form the major modality in human communication.

An overview of the methodologies used for automatic analysis of facial expression can be found in [5]. A usual approach to measuring deformation, fortified by the fact that there are inter-personal variations of facial action amplitude, is to refer to the neutral – expression face of a given person. An important parameter of this approach is the effectiveness of the image processing procedures. In actual situations, such as processing visual data from talk shows, many kinds of noise may hinder feature extraction: subjects turning their heads or moving their hands may lead to feature occlusion or bad and uneven lighting may hamper edge- or color-based feature extraction algorithms. As a result, the appearance and deformation of one or more features may not be available for a given frame of a video sequence; worse yet, an erroneous deformation estimate may be unknowingly fed into the knowledge representation infrastructure.

An ideal recognition system should be able to classify all visually distinguishable facial expressions; a robust and extensible face and facial action model is a vital requirement. Ideally, this would result in a particular face model setup uniquely describing a particular facial expression [6]. A usual reference point is provided by the 44 facial actions defined in FACS (Facial Action Coding System) whose combinations form a complete set of facial expressions and facial expressions with a similar facial appearance [7]. It has to be noted though, that some of the facial action tokens included in FACS may not appear in meaningful facial expressions, since the purpose of FACS is to describe any visually distinguishable facial action and not to concentrate on emotional expressions [8].

When put to practice, these principles typically suffer from the imperfection of the image or video processing components, that cannot always detect all the required facial features correctly (detected point on the face mapped to correct feature) and accurately (the exact position of the point on the face detected with absolute precision). Thus, errors, noise and uncertainty in general are inserted in the process of expression analysis from the very first step and are therefore inherent in the whole process.

2 Methodology outline
A very important requirement for an ideal facial expression architecture is that all of the processes therein have to be performed without any or with the least possible user intervention. This typically involves initial detection of the face, extraction and tracking of relevant facial information, and facial expression classification. In this framework, actual implementation and integration details are enforced by the particular application. For example, if the application domain of the integrated system is behavioral science, real-time performance may not an essential property of the system.

In the framework of MPEG-4 standard, parameters have been specified for Face and Body Animation (FBA) by defining specific Face and Body nodes in the scene graph; the initial goal of FBA definition is the animation of both realistic and cartoonist characters. Thus, MPEG-4 has defined a large set of parameters and the user can select subsets of these parameters according to the application.

MPEG-4 specifies 84 feature points on the neutral face, which provide spatial reference for Facial Animation Parameter (FAP) definition; these feature points are

presented in Figure 1. FAPs are defined through the comparison of distances between pairs of feature points on the observed and the neutral face. Most of the techniques for facial animation are based the well-known system for describing "all visually distinguishable facial movements" FACS. FACS is an anatomically oriented coding system, based on the definition of "Action Units" (AU) of a face that cause facial movements. An Action Unit could combine the movement of two muscles or work in the reverse way, i.e., split into several muscle movement. The FACS model has inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard [9]. In particular, the Facial Definition Parameter (FDP) and the Facial Animation Parameter set were designed in the MPEG-4 framework to allow the definition of a facial shape and texture through FDPs, thus eliminating the need for specifying the topology of the underlying geometry, and the animation of faces through FAPs, thus reproducing expressions, emotions and speech pronunciation.

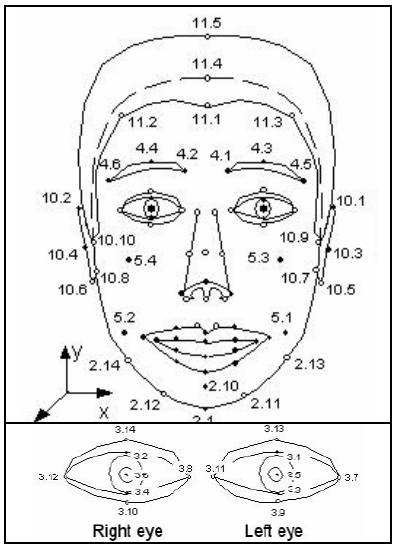

Figure 1. *MPEG-4 defined feature points on the neutral face.*

A long established tradition attempts to define facial expression in terms of qualitative targets, i.e. static positions capable of being displayed in a still photograph. The still image usually captures the apex of the expression, i.e. the instant at which the indicators of emotion are most marked. More recently emphasis, has switched towards descriptions that emphasize gestures, i.e. significant movements of facial features. Either way, analysis of the emotional expression of a human face requires a number of pre-processing steps. Following the most recent approach that emphasizes facial gestures, the required raw processing steps are to detect or track the face, to locate characteristic facial regions such as eyes, mouth and nose on it, to extract and follow the movement of facial features, such as characteristic points in these regions, or model facial gestures using anatomic information about the face. Continuing, extracted information needs to be combined with higher level knowledge, mapping detected facial feature movements to their corresponding facial expressions.

In our approach, we start by detecting the face, initially by using variance, mean color and number of skin color pixels in order to discard most candidate frames and applying more sophisticated face detection techniques on the rest. The detected face is pre-processed in order to roughly estimate regions of interest for i) the eyes and eyebrows and ii) the mouth. Each one of these ROIs is processed by a variety of methodologies in order to extract required facial features as accurately and certainly as possible. Distances between these feature points define the FAPs, which are combined in fuzzy rules in order to provide an estimation of the observed facial expression. The step in which uncertainty is most inherent, i.e. that of image processing for feature extraction, is analyzed in the following section.

3 Feature extraction

Besides expression representation, an important parameter of the expression analysis process is the effectiveness of the image processing procedures. Automatic analysis systems usually require good input to avoid misclassification or errors which is often ensured by the use of specific environment conditions such as in [13]. In actual situations, such as processing visual data from talk shows, many kinds of noise may hinder feature extraction: subjects turning their heads or moving their hands may lead to feature occlusion or bad and uneven lighting may hamper edge- or color-based feature extraction algorithms. As a result, the appearance and deformation of one or more features may not be available for a given frame of a video sequence; worse yet, an erroneous deformation estimate may be unknowingly provided as input to the subsequent expression analysis and classification procedures.

In this work, precise facial feature extraction is performed resulting in a set of masks, i.e. binary maps indicating the position and extent of each facial feature. 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 used to define the feature points. For the nose and each of the eyebrows, a single mask is created. On the other hand, since the detection of eyes and mouth can be problematic in low-quality images, a variety of methods is used, each resulting in a different mask. In total, we have four masks for each eye, three for the mouth and one for each one of the eyebrows. The methodologies applied in the extraction of these masks include:

- 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.
- A second neural network, with similar architecture to the first one, trained to identify mouth regions.
- Luminance based masks, which identify eyelid and sclera regions.
- Edge-based masks.
- A region growing approach based on standard deviation

Since, as we already mentioned, the detection of a mask using any of these applied methods can be problematic, all detected masks have to be validated against a set of criteria; of course, different criteria are applied to masks of different facial features. Each one of the criteria examines the masks in order to decide whether they have acceptable

size and position for the feature they represent. This set of criteria consist of relative anthropometric measurements, such as the relation of the eye and eyebrow vertical positions, which when applied to the corresponding masks produce a value in the range [0,1] with zero denoting a totally invalid mask; in this manner, a validity confidence degree is generated for each one of the initial feature masks. For example, two criteria that can be used for the validation of the eye masks are the following:

$$M_{eye}^{1c} = 1 - \left| 1 - \frac{d_2 / d_6}{0.49} \right| \text{ and } M_{eye}^{2c} = 1 - \frac{|d_4|}{d_5}$$

where M_{eye}^{1c} and M_{eye}^{2c} are the confidence degrees acquired trough the application of each validation criterion on an eye mask. The former of the two criteria is based on [11], where the ration of eye width over bipupil breadth is reported as constant and equal to 0.49. In almost all cases these validation criteria, as well as the other criteria utilized in mask validation, produce confidence values in the [0,1] range. In the rare cases that the estimated value exceeds the limits, it is set to the closest extreme value, zero for negative values and one for values exceeding one..

For the features for which more than one masks have been detected using different methodologies, the multiple masks have then to be fused together to produce a final mask. The choice for mask fusion, rather than simple selection of the mask with the greatest validity confidence, is based on the observation that the methodologies applied in the initial masks' generation produce different error patterns from each other, since they rely on different image information or exploit the same information in fundamentally different ways. Thus, they provide independent information on the location on the mask; combining information from independent sources has the property of alleviating a portion of the uncertainty present in the individual information components. In other words, the final masks that are acquired via mask fusion are accompanied by lesser uncertainty than each one of the initial masks.

The fusion algorithm is based on a Dynamic Committee Machine structure that combines the masks based on their validity confidence, thus producing a final mask together with the corresponding estimated confidence. As already explained, this confidence degree is always higher than the degree of any of the considered initial masks. A final, more refined, confidence value can be acquired when also taking into account the temporal information from the video sequence. The final confidence for each feature mask is based on three parameters: absolute anthropometric measurements based on [11], face symmetry exploitation and examination of the facial feature size constancy over a period of ten frames. The outcome of this procedure is a set of final masks along with the final confidence of their validity.

A way to evaluate our feature extraction performance is Williams' Index (WI) [12], which compares the agreement of an observer with the joint agreement of other observers. An extended version of WI which deals with multivariate data can be found in [13]. The modified Williams' Index *I'* divides the average number of agreements (inverse disagreements, $D_{j,j'}$) between the computer (observer 0) and *n-1* human observers (*j*) by the average number of agreements between human observers:

$$WI = \frac{\frac{1}{n}\sum_{j=1}^{n} \frac{1}{D_{0,j}}}{\frac{2}{n(n-1)} \sum_{j} \sum_{j':j'>j} \frac{1}{D_{j,j'}}}$$

and in our case we define the average disagreement between two observers *j,j'* as

$$D_{j,j'} = \frac{1}{D_{bp}} \left\| M_j^x \veebar M_{j'}^x \right\|$$

where $\veebar$ denotes the pixel-wise xor operator, $\left\| M_j^x \right\|$ denotes the cardinality of feature mask *x* constructed by observer *j*, and D_{bp} (see table 1) is used as a normalization factor to compensate for camera zoom on video sequences.

These feature masks are used to extract the Feature Points (FPs) considered in the definition of the FAPs used in this work. 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. Continuing, 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 is accepted by default as one displaying no facial deformations. For example, FAP F_{37} is estimated as:

$$F_{37} = \left\| FP_{4.5}^n - FP_{3.11}^n \right\| - \left\| FP_{4.5} - FP_{3.11} \right\|$$

where FP_i^n, FP_i are the locations of feature point i on the neutral and the observed face, respectively, and $\left\| FP_i - FP_j \right\|$ is the measured distance between feature points i and j. Obviously, the uncertainty in the detection of the feature points propagates in the estimation of the value of the FAP as well. Thus, the confidence in the value of the FAP, in the above example, is estimated as $F_{37}^c = \min(FP_{4.5}^c, FP_{3.11}^c)$. On the other hand, some FAPs may be estimated in different ways. For example, FAP F_{31} is estimated as:

$$F_{31}^1 = \left\| FP_{3.1}^n - FP_{3.3}^n \right\| - \left\| FP_{3.1} - FP_{3.3} \right\|$$

or as

$$F_{31}^2 = \left\| FP_{3.1}^n - FP_{9.1}^n \right\| - \left\| FP_{3.1} - FP_{9.1} \right\|$$

As argued above, considering both sources of information for the estimation of the value of the FAP alleviates some of the initial uncertainty in the output. Thus, for cases in which two distinct definitions exist for a FAP, the final value and confidence for the FAP are as follows: $F_i = \frac{F_i^1 + F_i^2}{2}$.

The amount of uncertainty contained in each one of the distinct initial FAP calculations can be estimated by $E_i^1 = 1 - F_i^{1c}$ for the first FAP and similarly for the other. The uncertainty present after combining the two can be given by some t-norm operation on the two

$E_i = t(E_i^1, E_i^2)$ .The Yager t -norm with parameter $w = 5$ gives reasonable results for this operation:

$$E_i = 1 - \min\left(1, \left((1-E_i^1)^w + (1-E_i^2)^w\right)^w\right)$$

The overall confidence value for the final estimation of the FAP is then acquired as $F_i^c = 1 - E_i$. While evaluating the expression profiles, FAPs with greater uncertainty must influence less the profile evaluation outcome, thus each FAP must include a confidence value. This confidence value is computed from the corresponding FPs which participate in the estimation of each FAP.

Finally, FAP measurements are transformed to antecedent values x_j for the fuzzy rules using the fuzzy numbers defined for each FAP, and confidence degrees x_j^c are inherited from the FAP $x_j^c = F_i^c$ where F_i is the FAP based on which antecedent x_j is defined.

4 Expression recognition

In previous work we have defined expression vocabularies, i.e. the set of FAPs that each may be activated for each expression, and expression profiles, i.e. sets of FAP values, each profile representing a specific instance of the expression [14].

Each profile is easily transformed into a fuzzy rule, thus leading to the generation of a neurofuzzy classifier that, given the FAP values extracted from a still image as input, provides an estimation of the user expression as output.

In order to further reduce the uncertainty in the facial expression estimation, one may consider that although expression varies rapidly, emotion – and thus groups of emotion – do not vary equally rapidly. Based on this observation evidence theory can be used in order to combine the findings of the analysis, when applied on consecutive or almost consecutives frames of a shot [15].

5 Conclusions

Automatic analysis of facial pictures and video is an essential tool towards estimation of the human emotion in HCI. In this paper we have presented the abstract description of a system that is able to tackle the task, taking into account the uncertainty that is inherent in the comprising steps and utilizing anthropometric criteria and dynamic committee machines in order to alleviate a part of it.

The theory and methodologies presented herein have been developed in the framework of the ERMIS IST [16] project and are also being applied and extended in the HUMAIN European Network of Excellence [17], through the participation of the Image, Video and Multimedia Systems Laboratory of the National Technical University of Athens.

6 References

1. Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J. (1999). Measuring Facial Expressions by Computer Image Analysis, *Psychophysiology*, 36, 253-263.
2. Cowie, R., Douglas-Cowi,e E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. (2001). Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*.
3. Ekman, P. (1982). *Emotion in the Human Face*. Cambridge Univ. Press.
4. Ekman, P., Friesen, W.V. (1978) *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto.
5. Mehrabian, A. (1968). Communication without Words, *Psychology Today*, 2(4), 53-56.
6. Picard, R.W. (1997). *Affective Computing*, MIT Press, Cambridge, MA.
7. Picard, R.W., Vyzas, E. (1999) Offline and Online Recognition of Emotion Expression from Physiological Data, *Emotion-Based Agent Architectures Workshop Notes, Int'l Conf. Autonomous Agents*, 135-142.
8. Raouzaiou, A., Tsapatsoulis, N., Karpouzis, K., Kollias, S. (2002). Parameterized facial expression synthesis based on MPEG-4, *Eurasip Journal on Applied Signal Processing*, 10, 1021-1038.
9. Tekalp, A.M., Ostermann, J., (2000). Face and 2-D Mesh Animation in MPEG-4, *Signal Processing: Image Communication*, 15, 387-421.
10. Young, J.W. (1993). Head and face anthropometry of adult U.S. civilians, *FAA Civil Aeromedical Institute*.
11. Williams, G.W. (1976). Comparing the joint agreement of several raters with another rater, *Biometrics*, 32, 619-627.
12. Chalana, V., Kim, Y. (1997). A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images, *IEEE Transactions on Medical Imaging*, 16, 5.
13. Pantic, M. Rothkrantz, L.J.M. (2000). Expert system for automatic analysis of facial expressions, *Image and Vision Computing*, 18, 881-905.
14. Pantic, M. Rothkrantz, L.J.M. (2000). Automatic Analysis of Facial Expressions: The State of the Art, Maja Pantic, Leon J.M. Rothkrantz, *IEEE Transactions on PAMI*, 22(12).
15. Wallace, M., Raouzaiou, A., Tsapatsoulis, N., Kollias, S. (2004). Facial Expression Classification based on MPEG-4 FAPs: The use of evidence and prior knowledge for uncertainty removal, *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Budabest, Hungary.
16. IST Project: *Emotionally Rich Man-Machine Interaction Systems* (ERMIS), 2001-2003. http://www.image.ntua.gr/ermis/
17. NoE: *Human-Machine Interaction Network on Emotion* (HUMAINE), 2004-2007. http://emotion-research.net/