COMBINATION OF MULTIPLE EXTRACTION ALGORITHMS IN THE DETECTION OF FACIAL FEATURES

Spiros Ioannou, Manolis Wallace, Kostas Karpouzis, Amaryllis Raouzaiou and Stefanos Kollias*

National Technical University of Athens, 9, Iroon Polytechniou Str., 157 80 Zographou, Athens, Greece
*also with University of Indianapolis, Athens Campus, 9, Ipitou Str., 105 57 Syntagma, Athens, Greece
e-mail: {sivann,wallace, araouz}@image.ntua.gr, kkarpou@softlab.ntua.gr, stefanos@cs.ntua.gr

ABSTRACT

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 existing rule-based expression recognition approaches, different feature extraction techniques have been tested that allow for the automatic detection of feature points, providing the required input for a rule based expression analysis; each one of these techniques outperforms others under specific constraints. In this paper we propose a feature extraction system which combines analysis from multiple channels based on their confidence, to result in better, error resilient facial feature boundary detection. The proposed approach has been implemented as an extension to an existing expression analysis system in the framework of the IST ERMIS project.

1. INTRODUCTION

Despite common belief, social psychology research has shown that conversations are usually dominated by facial expressions, indicating the speaker's predisposition towards the listener. For example, Mehrabian indicated that the linguistic part of a message contributes only for seven percent to the effect of the message as a whole; the paralinguistic part, 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 facial expressions form a 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 resolution. A 4:2:0 video signal (eg. H-261, MPEG-2 where C_r and C_b are each subsampled by a factor of 2 both horizontally and vertically) is still considered to be a very good quality signal; moreover, the face usually 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 usually feasible to detect the presence and location of face and all facial features with high accuracy, it is very difficult in such conditions to find the exact boundary of each one (eye, eyebrow, mouth) in order to estimate its deformation from a neutral-expression frame [7].

To accommodate for such problems, in this work we propose a new facial feature extraction method which relies on the fusion of several facial feature masks derived from multiple feature extractors. 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. For each feature, extracted feature masks are fused together by a dynamic committee machine which uses their evaluation to calculate weights; input image quality in the form of resolution and color quality are used to estimate the gating variables. The resulting enhanced accuracy of the extracted features naturally leads to enhanced results of the overall process of facial expression analysis as well.

2. FEATURE EXTRACTION

An overview of the system is given in Figure 1. At first, face detection is performed using nonparametric discriminant analysis with a Support Vector Machine which classifies face and non-face areas by reducing the training problem dimension to a fraction of the original with negligible loss of classification performance [6]. This step provides us with a rectangle head boundary which includes the whole face area. The latter is segmented roughly using static anthropometric rules into three overlapping rectangle regions of interest which include both facial features and facial background [1]; these three feature-candidate areas include the left eye/eyebrow, the right eye/eyebrow and the mouth. We utilize these areas to initialize the feature extraction process.

Facial feature extraction performance depends on head pose, thus head pose needs to be detected and the head restored in the upright position; in this work we are mainly concerned with roll rotation, since it is the most frequent rotation encountered in real life video sequences.

To estimate the head pose we first locate the left and right eyes in the corresponding eye candidate areas and estimate head roll rotation by calculating the angle between the horizontal plane and the line defined by the eye centers.

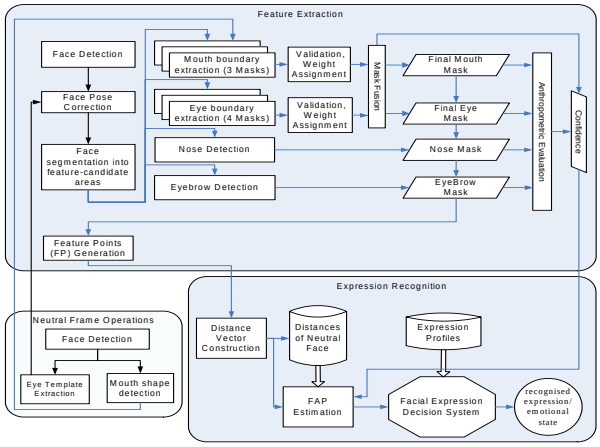

Figure 1: System Overview

For eye localization we use a feed-forward back propagation neural network with a sigmoidal activation function. The multi-layer perceptron (MLP) we adopt employs Marquardt-Levenberg learning [8], while the optimal architecture obtained through pruning has two 20 node hidden layers for 13 inputs. The network is applied separately on the left and right eye-candidate face regions; for each pixel in those regions the 13 NN inputs are the luminance Y, the Cr & Cb chrominance values and the 10 most important DCT coefficients (with zigzag selection) of the neighboring 8x8 pixel area. Using additional input color spaces, such as Lab, RGB or HSV to train the network, has not increased its distinction efficiency. The MLP has two outputs, one for each class, namely eye and non-eye, and it has been trained with more than 100 hand-made eye masks that depict eye and non-eye area in random frames from the ERMIS database [5], in images of diverse quality, resolution and lighting conditions. The network's output for facial images outside the training set is good for locating the eye; however it cannot provide accurate information near the eye boundaries.

The output of the aforementioned eye localization process is used as input in order to create facial feature 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 considered feature points (FPs). 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 more problematic, a variety of methods is 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 thus he had to avoid utilizing complex or time-consuming feature extractors. The feature extractors developed for this work are briefly described in the following.

2.1. Mask Extraction

Eyebrows are detected with a procedure involving morphological edge detection and feature selection using data from [1]. Nose detection is based on nostril localization. Nostrils are easy to detect due to their low intensity [9]. Connected objects (i.e. nostril candidates) are labeled based on their vertical proximity to the left or right eye, and the best pair is selected according to its position, luminance and geometrical constraints from [1]. For the eyes the following masks are constructed:

• A refined version of the original neural-network derived mask. The initial eye mask is extended by using an adaptive low-luminance threshold on an area defined from the neural network high-confidence output. This mask includes the top and bottom eyelids in their full extent that are usually missing from the initial mask. (Figure 3e)

• A mask expanding in the area between the upper and lower eyelids. Since the eye-center is almost always detected correctly from the neural network, the horizontal edges of the eyelids in the eye area are used to limit the eye mask in the vertical direction. A modified Canny edge operator is used due to its property of providing good localization. The operator is limited to ignore movements in the most vertical directions. (Figure 3b)

• A region-growing technique that takes advantage from the fact that texture complexity inside the eye is higher compared to the rest of the face. This process consists of thresholding the iteratively reduced grayscale eye image with its 3x3 standard deviation map, while the resulting binary eye mask center remains close to the original. This process is found to perform very well for images of very-low resolution and low color quality. (Figure 3c)

• A mask computed using the normal probability of luminance using a simple adaptive threshold on the eye area. This mask includes the darkest areas of the eye area which usually include the sclera and eyelashes but can extend outside the eye area when illumination is not uniform, thus it is cut vertically at its thinnest points from both sides of the eye centre and the convex hull of the result is used. (Figure 3d)

Finding the extent of a closed mouth in a still image is a relatively easy accomplished task [10]. In case of an open mouth, several methods have been proposed which make use of intensity [11] or color information [12]. In this work, we propose three different approaches that are then

fused in order to produce the final mask:

- An MLP neural network is trained to identify the mouth region using the neutral image. The network has similar architecture as the one used for the eyes. The train data are acquired from the neutral image (where the mouth is closed) as follows: the mouth-candidate ROI is first filtered with Alternating Sequential Filtering by Reconstruction (ASFR) to simplify and create connected areas of similar luminance. Simple but effective luminance thresholding is then used to find the area between the lips in the neutral image where the mouth is closed. This area is dilated vertically and the data depicted by this area are used to train the network.
- A horizontal morphological gradient is calculated in the mouth area and the longest connected object which comply with constrains from [6] and the nose position is selected as a possible mouth mask
- This final approach takes advantage of the relative low luminance of the lip corners and contributes to the correct identification of horizontal mouth extent which is not always detected by the previous methods in cases of smiling and apparent teeth. A short summary of the procedure is as follows: The image is simplified and thresholded and connected objects are labeled. Two cases are examined separately: either we have no apparent teeth and the mouth area is denoted by a cohesive dark area or there are teeth and thus two dark areas appear at both sides of the teeth. In the first case mouth extend is straightforward to detect; in the latter mouth centre proximity of each object is assessed through [6] and the appropriate objects are selected. The convex hull of the result is then merged through morphological reconstruction with an horizontal edge map to include the upper and bottom lips. The result is the third mouth mask.

2.2. Mask Fusion

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. 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.

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.

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 feature extractors briefly described above are merely the ones developed for this specific work. 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 [4]. 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 [3]:

$$(y_{comb} - t)^2 = \frac{1}{M}\sum_i (y_i - t)^2 - \frac{1}{M}\sum_i (y_i - y_{comb})^2 \quad (0.1)$$

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.

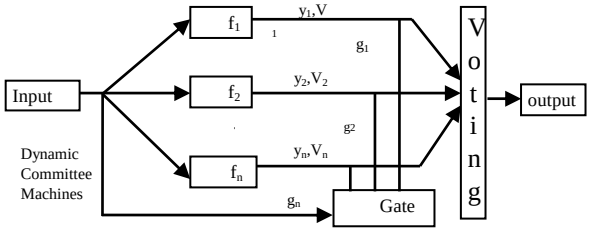

Figure 2: Dynamic Committee Machine Architecture

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 $\mathbf{M}_f^{\mathrm{x}}$ is calculated from the n masks as:

$$m_f^x = \frac{1}{n}\sum_{i=1}^{n} m_i^x M_f^{cx_i} h^i g^i \qquad (0.2)$$

$$h^k = \begin{cases} 1, M_f^{c,x_k} \geq \left(\mathrm{t_{vd}} \cdot \left\langle M_q^{c,x_k} \right\rangle_{\mathrm{q}} \right) \\ 0, M_f^{c,x_k} < \left(\mathrm{t_{vd}} \cdot \left\langle M_q^{c,x_k} \right\rangle_{\mathrm{q}} \right) \end{cases} \qquad (0.3)$$

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 $M_f^{c,x_k} < \left(\mathrm{t_{vd}} \cdot \left\langle M_q^{c,x_k} \right\rangle_{\mathrm{q}} \right)$ to contribute to the final mask. A sufficient value for $\mathrm{t_{vd}}$ is 0.8.

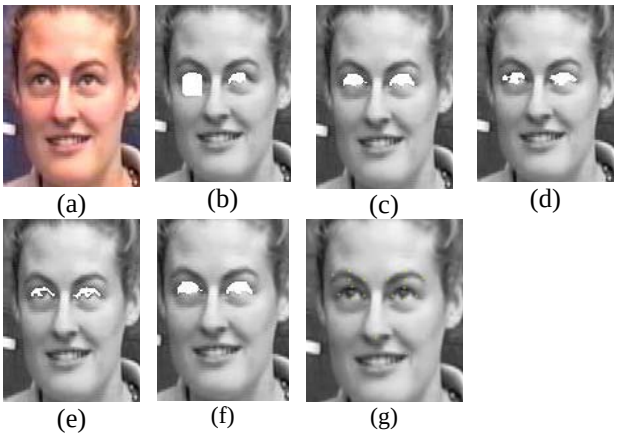

Figure 3. (a):original frame, (b),(c),(d),(e): the four detected masks, (f):final mask for the eyes, (g):all detected feature points from the final masks

The role of the gating variable g^i is to favor the color-based feature extraction methods ($\mathbf{M}_1^{\mathrm{e}}$, $\mathbf{M}_1^{\mathrm{m}}$) 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 [13].

3. Results and Concluding Remarks

In Figure 3 we briefly present some indicative results from the application of the proposed methodology to facial feature extraction. Figure 3(a) is the region specified by the face detection step. We can see in Figures 3(b,c,d,e) that not all eye detection approaches perform equally well for the frame in question. The mask fusion committee machine, taking advantage from the information available in the evaluation of the masks, provides the overall masks of Figure 3(f). Figure 3(g) presents the FPs extracted from the specific frame, when all masks and facial features are considered.

It is worth noting that although in this example the overall mask is very similar to that of Figure 3(c), other extraction approaches often perform better in other frames. This validates our approach to dynamically estimate the overall mask using the proposed methodology. The result is that we can produce a system that is able to provide better feature extraction results, with higher confidence, and with great resilience to errors occurring in some of the considered extraction methodologies. Early tests within the framework ERMIS [5] on both low and high quality video from the ERMIS database have been very promising: the algorithm can perform fully unattended feature extraction and overcomes errors occurring in individual modules.

References

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

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

[3] 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.

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

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

[6] 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

[7] 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

[8] S. Kollias and D. Anastassiou. "An adaptive least squares algorithm for the efficient training of artificial neural networks". IEEE Transactions on Circuits and Systems, Volume: 36 , Issue: 8 , Aug. 1989 Pages:1092 – 1101

[9] Dmitry O. Gorodnichy. On Importance of Nose for Face Tracking, Proc. Intern. Conf. on Automatic Face and Gesture Recognition (FG'2002), Washington DC, May 20-21, 2002.

[10] Hagan, M. T., and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, vol. 5, no. 6, pp. 989-993, 1994.

[11] Lijun Yin, Generating Realistic Facial Expressions with Wrinkles for Model-Based Coding, Computer Vision and Image Understanding 84, 201-240 (2001)

[12] Leung et al: Lip image segmentation using fuzzy clustering incorporating an alliptic shape function, IEEE Trans. on image processing, vol.13, No.1, January 2004

[13] 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.