Confidence-Based Fusion of Multiple Feature Cues for Facial Expression Recognition

Spiros Ioannou[1], Manolis Wallace[2], Kostas Karpouzis[1], Amaryllis Raouzaiou[1] and Stefanos Kollias[1]

[1]National Technical University of Athens,
9, Iroon Polytechniou Str., 157 80 Zographou, Athens, Greece
[2]University of Indianapolis, Athens Campus,
9, Ipitou Str., 105 57 Syntagma, Athens, Greece

***Abstract*— Since facial expressions are a key modality in human communication, the automated analysis of facial images for the estimation of the displayed expression is essential in the design of intuitive and accessible human computer interaction systems. In most existing rule-based expression recognition approaches, analysis is semi-automatic or requires high quality video. In this paper we propose a feature extraction system which combines analysis from multiple channels based on their confidence, to result in better facial feature boundary detection. The facial features are then used for expression estimation. The proposed approach has been implemented as an extension to an existing expression analysis system in the framework of the IST ERMIS project.**

***Index Terms*— Facial feature extraction, confidence, multiple cue fusion, human computer interaction**

I. 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" [15]. 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) [16].

Interpersonal communication is for the most part completed via the face. 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 [2]. This implies that the facial expressions form the major modality in human communication, and need to be considered by HCI/MMI systems.

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. When dealing with such input we have to accept that color quality and video resolution will be very poor. 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. Moreover it is very difficult to fit a precise model to each feature or to employ tracking since high-order frequency information is missing in such situations. A way to overcome this limitation is to combine 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 them produces different error patterns.

II. Expression Representation

An automated emotion recognition through facial expression analysis system, must deal mainly with two major research areas: automatic facial feature extraction and facial expression recognition. Thus, it needs to combine low-level image processing with the results of psychological studies about facial expression and emotion perception.

Most of the existing expression recognition systems can be classified in two major categories: the former includes techniques which examine the face in its entirety (holistic approaches) and take into account properties such as intensity [9] or optical flow distributions and the latter includes methods which operate locally, either by analyzing the motion of local features, or by separately recognizing, measuring, and combining the various facial element properties (analytic approaches). A good overview of the current state of the art is presented in [4][10].

In this work we estimate facial expression through the estimation of the MPEG FAPs. FAPs are measured through detection of movement and deformation of local intransient facial features such as mouth, eyes and eyebrows in single frames. Feature deformations are estimated by comparing their states to some frame, in which the person's expression is known to be neutral. Although FAPs [1] provide all the necessary elements for MPEG-4 compatible animation, we cannot use them directly for the analysis of expressions from

video scenes, due to the absence of a clear quantitative definition framework. In order to measure FAPs in real image sequences, we have to define a mapping between them and the movement of specific FDP feature points (FPs), which correspond to salient points on the human face.

III. Feature Extraction

An overview of the system is given in Figure 1. 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, to define the considered 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 are used, each resulting in a different mask. In total, we have four masks for each eye and three for the mouth. These masks have to be calculated in near-real time; 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 to detect regions of high texture based on standard deviation

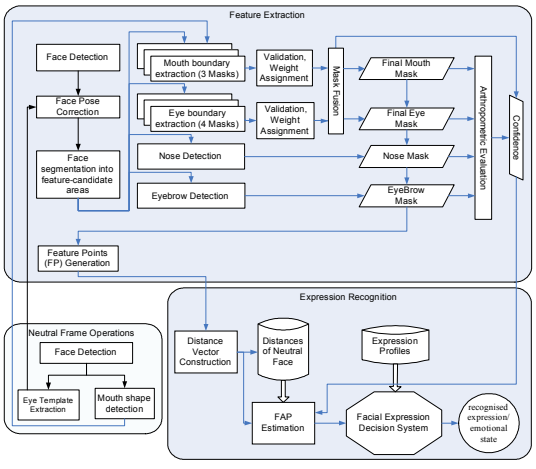


Figure 1: System Overview

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. A subset of the distances used to form the acceptance criteria of the eyes is shown in the following example:

d_1	Eye width
d_2	Distance of eye's middle vertical coordinate and eyebrow's middle vertical coordinate
d_3	Eyebrow width
d_4	D_{bp}, Bipupil breadth

$$M_{eye_1}^{c_1} = 1 - \left| 1 - \left(d_1 / d_4 \right) / 0.49 \right| \qquad (0.1)$$

and

$$M_{eye_1}^{c_2} = 1 - \left| d_2 \right| / d_3 \qquad (0.2)$$

where $M_{eye_1}^{c_1}$ and $M_{eye_1}^{c_2}$ are the confidence degrees acquired trough the application of each validation criterion on eye mask M_{eye_1}. The former of the two criteria is based on [7], where the mean ratio of eye width over bipupil breadth is reported as 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, 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, producing a final mask together with the corresponding estimated confidence [18] 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 that the combination error $y_{comb} - t$ from different machines f_i is guaranteed to be lower than the average error:

$$(y_{comb} - t)^2 = \frac{1}{M}\sum_i (y_i - t)^2 - \frac{1}{M}\sum_i (y_i - y_{comb})^2 \quad (0.3)$$

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.

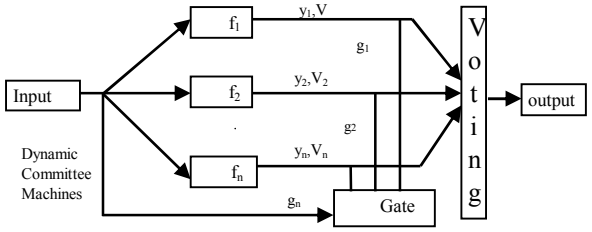


Figure 2: Dynamic Committee Machine Architecture

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:

$$m_f^x = \frac{1}{n}\sum_{i=1}^{n} m_i^x M_f^{cx_i} h^i g^i, \quad (0.4)$$

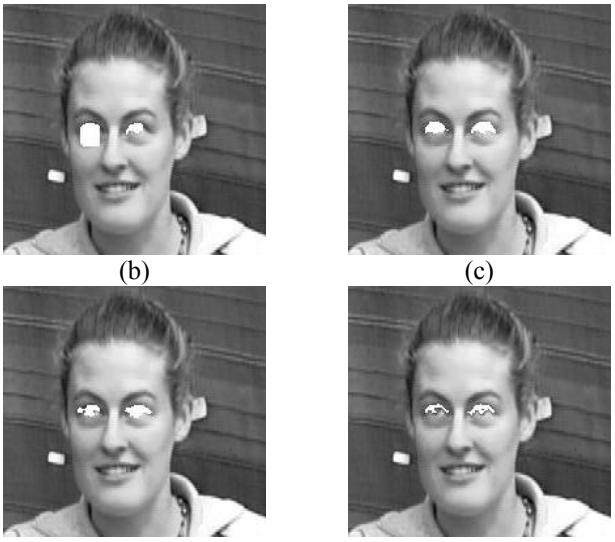
$$h^k = \begin{cases} 1, M_f^{c,x_k} \geq \left(t_{vd} \cdot \left\langle M_q^{c,x_k} \right\rangle_q\right) \\ 0, M_f^{c,x_k} < \left(t_{vd} \cdot \left\langle M_q^{c,x_k} \right\rangle_q\right) \end{cases} \quad (0.5)$$

Where m_i^x is the element of mask M_i^x, M_f^{c,x_i} the final 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; since non-synthetic training data for the latter is difficult to acquire, in our first implementation, the gating output of variable g^i is not trained but it is defined manually as follows:

$$g^i = \begin{cases} n, i = 1, D_{bp} > 128, \sigma_{cr} < t_\sigma, \sigma_{cb} < t_\sigma \\ 1/n, i \neq 1, D_{bp} > 128, \sigma_{cr} < t_\sigma, \sigma_{cb} < t_\sigma \\ 1, \text{otherwise} \end{cases} \quad (0.6)$$

where D_{bp} the bipupil width in pixels and σ_{cr}, σ_{cb} the standard deviation of the C_r, C_b channels respectively inside the facial area. It has been found that σ_{cr}, σ_{cb} in the same image is less than $5 \cdot 10^{-3}$ for good color quality and much larger for poor quality images.

(a)



(b) (c) (d) (e)

Figure 3. Original frame (a) and the four detected masks for the eyes in frame 3528 of the "Alyssa" sequence [7]

Figure 4. Final mask for the eyes

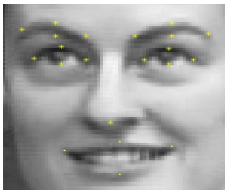


Figure 5. All detected feature points from the final masks

IV. Expression Analysis

The 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} *(squeeze_l_eyebrow)* is estimated as:

$$F_{37} = \left\| FP^n_{4.5} - FP^n_{3.11} \right\| - \left\| FP_{4.5} - FP_{3.11} \right\| \qquad (0.7)$$

where FP^n_i, 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.

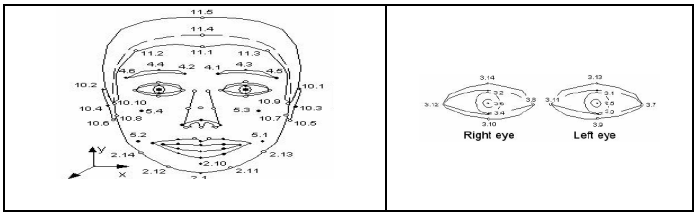


Figure 6. MPEG-4 Feature Points (FPs)

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

$$F^c_{37} = \min(FP^c_{4.5}, FP^c_{3.11}) \qquad (0.8)$$

On the other hand, some FAPs may be estimated in different ways. For example, FAP F_{31} is estimated as:

$$F^1_{31} = \left\| FP^n_{3.1} - FP^n_{3.3} \right\| - \left\| FP_{3.1} - FP_{3.3} \right\| \qquad (0.9)$$

or as

$$F^2_{31} = \left\| FP^n_{3.1} - FP^n_{9.1} \right\| - \left\| FP_{3.1} - FP_{9.1} \right\| \qquad (0.10)$$

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:

$$F_i = \frac{F^1_i + F^2_i}{2} \qquad (0.11)$$

The amount of uncertainty contained in each one of the distinct initial FAP calculations can be estimated by

$$E^1_i = 1 - F^{1c}_i \qquad (0.12)$$

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:

$$E_i = t(E^1_i, E^2_i) \qquad (0.13)$$

The Yager t-norm with parameter $w=5$ gives reasonable results for this operation:

$$E_i = 1 - \min\left(1, \left((1-E^1_i)^w + (1-E^2_i)^w\right)^w\right) \qquad (0.14)$$

The overall confidence value for the final estimation of the FAP is then acquired as

$$F^c_i = 1 - E_i \qquad (0.15)$$

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:

$$x_j^c = F_i^c \tag{0.16}$$

where F_i is the FAP based on which antecedent x_j is defined. More information about the used expression profiles can be found in [3][8].

V. Experimental Results

Facial feature extraction can be seen as a subcategory of image segmentation, i.e. image segmentation into facial features. Zhang [20] reviewed a number of simple discrepancy measures of which, if we consider image segmentation as a pixel classification process, only one is applicable here: the number of misclassified pixels on each facial feature. While manual feature extraction do not necessarily require expert annotation, it is clear in especially in low-resolution images manual labeling introduces an error. It is therefore desirable to obtain a number of manual interpretations in order to evaluate the inter-observer variability. A way to compensate for the latter is Williams' Index (WI) [6], 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 [19]. The modified Williams' Index 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:

$$WI = \frac{\frac{1}{n}\sum_{j=1}^{n}\frac{1}{D_{0,j}}}{\frac{2}{n(n-1)}\sum_{j}\sum_{j':j'>j}\frac{1}{D_{j,j'}}} \tag{0.17}$$

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

$$D_{j,j'} = \frac{1}{D_{bp}}\left\| M_j^x \veebar M_{j'}^x \right\| \tag{0.18}$$

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} (bibupil width) is used as a normalization factor to compensate for camera zoom on video sequences.

From a dataset of about 50000 frames, 250 frames were selected at random and were manually labeled from two observers. Distribution of WI is shown in **Figure 7**. At a value of 0, the computer mask is infinitely far from the observer mask. When the index is larger than 1, the computer generated mask disagrees less with the observers than the observers disagree with each other. TABLE 1 summarizes the results. For the eyes and mouth *WI* has been calculated for the both the final mask and each of the intermediate masks. WI_x denotes WI for single mask *x* and WI_f is the WI for the final mask for each facial feature; $\langle WI_x \rangle$ denotes the average *WI* for mask *x* calculated over all test frames. **Figure 7** illustrates the *WI* distribution on the test frames, calculated on each frame as the average *WI* of all the final feature masks.

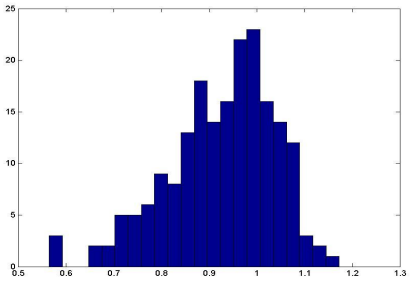


Figure 7
Williams Index distribution
(average on eyes and mouth)

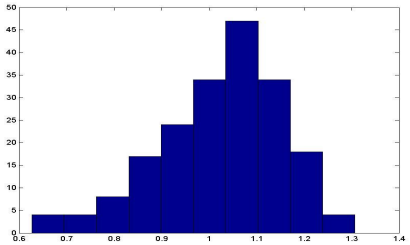


Figure 8
Williams Index distribution
(average on left and right eyebrows)

VI. Conclusions

Automatic recognition of FAPs is a difficult problem, and relatively little work has been reported [21]. Within the ERMIS [5] framework the majority of collected data have had the aforementioned quality problems; sometimes one has to compromise between quality and the use of intrusive equipment. In both the study of emotional cues and HCI video quality has to be sacrificed. The procedure we have described can exploit anthropometric knowledge [7] to evaluate a set of extracted features based on different techniques in order to improve overall performance. Early tests on both low and high quality video from the ERMIS database have been very promising: the algorithm can perform fully unattended FAP extraction and self-recovers in cases of false detections. The system runs currently in MATLAB and the performance is in the order of a few seconds per frame.

TABLE 1
RESULT SUMMARY

	Mask #	$\langle WI_x \rangle$	$\langle WI_f \rangle$	$\frac{\langle WI \rangle_f}{\langle WI \rangle_x}$	σ^2	% of frames where $WI_f > WI_x$	$\langle WI \rangle$ in frames where $WI_f < WI_x$	$\langle WI \rangle$ in frames where $WI_f > WI_x$
Left Eye								
NN[1]		0.6771		1.287	0.103	74.2	0.697	0.885
	1	0.7016		1.216	0.056	78.8	0.731	0.868
	2	0.8219	0.8388	1.029	0.027	82.4	0.770	0.887
	4	0.7416		1.131	0.057	76.2	0.811	0.847
	3	0.8708		0.979	0.026	44.3	0.812	0.867
Right Eye								
NN[1]		0.8008		1.093	0.020	75.2	0.672	0.946
	1	0.7185		1.243	0.084	81.4	0.674	0.929
	2	0.7740	0.8756	1.140	0.021	58.2	0.836	0.883
	3	0.6504		1.346	0.028	84.5	0.632	0.920
	4	0.8939		0.982	0.02	48.4	0.778	0.996
Mouth								
	1	0.7632		1.051	0.046	59.2	0.752	0.772
	2	0.8231	0.7803	0.963	0.038	44.8	0.721	0.852
	3	0.5703		1.446	0.204	96.9	0.510	0.793
Eyebrows								
left			1.0340					
right			1.0139					

WI_x denotes WI for single mask x and WI_f is the WI for the final mask for each facial feature.

[1]NN denotes the eye mask derived from the eye detection neural network output

REFERENCES

[1] A. M. Tekalp, J. Ostermann, "Face and 2-D Mesh Animation in MPEG-4", Signal Processing: Image Communication, Vol. 15, pp. 387-421, 2000.

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

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

[4] B. Fasel, et al, "Automatic Facial Expression Analysis: A Survey", Pattern Recognition, 36, pp 259-275, 2003

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

[6] G. W. Williams, Comparing the joint agreement of several raters with another rater", Biometrics, vol32, pp. 619-627, 1976

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

[8] K. Karpouzis, A. Raouzaiou, A. Drosopoulos, S. Ioannou, T. Balomenos, N. Tsapatsoulis and S. Kollias. "Facial expression and gesture analysis for emotionally-rich man-machine interaction" N. Sarris,

[9] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In Proc. of Computer Vision and Pattern Recognition, pages 586-591. IEEE, June 1991b.

[10] M. H. Yang, D. Kriegman, N. Ahuja, "Detecting Faces in Images:A Survey", PAMI, Vol.24(1), pp. 34-58, 2002.

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

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias , W. Fellenz, J. Taylor. Emotion Recognition in Human-Computer Interaction, IEEE Signal Processing Magazine, 2001, pp. 32-80.

[13] R. Fransens, Jan De Prins, SVM-based Nonparametric Discriminant Analysis, An Application to Face Detection, Ninth IEEE International Conference on Computer Vision Volume 2, October 13 - 16, 2003

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

[15] R.W. Picard, Affective Computing, MIT Press, Cambridge, MA.

[16] R.W. Picard,, Vyzas E., Offline and Online Recognition of Emotion Expression from Physiological Data, Emotion-Based Agent Architectures Workshop Notes, Int'l Conf. Autonomous Agents, pp. 135-142, 1999.

[17] S.Ioannou, A. Raouzaiou, K. Karpouzis, M. Pertselakis, N. Tsapatsoulis, S.Kollias save Adaptive Rule-Based Facial Expression Recognition G. Vouros, T. Panayiotopoulos (Eds.), Lecture Notes in Artificial Intelligence, Vol. 3025, Springer-Verlag, pp. 466 - 475, 2004.

[18] T.G. Dietterich, Ensemble methods in machine learning, Proceedings of First International Conference on Multiple Classifier Systems, 2000.

[19] Vikram Chalana and Yongmin Kim, A Methodology for Evaluation of Boundary Detection Algorithms on Medical Images, IEEE Transactions on Medical Imaging, Vol.16, No.5 October 1997

[20] Y.J.Zhang, A Survey on Evaluation Methods for Image Segmentation, Pattern Recognition, Vol 29, No. 8, pp1334-1346, 1996

[21] 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