Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2007, Article ID 29081, 22 pages
doi:10.1155/2007/29081

Research Article

Robust Feature Detection for Facial Expression Recognition

Spiros Ioannou, George Caridakis, Kostas Karpouzis, and Stefanos Kollias

*Image, Video and Multimedia Systems Laboratory, National Technical University of Athens,
9 Iroon Polytechniou Street, 157 80 Zographou, Athens, Greece*

Received 1 May 2006; Revised 27 September 2006; Accepted 18 May 2007

Recommended by Jörn Ostermann

This paper presents a robust and adaptable facial feature extraction system used for facial expression recognition in human-computer interaction (HCI) environments. Such environments are usually uncontrolled in terms of lighting and color quality, as well as human expressivity and movement; as a result, using a single feature extraction technique may fail in some parts of a video sequence, while performing well in others. The proposed system is based on a multicue feature extraction and fusion technique, which provides MPEG-4-compatible features assorted with a confidence measure. This confidence measure is used to pinpoint cases where detection of individual features may be wrong and reduce their contribution to the training phase or their importance in deducing the observed facial expression, while the fusion process ensures that the final result regarding the features will be based on the extraction technique that performed better given the particular lighting or color conditions. Real data and results are presented, involving both extreme and intermediate expression/emotional states, obtained within the sensitive artificial listener HCI environment that was generated in the framework of related European projects.

1. INTRODUCTION

Facial expression analysis and emotion recognition, a research topic traditionally reserved for psychologists, has gained much attention by the engineering community in the last twenty years. Recently, 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.” The reasons include the need for quantitative facial expression description [1] as well as automation of the analysis process [2] which is strongly related to ones’ emotional and cognitive state [3].

Automatic estimation of facial model parameters is a difficult problem and although a lot of work has been done on selection and tracking of features [4], relatively little work has been reported [5] on the necessary initialization step of tracking algorithms, which is required in the context of facial feature extraction and expression recognition. Most facial expression recognition systems use the facial action coding system (FACS) model introduced by Ekman and Friesen [3] for describing facial expressions. FACS describes expressions using 44 action units (AU) which relate to the contractions of specific facial muscles. In addition to FACS, MPEG-4 metrics [6] are commonly used to model facial expressions and underlying emotions. They define an alternative way of modeling facial expressions and the underlying emotions, which is strongly influenced by neurophysiologic and psychological studies. MPEG-4, mainly focusing on facial expression synthesis and animation, defines the facial animation parameters (FAPs) that are strongly related to the action units (AUs), the core of the FACS. A comparison and mapping between FAPs and AUs can be found in [7].

Most facial expression recognition systems attempt to map facial expressions directly into archetypal emotion categories while been unable to handle expressions caused by intermediate or nonemotional expressions. Recently, several automatic facial expression analysis systems that can also distinguish facial expression intensities have been proposed [8–11], but only a few are able to employ model-based analysis using the FAP or FACS framework [5, 12]. Most existing approaches in facial feature extraction are either designed to cope with limited diversity of video characteristics or require manual initialization or intervention. Specifically, [5] depends on optical flow, [13–17] depend on high resolution or noise-free input video, [18–20] depend on color information, [15, 21] require manual labeling or initialization, [12] requires markers, [14, 22] require manual selection of feature points on the first frame, [23] requires two head-mounted cameras, [24–27] require per-user or per-expression training either on the expression recognition or the feature extraction or cope only with fundamental emotions. From the above, [8, 13, 21, 23, 25, 27] provide success results solely on expression recognition and not on the feature extraction/recognition. Additionally very few approaches can perform in near real time.

Fast methodologies for face and feature localization in image sequences are usually based on calculation of the skin color probability. This is usually accomplished by calculating the a posteriori probability of a pixel belonging to the skin class in the joint Cb/Cr domain. Several other color spaces have also been proposed which exploit specific color characteristics of various facial features [28]. Video systems, on the other hand, convey image data in the form of one component that represents lightness (luma) and two components that represent color (chroma), disregarding lightness. Such schemes exploit the poor color acuity of human vision: as long as luma is conveyed with full detail, detail in the chroma components can be reduced by subsampling (filtering or averaging). Unfortunately, nearly all video media have reduced vertical and horizontal color resolutions. A 4 : 2 : 0 video signal (e.g., H-261, MPEG-2 where each of Cr and Cb are subsampled by a factor of 2 both horizontally and vertically) is still considered to be a very good quality signal. The perceived video quality is good indeed, but if the luminance resolution is low enough—or the face occupies only a small percentage of the whole frame—it is not rare that entire facial features share the same chrominance information, thus rendering color information very crude for facial feature analysis. In addition to this, overexposure in the facial area is common due to the high reflectivity of the face and color alteration is almost inevitable when transcoding between different video formats, rendering Cb/Cr inconsistent and not constant. Its exploitation is therefore problematic in many real-life video sequences; techniques like the one in [29] have been proposed in this direction but no significant improvement has been observed.

In the framework of the European Information Technology projects, ERMIS [30] and HUMAINE [31], a large audiovisual database was constructed which consists of people driven to emotional discourse by experts. The subjects participating in this experiment were not faking their expressions and the largest part of the material is governed by subtle emotions which are very difficult to detect even for human experts, especially if one disregards the audio signal.

The aim of our work is to implement a system capable of analyzing nonextreme facial expressions. The approach has been tested in a real human-computer interaction framework, using the SALAS (sensitive artificial listener) testbed [30, 31], which is briefly described in the paper. The system should be able to evaluate expressions even when the latter are not extreme and should be able to handle input from various speakers. To overcome the variability in terms of luminance and color resolution in our material, an analytic approach that allows quantitative and rule-based expression profiling and classification was developed. Facial expression is estimated through analysis of MPEG FAPs [32], the latter being measured through detection of movement and deformation of local intransient facial features such as mouth, eyes, and eyebrows through time, assuming availability of a person's neutral expression. The proposed approach is capable of detecting both basic and intermediate expressions (e.g., boredom, anger) [7] with corresponding intensity and confidence levels.

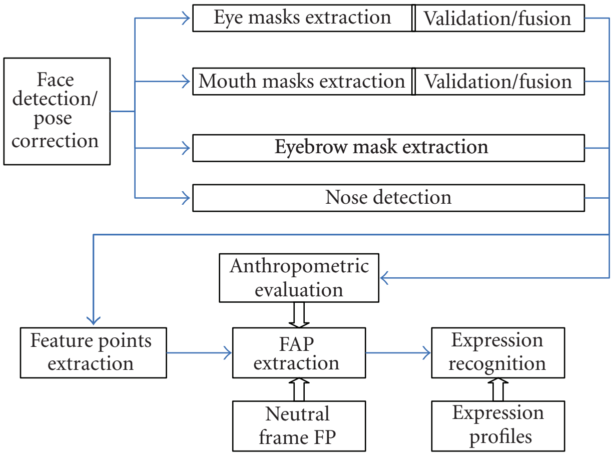


Figure 1: Diagram of the proposed methodology.

An overview of the proposed expression and feature extraction methodologies is given in Section 2 of the paper. Section 3 describes face detection and pose estimation while Section 4 provides detailed analysis of automatic facial feature boundary extraction and construction of multiple masks for handling different input signal variations. Section 5 describes the multiple mask fusion process and confidence generation. Section 6 focuses on facial expression/emotional analysis, and presents the SALAS human-computer interaction framework while Section 7 presents the obtained experimental results. Section 8 draws conclusions and discusses future work.

2. AN OVERVIEW OF THE PROPOSED APPROACH

An overview of the proposed methodology is illustrated in Figure 1. The face is first located, so that approximate facial feature locations can be estimated from the head position and rotation. Face roll rotation is estimated and corrected and the head is segmented focusing on the following facial areas: left eye/eyebrow, right eye/eyebrow, nose, and mouth. Each of those areas, called feature-candidate areas, contains the features whose boundaries need to be extracted for our purposes. Inside the corresponding feature-candidate areas precise feature extraction is performed for each facial feature, that is, eyes, eyebrows, mouth, and nose, using a multicue approach, generating a small number of intermediate feature masks. Feature masks generated for each facial feature are fused together to produce the final mask for that feature. The mask fusion process uses anthropometric criteria [33] to perform validation and weight assignment on each intermediate mask; each feature's weighted masks are then fused to produce a final mask along with confidence level estimation.

Measurement of facial animation parameters (FAPs) requires the availability of a frame where the subject's expression is found to be neutral. This frame will be called the *neutral frame* and is manually selected from video sequences to be analyzed or interactively provided to the system when initially brought into a specific user's ownership. The final feature masks are used to extract 19 feature points (FPs) [7]. Feature points obtained from each frame are compared to FPs obtained from the neutral frame to estimate facial deformations and produce the facial animation parameters (FAPs). Confidence levels on FAP estimation are derived from the equivalent feature point confidence levels. The FAPs are used along with their confidence levels to provide the facial expression estimation.

3. FACE DETECTION AND POSE ESTIMATION

In the proposed approach, facial features including eyebrows, eyes, mouth, and nose are first detected and localized. Thus, a first processing step of face detection and pose estimation is carried out, as described below, to be followed by the actual facial feature extraction process described in Section 4. At this stage, it is assumed that an image of the user at neutral expression is available, either a priori or captured before interaction with the proposed system starts.

The goal of face detection is to determine whether or not there are faces in the image, and if yes, return the image location and extent of each face [34]. Face detection can be performed with a variety of methods [35–37]. In this paper, we used nonparametric discriminant analysis with a *support vector machine* (SVM) which classifies face and nonface areas reducing the training problem dimension to a fraction of the original with negligible loss of classification performance [30, 38].

800 face examples from the NIST Special Database 18 were used for this purpose. All examples were aligned with respect to the coordinates of the eyes and mouth and rescaled to the required size. This set was virtually extended by applying small scale, translation, and rotation perturbations and the final training set consisted of 16 695 examples.

The face detection step provides a rectangle head boundary which includes all facial features as shown in Figure 2. The latter can be then segmented roughly using static anthropometric rules (Figure 2, Table 1) into three overlapping rectangle regions of interest which include both facial features and facial background; these three *feature-candidate areas* include the left eye/eyebrow, the right eye/eyebrow, and the mouth. In the following, we utilize these areas to initialize the feature extraction process. Scaling does not affect feature-candidate area detection, since the latter is proportional to the head boundary extent, extracted by the face detector.

The accuracy of feature extraction depends on head pose. In this paper, we are mainly concerned with roll rotation, since it is the most frequent rotation encountered in real-life video sequences. Small head yaw and pitch rotations which do not lead to feature occlusion do not have a significant impact on facial expression recognition. The face detection techniques described in the former section is able to cope with head roll rotations up to 30°. This is a quite satisfactory range in which the feature-candidate areas are large enough so that the eyes reside in the eye-candidate search areas defined by the initial segmentation of a rotated face.

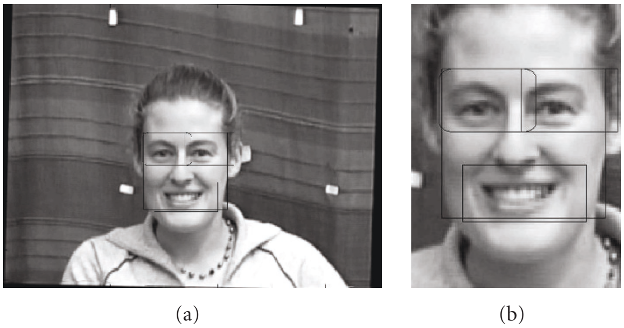


Figure 2: Feature-candidate areas: (a) full frame (352 × 288), (b) Zoomed (90 × 125).

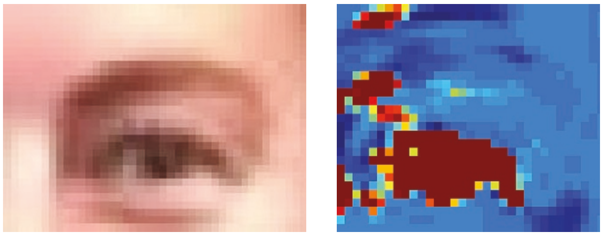
Table 1: Anthropometric rules for feature-candidate facial areas. W_f, H_f represent face width and face height, respectively.

Area	Location	Width	Height
Eyes and eyebrows	Top left and right parts of the face	$0.6W_f$	$0.5H_f$
Nose and mouth	Bottom part of the face	W_f	$0.5H_f$

To estimate the head pose, we first locate the left and right eyes in the detected corresponding eye candidate areas. After locating the eyes, we can estimate head roll rotation by calculating the angle between the horizontal plane and the line defined by the eye centers. For eye localization, we propose an efficient technique using a feed-forward backpropagation neural network with a sigmoidal activation function. The multilayer perceptron (MLP) we adopted employs Marquardt-Levenberg learning [39, 40] while the optimal architecture obtained through pruning has two 20-node hidden layers and 13 inputs. We apply the network separately on the left and right eye-candidate face regions. For each pixel in these 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 8 × 8 pixel area. Using alternative input color spaces such as Lab, RGB or HSV to train the network has not changed its distinction efficiency. The MLP has two outputs, one for each class, namely, eye and noneye, and it has been trained with more than 100 hand-made eye masks that depict eye and noneye area in random frames from the ERMIS [30] database, in images of diverse quality, resolution, and lighting conditions.

The network's output in randomly selected facial images outside the training set is good for locating the eye, as shown in Figure 3(b). However, it cannot provide exact outliers, that is, point locations at the eye boundaries; estimation of *feature points (FP)* is further analyzed in the next section.

To increase speed and reduce memory requirements, the eyes are not detected on every frame using the neural network. Instead, after the eyes are located in the first frame, two square grayscale eye templates are created, containing each of



(a) (b)

Figure 3: (a) Left eye input image (b) network output on left eye, darker pixels correspond to higher output.

the eyes and a small area around them. The size of the templates is half the eye-center distance (bipupil breadth, D_{bp}). For the following frames, the eyes are located inside the two eye-candidate areas, using template matching which is performed by finding the location where the sum of absolute differences (SAD) is minimized.

After head pose is computed, the head is rotated to an upright position and new feature-candidate segmentation is performed on the head using the same rules shown in Table 1, so as to ensure facial features reside inside their respective candidate regions. These regions containing the facial features are used as input for the facial feature extraction stage, described in the following section.

4. AUTOMATIC FACIAL FEATURE DETECTION AND BOUNDARY EXTRACTION

To be able to compute MPEG-4 FAPs, precise feature boundaries for the eyes, eyebrows, and mouth have to be extracted. Eye boundary detection is usually performed by detecting the special color characteristics of the eye area [28], by using luminance projections, reverse skin probabilities, or eye model fitting [17, 41]. Mouth boundary detection in the case of a closed mouth is a relatively easily accomplished task [40]. In case of an open mouth, several methods have been proposed which make use of intensity [17, 41] or color information [18, 28, 42, 43]. Color estimation is very sensitive to environmental conditions, such as lighting or capturing camera's characteristics and precision. Model fitting usually depends on ellipse or circle fitting, using Hough-like voting or corner detection [44]. Those techniques while providing accurate results in high-resolution images are unable to perform well with low video resolution which lack high-frequency properties; such properties which are essential for efficient corner detection and feature border trackability [4] are usually lost due to analogue video media transcoding or low-quality digital video compression.

In this work, nose detection and eyebrow mask extraction are performed in a single stage, while for eyes and mouth which are more difficult to handle, multiple (four in our case) masks are created taking advantage of our knowledge about different properties of the feature area; the latter are then combined to provide the final estimates as shown in Figure 1. Tables 2 and 5 summarize extracted eye and mouth mask notation, respectively, while providing a short qualitative description. In the following, we use the notation $\mathbf{M}^x_k$ to denote the binary mask k of facial feature x, where x is e for eyes, m for mouth, n for nose, and b for eyebrows, and $\mathbf{L}^x$ denotes the respective luminance masks. Additionally, feature size and position validation depends on several relaxed anthropometric constraints; these include t^m_{asf}, t^e_c, t^b_1, t^b_2, t^m_{b1}, t^m_{c2}, t^m_{b2}, t^n_2, t^n_3, t^n_4 defined in Table 3, while other thresholds defined in text are summarized in Table 4.

4.1. *Eye boundary detection*

4.1.1. *Luminance and color information fusion mask*

This step tries to refine eye boundaries extracted by the neural network described in Section 3 and denoted as $(\mathbf{M}^e_{\text{nn}})$, building on the fact that eyelids usually appear darker than skin due to eyelashes and are almost always adjacent to the iris.

At first, luminance information inside the area depicted by a dilated version of $\mathbf{M}^e_{\text{nn}}$ is used to find a luminance threshold t^e_b:

$$t^e_b = \frac{1}{3}\left(2\left\langle f_c\left(\mathbf{L}^e, \mathbf{M}^e_{\text{nn}}\right)\right\rangle + \min\left(\mathbf{L}^e\right)\right), \qquad (1)$$

$$f_c(\mathbf{A}, \mathbf{B}) = \{c_{ij}\}, \quad c_{ij} = \begin{cases} a_{ij}, & b_{ij} \neq 0, \\ 0, & b_{ij} = 0, \end{cases} \qquad (2)$$

where $\mathbf{L}^e$ is the luminance channel of the eye-candidate area and $\langle \bullet \rangle$ denotes the average over an image area, and $\min(X)$ denotes the minimum value of area X.

When threshold t^e_b is applied to $\mathbf{L}^e$, a new mask is derived, denoted as $\mathbf{M}^e_{\text{npp}}$. This map includes dark objects near the eye centre, namely, the eyelashes and the iris. From the connected components in $\mathbf{M}^e_{\text{npp}}$ we can robustly locate the one including the iris by estimating its thickness. In particular, we apply a distance transform using the Euclidean distance metric and select the connected component where distance transform obtains its maximum value $\text{DT}_{\max}$, to produce $\mathbf{M}^e_1$ mask as illustrated in Figure 4. The latter includes the iris and adjacent eyelashes. The point where the distance transform equals to $\text{DT}_{\max}$ accurately computes the iris centre.

4.1.2. *Edge-based mask*

This second approach is based on eyelid detection. Eyelids reside above and below the eye centre, which has already been estimated by the neural network. Taking advantage of their mainly horizontal orientation, eyelids are easily located through edge detection.

We use the canny edge detector [45] mainly because of its good localization performance and its ability to minimize multiple responses to a single edge. Since the canny operator follows local maxima, it usually produces closed curves. Those curves are broken apart into horizontal parts by morphological opening using a 3 × 1 structuring element; let us denote the result as $\mathbf{M}^e_{b_1}$. Since morphological opening can break edge continuity, we enrich this edge mask by performing edge detection, using a modified canny edge detector. The

Table 2: Summary of eye masks.

Described in	Detects	Depends on	Results
Section 4.1.1	Iris and surrounding dark areas including eyelashes	$\mathbf{L}^e, \mathbf{M}^e_{\text{nn}}$	$\mathbf{M}^e_1$
Section 4.1.2	Horizontal edges produced by eyelids, residing above and below eye centre	$\mathbf{L}^e$, eye centre	$\mathbf{M}^e_2$
Section 4.1.3	Areas of high texture around the iris	$\mathbf{L}^e$	$\mathbf{M}^e_3$
Section 4.1.4	Area with similar luminance to eye area defined by mask $\mathbf{M}^e_{\text{nn}}$	$\mathbf{L}^e, \mathbf{M}^e_{\text{nn}}$	$\mathbf{M}^e_4$

Table 3: Relational anthropometric constraints.

Variable	Value	Refers to
t^m_{asf}	1%	W_f
t^e_c	5%	W_f
t^b_2	5%	D_{bp}
t^n_2	10%	D_{bp}
t^m_{b1}	10%	I_w
t^n_4	15%	D_{bp}
t^m_{c2}	25%	D_{bp}
t^n_3	20%	D_{bp}
t^b_1	30%	D_{bp}
t^m_{b2}	50%	I_w

Table 4: Adaptive thresholds.

Variable	Value	Refers to
t^e_b	$\frac{1}{3}(2\langle f_c(\mathbf{L}_e, \mathbf{M}^e_{\text{nn}})\rangle + \min(\mathbf{L}_e))$	L
t^b_E	$\langle \mathbf{M}^b_{E_1} \rangle + \sqrt{\langle (\mathbf{M}^b_{E_1})^2 \rangle - \langle \mathbf{M}^b_{E_1} \rangle^2}$	L
t^m_{c1}	$\langle \mathbf{L}^{\text{asfr}}_m \rangle - \sqrt{\langle [\mathbf{L}^{\text{asfr}}_m]^2 \rangle - \langle \mathbf{L}^{\text{asfr}}_m \rangle^2}$	L
t^m_1	$\frac{1}{3}(2\overline{\mathbf{L}_m} + \min(\mathbf{L}_m))$	L
t^n_1	$\frac{1}{3}(\overline{\mathbf{L}_n} + 2\min(\mathbf{L}_n))$	L
t^m_2	90%	NN output
t^e_d	90%	L
	Thresholds	
Variable	Value	
t_σ	10^{-3}	
t_r	128	
t_{vd}	0.8	

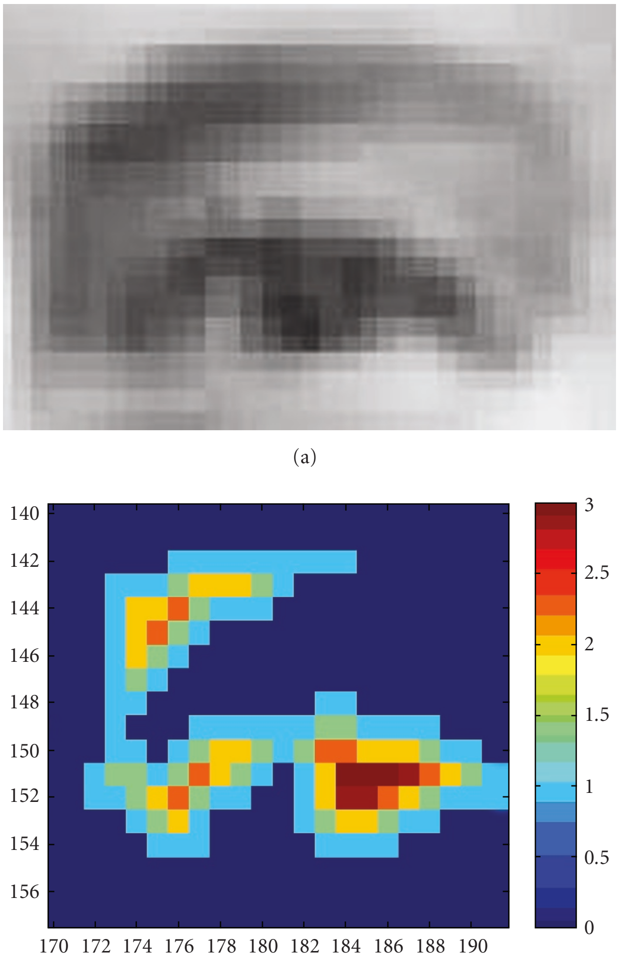
L_x: Luminance image of feature x.

latter looks for gradient continuity only in the vertical direction, thus following half of the possible operator movements. Since edge direction is perpendicular to the gradient, this modified canny operator produces mainly horizontal edge lines, resulting in a mask denoted as $\mathbf{M}^e_{b_2}$.

The binary maps $\mathbf{M}^e_{b_1}$ and $\mathbf{M}^e_{b_2}$ are then combined,

$$\mathbf{M}^e_{b_3} = \mathbf{M}^e_{b_1} + \mathbf{M}^e_{b_2}, \tag{3}$$

to produce map $\mathbf{M}^e_{b_3}$ illustrated in Figure 5(a). Edges directly above and below the eye centre in map $\mathbf{M}^e_{b_3}$, which are depicted by arrows in Figure 5(a), are selected as eyelids and the space between them as $\mathbf{M}^e_2$, as shown in Figure 5(b).



(a)

(b)

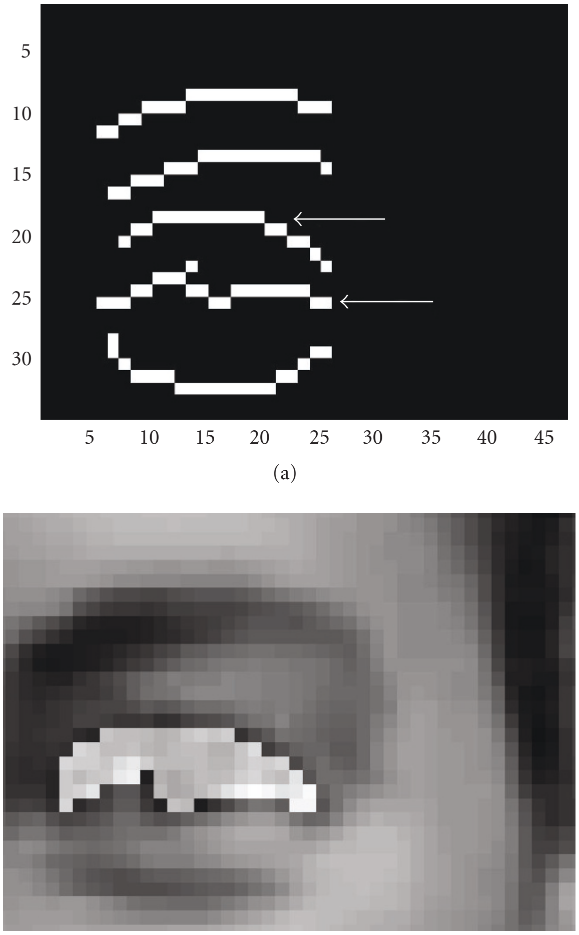
Figure 4: (a) Left eye input image (cropped). (b) Left eye mask $\mathbf{M}^e_1$ depicting distance transform values of selected object.

4.1.3. *Standard-deviation-based mask*

A third mask is created for each of the eyes to strengthen the final mask fusion stage. This mask is created using a region growing technique; the latter usually gives very good segmentation results corresponding well to the observed edges. Construction of this mask relies on the fact that facial texture is more complex and darker inside the eye area and especially in the eyelid-sclera-iris borders than in the areas around them. Instead of using an edge density criterion, we developed a simple but effective new method to estimate both the eye centre and eye mask.

Table 5: Summary of mouth masks.

Described in	Detects	Depends on	Results
Section 4.4.1	Lips and mouth with similar properties to ones trained from the neutral frame	$\mathbf{M}_t^m$, Mouth-candidate image (color)	$\mathbf{M}_1^m$
Section 4.4.2	Horizontal edges caused by lips	$\mathbf{L}^m$	$\mathbf{M}_2^m$
Section 4.4.3	Mouth horizontal extent through lip corner detection. Mouth opening through lip edge detection	$\mathbf{L}^m$	$\mathbf{M}_3^m$



(a)

(b)

Figure 5: (a) Modified canny result. (b) Detected mask $\mathbf{M}_2^e$.

We first calculate the standard deviation of the luminance channel $\mathbf{L}^e$ in $n\times n$ sliding blocks resulting in $\mathbf{I}^e_{\text{std}_n}$. $\mathbf{I}^e_{\text{std}_n}$ is iteratively thresholded with $(1/d)\mathbf{L}^e$, where d is a divisor increasing in each iteration, resulting in $\mathbf{M}^e_{s_{n,d}}$. While d increases, areas in $\mathbf{M}^e_{s_{n,d}}$ dilate, tending to connect with each other.

This operation is performed at first for $n = 3$. The eye centre is selected on the first iteration as the centre of the largest component; for iteration i, the estimated eye centre is denoted as $\mathbf{c}_i$ and the procedure continues while $\|\mathbf{c}_1 - \mathbf{c}_i\| \leq W_f t_c^e$ resulting in binary map $\mathbf{M}^e_{s_{3,f}}$, as illustrated in Figure 6(a). This is an indication that eye area has exceeded its actual borders and is now connected to other subfeatures. The same process is repeated with $n = 6$ resulting in map $\mathbf{M}^e_{s_{6,f}}$ illustrated in Figure 6(b). Different block sizes are used to raise the procedure's robustness to variations of image resolution and eye detail information. Smaller block sizes converge slower to their final map but the combination of both type of maps results in map $\mathbf{M}_3^e$, as in the case of Figure 6(c), ensuring a better result in case of outliers. Examples of outliers include compression artifacts, which induce abrupt illumination variations. For pixel coordinates (i, j), the above are implemented as follows:

$$\begin{aligned}
\mathbf{L}^e &= \{l_{i,j}\}, \\
\mathbf{I}^e_{\text{std}_n} &= \{i_{n,i,j}\}, \quad i_{n,i,j} = \sqrt{\langle l_{i,j}^2 \rangle - \langle l_{i,j} \rangle^2}, \\
m_{n,d,i,j} &= \begin{cases} 1, & \dfrac{l_{i,j}}{d} > i_{n,i,j}, \\ 0, & \dfrac{l_{i,j}}{d} < i_{n,i,j}, \end{cases} \quad n = 3, 6, \\
\mathbf{M}^e_{s_{n,d}} &= \{m_{n,d,i,j}\},
\end{aligned} \tag{4}$$

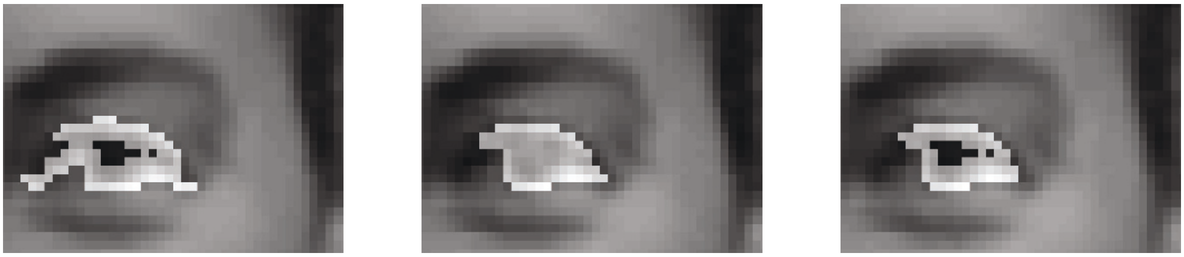
where $d \in (0, \max(\mathbf{L}^e)]$ and $\langle \bullet \rangle$ denotes the mean in the $n\times n$ area surrounding (i, j),

$$\begin{aligned}
f_a(\mathbf{A}, \mathbf{B}) &= \{c_{ij}\}, \quad c_{ij} = a_{ij} b_{ij}, \\
\mathbf{M}_3^e &= f_a\left(\mathbf{M}^e_{s_{2n,f}}, \mathbf{M}^e_{s_{n,f}}\right).
\end{aligned} \tag{5}$$

The above process is similar to a morphological bottom hat operation with the difference that the latter is rather sensitive to the structuring element size.

4.1.4. Luminance mask

Finally, a second luminance-based mask is constructed for eye/eyelid border extraction. In this mask, we compute the normal luminance probability of $\mathbf{L}^e$ resembling to the mean luminance value of eye area defined by the NN mask $\mathbf{M}^e_{\text{nn}}$. From the resulting probability mask, the areas with a confidence interval of t_d^e are selected and small gaps are closed with morphological filtering. The result is usually a blob depicting the boundaries of the eye. In some cases, the luminance values around the eye are very low due to shadows from the eyebrows and the upper part of the nose. To improve the outcome in such cases, the detected blob is cut vertically at its thinnest points from both sides of the eye centre; the resulting mask's convex hull is then denoted as $\mathbf{M}_4^e$ and illustrated in Figure 7.



(a) (b) (c)

Figure 6: (a) $\mathbf{M}^{e}_{s_{3,f}}$ eye mask for $n = 3$. (b) $\mathbf{M}^{e}_{s_{6,f}}$ eye mask for $n = 6$. (c) $\mathbf{M}^{e}_{3}$, combination of (a) and (b).

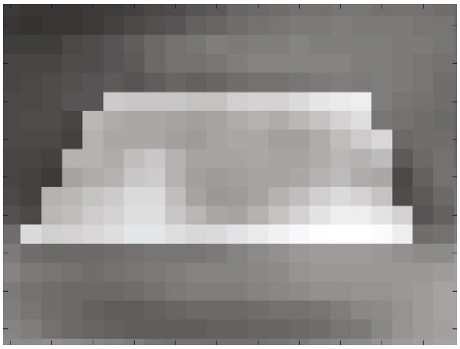
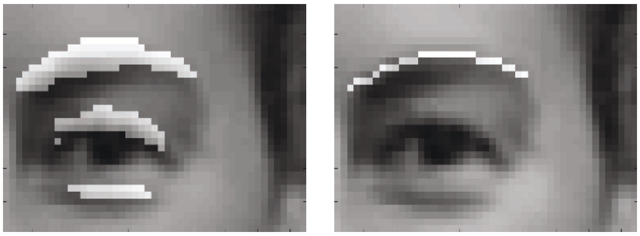


Figure 7: Left eye mask $\mathbf{M}^{e}_{4}$.



(a) (b)

Figure 8: (a) Eyebrow candidates. (b) Selected eyebrow mask $\mathbf{M}^{b}$.

4.2. *Eyebrow boundary detection*

Eyebrows are extracted based on the fact that they have a simple directional shape and that they are located on the forehead, which due to its protrusion, has a mostly uniform illumination. Each of the left and right eye and eyebrow-candidate images shown in Figure 2 is used for brow mask construction.

The first step in eyebrow detection is the construction of an edge map M^{b}_{E} of the grayscale eye/eyebrow-candidate image. This map is constructed by subtracting the dilation and erosion of the grayscale image using a line structuring element st^{b}_{2} pixels long and then thresholding the result as shown in Figure 8(a):

$$\begin{gathered}
\mathbf{M}^{b}_{E_1} = \delta_s(\mathbf{L}^{e}), -\varepsilon_s(\mathbf{L}^{e}),\\
t^{b}_{E} = \left(\langle \mathbf{M}^{b}_{E_1}\rangle + \sqrt{\left\langle (\mathbf{M}^{b}_{E_1})^{2}\right\rangle - \langle \mathbf{M}^{b}_{E_1}\rangle^{2}}\right),\\
\mathbf{M}^{b}_{E} = \mathbf{M}^{b}_{E_1} > t^{b}_{E},
\end{gathered} \tag{6}$$

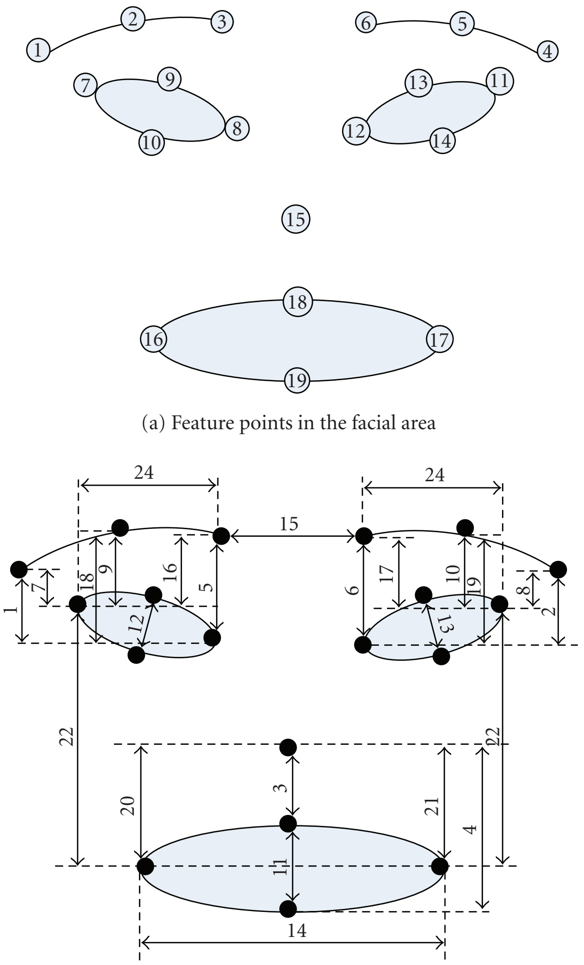
where δ_s, ε_s denote the dilation and erosion operators with structuring element s, and operator ">" denotes the thresholding operator to construct the binary mask $\mathbf{M}^{b}_{E}$. The selected edge detection mechanism is appropriate for eyebrows because it can be directional, it preserves the feature's original size and can be combined with a threshold to remove smaller skin anomalies such as wrinkles. The above procedure can be considered as a nonlinear high-pass filter.

Each connected component on the edge map is labeled and then tested against a set of filtering criteria. These criteria were formed through statistical analysis of the eyebrow lengths and positions on 20 persons of the ERMIS database [30]. Firstly, the major axis is found for each component through principal component analysis (PCA). All components whose major axis has an angle of more than 30 degrees with the horizontal plane are removed from the set. From the remaining components, those whose axis length is smaller than t^{b}_{1} are removed. Finally, components with a lateral distance from the eye centre more than $t^{b}_{1}/2$ are removed and the top-most remaining is selected resulting in the eyebrow mask $\mathbf{M}^{b}_{E_2}$. Since eyebrow area is of no importance for FAP calculation, the result can be simplified easily using (7) resulting in $\mathbf{M}^{b}$ which is depicted in Figure 8(b):

$$\begin{aligned}
\mathbf{M}^{b} &= \{m_{i,j}\},\\
\mathbf{M}^{b}_{E} &= \{m^{E}_{i,j}\},\\
m^{E}_{i,j} &= \begin{cases} 1, & (m_{i,j} = 1) \wedge (m_{i,j'} \neq 1),\ j' < j,\\ 0 & \text{otherwise.}\end{cases}
\end{aligned} \tag{7}$$

4.3. *Nose localization*

The nose is not used for expression estimation by itself, but is a fixed point that facilitates distance measurements for FAP estimation (Figure 9(a)), thus, its boundaries do not have to be precisely located. Nose localization is a feature frequently used for face tracking and usually based on nostril localization; nostrils are easily detected based on their low intensity [46].



(a) Feature points in the facial area

(b) Feature point distances

Figure 9

The facial area above the mouth-candidate components area is used for nose location. The respective luminance image is thresholded by t_1^n:

$$t_1^n = \frac{1}{3}\left(\langle \mathbf{L}^n \rangle + 2\min\left(\mathbf{L}^n\right)\right), \qquad (8)$$
$$\mathbf{L}^n : \text{luminance of nose-candidate region.}$$

Connected objects of the derived binary map are labeled. In bad lighting conditions, long shadows may exist along either side of the nose. For this reason, anthropometric data [47] about the distance of left and right eyes (bipupil breadth, D_{bp}) is used to reduce the number of candidate objects: objects shorter than t_2^n and longer than $t_3^n D_{\text{bp}}$ are removed. This has proven to be an effective way to remove most outliers without causing false negative results while generating the nostril mask $\mathbf{M}_1^n$ shown in Figure 10(a).

Horizontal nose coordinate is predicted from the coordinates of the two eyes. On mask $\mathbf{M}_1^n$, each of the connected component horizontal distances from the predicted nose centre is compared to the average internostril distance that is approximately $t_4^n D_{\text{bp}}$ [47], and components with the largest ones are considered as outliers. Those who qualify enter two separate lists, one including left-nostril candidates and one with right-nostril candidates based on their proximity to the left or right eye. Those lists are sorted according to their luminance and the two objects with the lowest values are retained from each list. The largest object is finally kept from each list and labeled as the left and right nostril, respectively, as shown in Figure 10(b). The nose centre is defined as the midpoint of the nostrils.

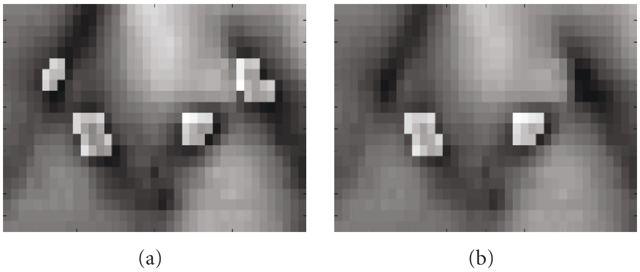


Figure 10: (a) Nostril candidates, (b) selected nostrils.

4.4. Mouth detection

4.4.1. Neural network lip and mouth detection mask

At first, mouth boundary extraction is performed on the mouth-candidate facial area depicted in Figure 2. An MLP neural network is trained to identify the mouth region using the neutral image. Since the mouth is closed in the neutral image, a long low-luminance region exists between the lips. The detection of this area, in this work, is carried out as follows.

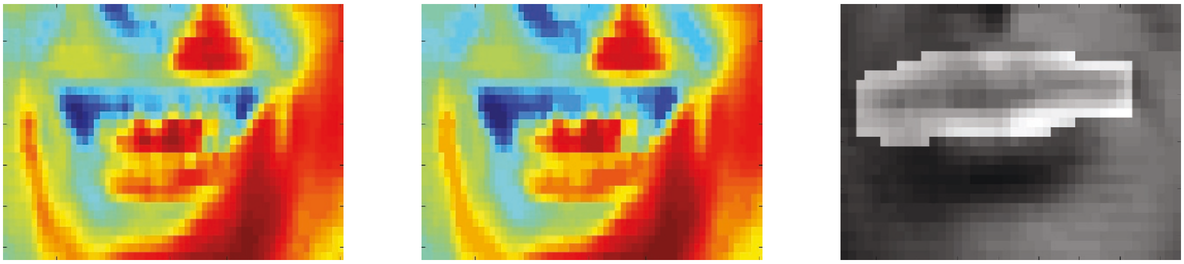
The initial mouth-candidate luminance image $\mathbf{L}^m$ shown in Figure 11(a) is simplified to reduce the presence of noise, remove redundant information, and produce a smooth image that consists mostly of flat and large regions of interest. Alternating sequential filtering by reconstruction (ASFR) (9) is thus performed on $\mathbf{L}^m$ to produce $\mathbf{L}^m_{\text{asfr}}$ shown in Figure 11(b). ASFR ensures preservation of object boundaries through the use of connected operators [48],

$$\begin{aligned} &f_{\text{asfr}}(I) = \beta_n \alpha_n \cdots \beta_2 \alpha_2 \beta_1 \alpha_1 (I), \quad n = 1, 2, \ldots, \\ &\alpha_r(I) = \rho^-(f \ominus rB \mid f), \qquad \beta_r(I) = \rho^+(f \oplus rB \mid f), \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad r = 1, 2, \ldots, \\ &\rho^{+(-)}(g \mid f): \text{ reconstruction closing (opening)} \\ &\qquad\qquad \text{of } f \text{ by marker } g, \end{aligned} \qquad (9)$$

where the operations $\oplus$ and $\ominus$ denote the Minkowski dilation and erosion.

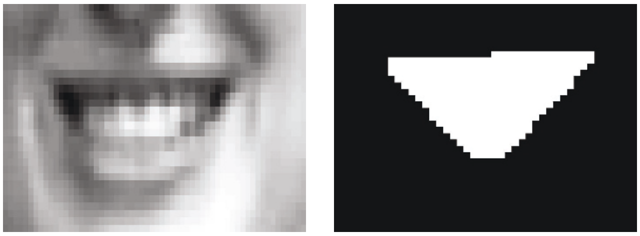
To avoid over simplification, the ASFR filter is applied with a scale of $n \le d_m^w \cdot t_{\text{asf}}^m$, where d_m^w is the width of $\mathbf{L}^m$. The luminance image is then thresholded by t_1^m:

$$t_1^m = \frac{1}{3}\left(2\overline{\mathbf{L}^m} + \min\left(\mathbf{L}^m_{\text{asfr}}\right)\right), \qquad (10)$$



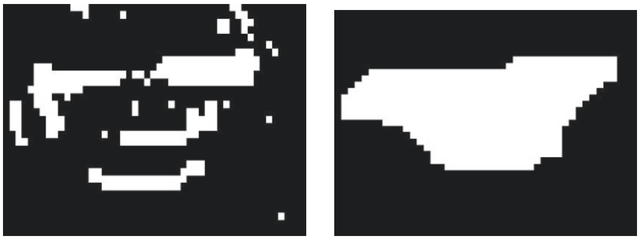
(a) (b) (c)

Figure 11: Extraction of training image: (a) initial luminance map $\mathbf{L}^m$, (b) filtered image $\mathbf{L}^m_{\text{asfr}}$, (c) extracted mask $\mathbf{M}^m_{t_1}$.



(a) (b)

Figure 12: (a) Luminance image, (b) NN mouth mask $\mathbf{M}^m_1$.



(a) (b)

Figure 13: (a) Initial binary edge map. (b) Output mask $\mathbf{M}^m_{b_2}$.

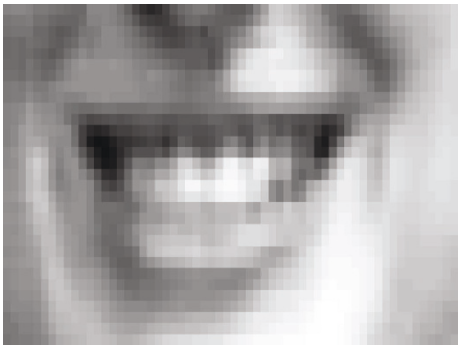


Figure 14: Mouth-candidate area depicting nonuniform illumination.

and connected objects on the resulting binary mask $\mathbf{M}^m_{t_1}$ are labeled as shown in Figure 11(c).

The major axis of each connected component is computed through PCA analysis, and the one with the longest axis is selected. The latter is subsequently dilated vertically and the resulting mask $\mathbf{M}^m_t$ is produced, which includes the lips. Mask $\mathbf{M}^m_t$ shown in Figure 11(c) is used to train a neural network to classify the mouth and nonmouth areas accordingly. The image area included by the mask corresponds to the mouth class and the image outside the mask to the nonmouth one. The perceptron has 13 inputs and its architecture is similar to that of the network used for eye detection.

The neural network trained on the neutral-expression frame is then used on other frames to produce an estimate of the mouth area: neural network output on the mouth-candidate image is thresholded by t^m_2 and those areas with high confidence are kept to form a binary map containing several small subareas. The convex hull of these areas is calculated to generate mask $\mathbf{M}^m_1$ as shown in Figure 12.

4.4.2. *Generic edge connection mask*

In this second approach, the mouth luminance channel is again filtered using ASFR for image simplification. The horizontal morphological gradient of $\mathbf{L}^m$ is then calculated similarly to the eyebrow binary edge map detection resulting in $\mathbf{M}^m_{b_1}$ shown in Figure 13(a). Since the nose has already been detected, its vertical position is known. The connected elements of $\mathbf{M}^m_{b_1}$ are labeled and those too close to the nose are removed. From the rest of the map, very small objects (less than $t^m_{b1}I_w$, where I_w is the map's width) are removed. Morphological closing is then performed so that those whose distance is less than $t^m_{b2}I_w$ connect together, in order to obtain mask $\mathbf{M}^m_{b_2}$ as shown in Figure 13(b). The longest of the remaining objects in horizontal sense is selected as mouth mask $\mathbf{M}^m_2$.

4.4.3. *Lip-corner luminance and edge information fusion mask*

The problem of most intensity-based methods that try to estimate mouth opening is the visibility of upper teeth, especially if they appear between the upper and lower lip altering saturation and intensity uniformity as illustrated in Figure 14.

A new method is proposed next to cope with this problem. First, the mouth-candidate luminance channel $\mathbf{L}^m$ is

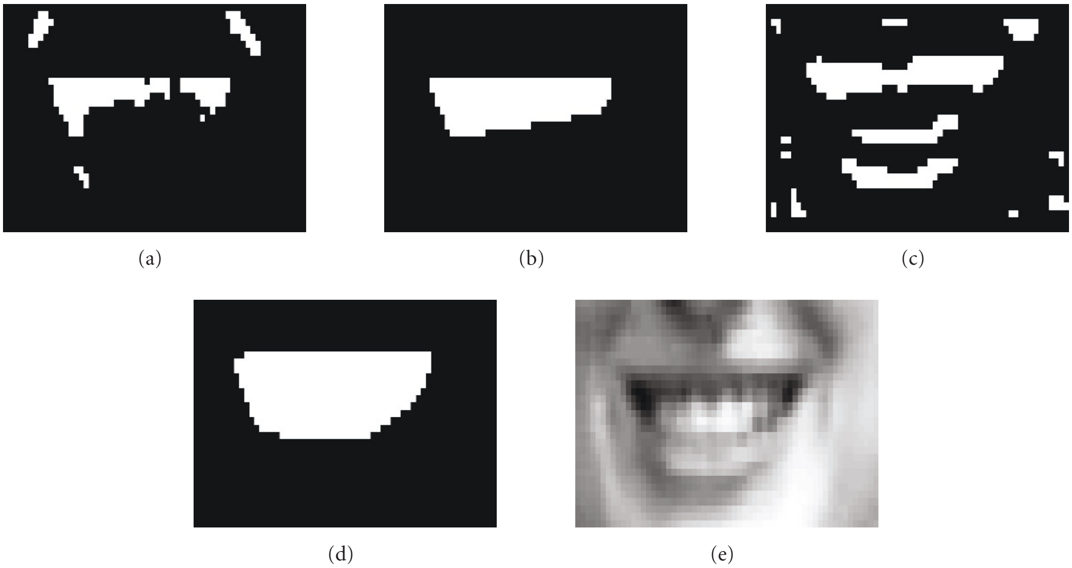


Figure 15: (a) Mask $\mathbf{M}^m_{c_1}$ with removed background outliers, (b) mask $\mathbf{M}^m_{c_2}$ with apparent teeth, (c) horizontal edge mask $\mathbf{M}^m_{c_3}$, (d) output mask $\mathbf{M}^m_3$, (e) input image.

thresholded using a low threshold t^m_{c1} providing an estimate of the mouth interior area, or the area between the lips in case of a closed mouth. The threshold used is estimated adaptively:

$$\mathbf{M}^m_{c_{1a}} = \mathbf{L}^m_{\text{asfr}} < t^m_{c1},$$
$$t^m_{c1} = \left(\langle \mathbf{L}^m_{\text{asfr}}\rangle - \sqrt{\left\langle \left[\mathbf{L}^m_{\text{asfr}}\right]^2\right\rangle - \langle \mathbf{L}^m_{\text{asfr}}\rangle^2}\right), \tag{11}$$

where operator "<" again stands for the thresholding process.

In the resulting binary map, all connected objects adjacent to the border are removed, thus removing facial background outliers, resulting in mask $\mathbf{M}^m_{c_1}$ shown in Figure 15(a). We now examine two cases separately: either we have no apparent teeth and the mouth area is denoted by a cohesive dark area (case 1) or teeth are apparent and thus two dark areas appear at both sides of the teeth (case 2). It should be noted that those areas appear even in large extensive smiles. The largest connected object is then selected from $\mathbf{M}^m_{c_1}$ and its centroid is found. If the horizontal position of its centroid is near the horizontal nose position, case 1 is selected, otherwise case 2 is assumed to occur and two dark areas appear at both sides of the teeth. To assess horizontal noise centre proximity, we use a distance threshold of $t^m_{c2}D_{\text{bp}}$. The two cases are quite distinguishable through this process. In case 2, the second largest connected object is also selected. A new binary map is created containing either one object in case 1 or both objects in case 2; the convex hull of this map is then calculated and mask $\mathbf{M}^m_{c_2}$ is produced, depicted in Figure 15(b).

The detected lip corners provide a robust estimation of mouth horizontal extent but are not adequate to detect mouth opening. Therefore, mask $\mathbf{M}^m_{c_2}$ is expanded to include the lower lips. An edge map is created as follows: the mouth image gradient is calculated in the horizontal direction, and is thresholded by the median of its positive values, as shown in Figure 15(c). This mask, denoted as $\mathbf{M}^m_{c_3}$, contains objects close to the lower middle part of the mouth, which are sometimes missed because of the lower teeth. The two masks, $\mathbf{M}^m_{c_2}$ and $\mathbf{M}^m_{c_3}$, have to be combined to a final mask. An effective way of achieving this is to keep from both masks objects which are close to each other. Since $\mathbf{M}^m_{c_2}$ may contain objects belonging to lower parts of the mouth area, it is expanded downwards by dilation with a nonsymmetric vertical structuring element, resulting in mask $\mathbf{M}^m_{c_{2d}}$. Morphological reconstruction [49] is then used to combine the masks together by using the area belonging to both $\mathbf{M}^m_{c_3}$ and $\mathbf{M}^m_{c_{2d}}$ as input and objects belonging to either mask (12) as marker. Final mask $\mathbf{M}^m_3$ is shown in Figure 15(d),

$$\begin{gathered}
\mathbf{M}^m_{c_{23}} = f_a\left(\mathbf{M}^m_{c_3}, \mathbf{M}^m_{c_{2d}}\right),\\
f_a(\mathbf{A},\mathbf{B}) = \{c_{ij}\}, \quad c_{ij} = \{a_{ij}b_{ij}\},\\
f_o(\mathbf{A},\mathbf{B}) = \{c_{ij}\}, \quad c_{ij} = \begin{Bmatrix} a_{ij}, b_{ij} = 0 \\ b_{ij}, a_{ij} = 0 \end{Bmatrix},\\
\mathbf{M}^m_3 = \rho\left(\mathbf{M}^m_{c_{23}}, f_o\left(\mathbf{M}^m_{c_{2d}}, \mathbf{M}^m_{c_3}\right)\right),
\end{gathered} \tag{12}$$

where $\rho(\mathbf{B},\mathbf{A})$ denotes the reconstruction of $\mathbf{A}$ with marker $\mathbf{B}$.

5. FINAL MASKS GENERATION AND CONFIDENCE ESTIMATION

Each facial feature's masks must be fused together to produce a final mask for that feature. The most common problems, especially encountered in low quality input images, include connection with other feature boundaries or mask dislocation due to noise, as depicted in Figure 16. In some cases, some masks may have completely missed their goal and provide a completely invalid result. Outliers such as illumination

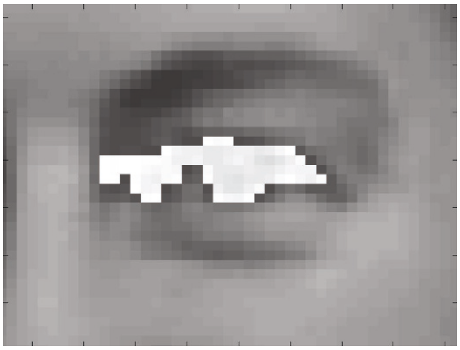


Figure 16: Noisy color and edge information cause problems in the extraction of this mask.

changes and compression artifacts cannot be predicted and so individual masks have to be re-evaluated and combined on each new frame.

5.1. *Validation of eye and mouth masks*

The proposed algorithms presented in Section 4 produce a mask $\mathbf{M}^b$ for each eyebrow, nose coordinates, four intermediate mask estimates $\mathbf{M}^e_{1\dots4}$ for each eye and three intermediate mouth mask estimates $\mathbf{M}^m_{1\dots3}$. The four masks for each eye and three mouth masks must be fused to produce a final mask for each feature. Since validation can only be done on the end result of each intermediate mask, we unfortunately cannot give different parts of each intermediate mask different confidence values, so each pixel of those masks will share the same value. We propose validation through testing against a set of anthropometric conformity criteria. Since, however, some of these criteria relate either to aesthetics or to transient feature properties, we cannot apply strict anthropometric judgment.

For each mask k of feature x, we employ a set of validation measurements $V^x_{k,i}$, denoted by i, which are then combined to a final validation tag $V^x_{k,f}$ for that mask. Each measurement produces a validation estimate value depending on how close it is to the usually expected feature shape and position, in the neutral expression. Expected values for these measurements are defined from anthropometry data [33] and from images extracted from video sequences of 20 persons in our database [30]. Thus, a validation tag between [0,1] is attached to each mask, with higher values denoting proximity to the most expected measurement values.

All validation measurements are based on distances defined in Table 6. Given these definitions, eye mask validation is based on four tags specified in Table 7, concerning individual eye dimensions, relations between the two eyes and relations between each eye and the corresponding eyebrow. Finally, mouth map validation is based on four tags referring to distance measurements specified in Table 8. In the following, validation value of measurement i for mask k of feature x will be denoted as $V^x_{k,i} \in [0,1]$ where $V^x_{k,i}$ is forced into [0,1], that is, if $V^x_{k,i} > 1$, then $V^x_{k,i} = 1$ and if $V^x_{k,i} < 0$, then $V^x_{k,i} = 0$.

Table 6: Mask validation distances.

d_1	Distance of eye's top horizontal coordinate and eyebrow's middle bottom horizontal coordinate
d_2	Eye width
d_3	Eye height
d_4	Distance of eye's middle vertical coordinate and eyebrow's middle vertical coordinate
d_5	Eyebrow width
d_6	D_{bp}, bipupil breadth
d_7	Distance of eye's middle vertical coordinate from mouth's middle vertical coordinate
d_8	Mouth width
d_9	Mouth height
d_{10}	Sellion-Stomion length
d_{11}	Sellion-Subnasion length

We want masks with very low validation tags to be discarded from the fusion process and thus those are also prevented from contribution on final validation tags; therefore, we ignore those with $V^x_{k,f} < (t_{vd} \cdot \langle V^x_{k,i} \rangle_i)$. Final validation tag for mask k is then calculated as follows:

$$V^x_{k,f} = \langle V^x_{k,i'} \rangle_{i'}, \quad i' : V^x_{k,i'} \geq t_{vd} \langle V^x_{k,i} \rangle_{i}, \; i \in \mathbb{N}_n. \tag{13}$$

5.2. *Mask fusion*

Each of the intermediate masks represents the best-effort result of the corresponding mask-extraction method used. Multiple eye and mouth masks must be merged to produce final mask estimates for each feature. The mask fusion method is based on the assumption that having multiple masks for each feature lowers the probability that all of them are invalid since each of them produces different error patterns. It has been proven in committee machine (CM) theory [50, 51] that for the desired output t the combination error $y_{\text{comb}} - t$ from different machines f_i is guaranteed to be lower than the average error:

$$\begin{gathered} y_{\text{comb}} = \frac{1}{M} \sum y_i, \\ \left(y_{\text{comb}} - t\right)^2 = \frac{1}{M} \sum_i \left(y_i - t\right)^2 - \frac{1}{M} \sum_i \left(y_i - y_{\text{comb}}\right)^2. \end{gathered} \tag{14}$$

Since intermediate masks have a validation tag which represents their "plausibility" of being actual masks for the feature they represent, it seems natural to combine them by giving more credit to those which have a higher validation value on one hand, and on the other to ignore those that we are sure will not contribute positively on the result. Furthermore, according to the specific qualities of each input, we would like to favor specific masks that are known to perform better on those inputs, that is, give more trust to color-based extractors when it is known that input has good color quality, or to the neural network-based masks when the face resolution is enough for the network to perform adequate border detection.

Regarding input quality, two parameters can be taken into account: image resolution and color quality; since

Table 7: Anthropometric validation measurements used for eye masks. Note that (eye width)/(bipupil breadth) = 0.49 [33].

Validation tag	Measurement	Description
$V^{e}_{k,1}$	$1 - \lvert (d_1)/(d_6/4) - 1 \rvert$	Distance of the eye's topmost centre from the corresponding eyebrow's bottom centre.
$V^{e}_{k,2}$	$1 - \lvert 1 - (d_2/d_6)/0.49 \rvert$	Eye width compared to left & right eye distance.
$V^{e}_{k,3}$	$0.3 - (d_3 - d_2)/d_2$	Relation of eye width and height
$V^{e}_{k,4}$	$1 - \lvert d_4 \rvert / d_5$	Horizontal alignment of the eye and respective eyebrow

Table 8: Anthropometric validation measurements used for mouth masks. Note that (bichelion breadth)/(bipupil breadth) = 0.82 and (stomion-subnasion length)/(bipupil breadth) = 0.344 [33].

Validation tag	Measurement	Description
$V^{m}_{k,1}$	$1 - \lvert d_7 \rvert / d_6$	Horizontal mouth centre, in comparison with the inter-eye centre coordinate.
$V^{m}_{k,2}$	$1 - \left\lvert \dfrac{d_8}{d_6}\dfrac{1}{0.82} - 1 \right\rvert$	Mouth width in comparison with bipupil breadth
$V^{m}_{k,3}$	1 if $d_9 < (1.3 d_6)$ else $d_9/(1.3 d_6)$	Mouth height in comparison with bipupil breadth
$V^{m}_{k,4}$	$1 - \left\lvert 1 - \dfrac{(d_{10} - d_{11})}{d_6}\dfrac{1}{0.344} \right\rvert$	Nose distance from top lip

nonsynthetic training data for the latter is difficult to acquire, we have found that a good estimator can be the chromatic deviation measured on the face skin area: very large variability in chromatic components is a good indicator for color noise presence. Therefore, σ_{Cr}, σ_{Cb} are less than t_σ for good color quality and much larger for poor quality images. Regarding resolution, we have found that the proposed neural-network-based detector performs very well in sequences where $D_{\mathrm{bp}} > t_r$ pixels, where D_{bp} denotes the bipupil breadth.

In the following, we use the following notation: final masks for left eye, right eye, and mouth are denoted as before as $\mathbf{M}^{e_L}_f$, $\mathbf{M}^{e_R}_f$, $\mathbf{M}^{m}_f$. For intermediate mask k of feature x, variable $V^x_{k,f}$ determines which masks are favored according to their final validation values and variable g^k determines which masks extractors are favored according to input characteristics. Moreover, each pixel-element on the final mask $\mathbf{M}^x_f$ is denoted as m^x_f and each pixel-element on the kth intermediate mask $\mathbf{M}^x_k$ as m^x_k, $k \in \mathbb{N}_n$, where pixel coordinates are omitted for clarity. Moreover, since we would like masks to be fused in a per-pixel basis, not all pixels on an output mask will necessarily derive from the same intermediate masks. Therefore, each pixel on the output mask will have a validation value v^x_f which will reflect mask validation and extractor suitability of the masks it derived; values of v^x_f for all pixels form validation values of final mask, V^x_f.

Let us denote the function between $m^x_f \in \{0,1\}$, $v^x_f \in [0,1]$, and $m^x_k \in \{0,1\}$ as

$$\begin{aligned} v^x_f &= f\left(m^x_k; V^x_{k,f}, g^k\right), \\ m^x_f &= F\left(v^x_f\right), \end{aligned} \tag{15}$$

then our requirements can be expressed as follows.

(1) If all masks k agree that a pixel m^x_k does not belong to the feature x, then this should be reflected on the fusion result regardless of validation tags $V^x_{k,f}$:

$$\text{if } \forall k \in \mathbb{N}_n, \quad m^x_k = 0 \Longrightarrow m^x_f = 0. \tag{16}$$

(2) We require that gating variable g^k should be balanced according to the number of masks:

$$\sum_{k=1}^{n} g^k = n. \tag{17}$$

(3) If all masks k agree that a pixel m^x_f does belong to feature x with maximum confidence, then this should be reflected on the fusion result:

$$\text{if } \forall k \in \mathbb{N}_n, \quad m^x_k = 1 \wedge V^x_{k,f} = 1 \Longrightarrow m^x_f = 1,\ v^x_f = 1. \tag{18}$$

(4) If all masks k have failed, then no mask should be created as a fusion result:

$$\forall k \in \mathbb{N}_n, \quad V^x_{k,f} = 0 \Longrightarrow m^x_f = 0. \tag{19}$$

(5) If one mask has failed, then the result should depend only on remaining masks:

$$\exists k_0 \in \mathbb{N}_n : V^x_{k_0,f} = 0 \Longrightarrow m^x_f = \underset{k \in \mathbb{N}_n - \{k_0\}}{f}\left(m^x_k; V^x_{k,f}, g^k\right). \tag{20}$$

(6) Fusion with a better input mask should produce a higher value on the output for the pixels deriving from this mask:

$$\begin{aligned} &\text{if } V^{x_1}_{k_0,f} > V^{x_2}_{k_0,f}, \quad \forall k \in \mathbb{N}_n - \{k_0\}, \\ &\text{it is } V^{x_1}_{k,f} = V^{x_2}_{k,f} \text{ and the same holds for all} \\ &m^{x_j}_k \neq 0, \quad g^{k,j}, \quad j = 1,2 \text{ then } v^{x_2}_f > v^{x_2}_f. \end{aligned} \tag{21}$$

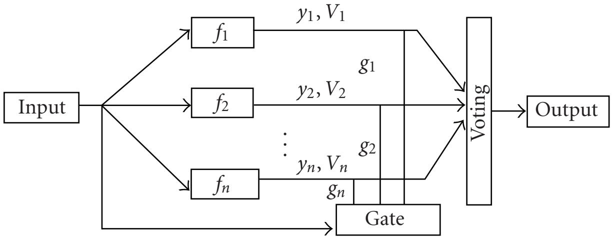


Figure 17: The dynamic committee machine model.

(7) If an input mask derives from a more trusted mask extractor, then pixels deriving from this mask should be associated with a higher value:

$$\text{if } g^{k,1} > g^{k,2}, \quad \forall k \in \mathbb{N}_n - \{k_0\} \text{ it is } V^{x_1}_{k,f} = V^{x_2}_{k,f},$$
$$\text{and the same holds for all } m^{x_j}_k \neq 0, \quad V^{x_j}_{k,f}, \quad j = 1,2 \quad (22)$$
$$\text{then } v^{x_2}_f > v^{x_2}_f.$$

To fulfill these requirements in this work, we propose a fusion method based on the idea of dynamic committee machines (DCM) which is depicted in Figure 17. In a static CM, the voting weight for a component is proportional to its error on a validation set. In DCMs, input is directly involved in the combining mechanism through a gating network (GN), which is used to modify those weights dynamically.

The machine's inputs are intermediate masks $\mathbf{M}^x_k$, $V^x_{k,f}$ is considered as the confidence of each input and variable g^k has a "gating" role. Final masks $\mathbf{M}^{e_L}_f$, $\mathbf{M}^{e_R}_f$, $\mathbf{M}^m_f$ are considered as the machine's output.

Each pixel-element m^x_f on the final mask $\mathbf{M}^x_f$ is calculated from the n masks as follows:

$$f\left(m^x_k; V^x_{k,f}, g^k\right) = \frac{1}{n} \sum_{k=1}^{n} m^x_k V^x_{k,f} g^k, \tag{23}$$

$$F\left(v^x_f\right) = \begin{cases} 0, & v^x_f < \left(\left\langle V^x_f \right\rangle \mid v^x_f > 0\right), \\ 1, & \text{otherwise.} \end{cases} \tag{24}$$

The role of gating variable g^k is used to favor color-aware feature extraction methods ($\mathbf{M}^e_1$, $\mathbf{M}^m_1$) in images of high-color quality and resolution; gating variable g^i is defined as follows:

$$g^k = \begin{cases} n - \left(\dfrac{n-1}{n}\right), & k = 1,\ D_{\text{bp}} > t_r,\ \sigma_{\text{Cr}} < t_\sigma,\ \sigma_{\text{Cb}} < t_\sigma, \\ \dfrac{1}{n}, & k \neq 1,\ D_{\text{bp}} > t_r,\ \sigma_{\text{Cr}} < t_\sigma,\ \sigma_{\text{Cb}} < t_\sigma, \\ 1, & \text{otherwise,} \end{cases} \tag{25}$$

where D_{bp} the bipupil width in pixels, σ_{Cr}, σ_{Cb} the standard deviation of the Cr, Cb channels, respectively, inside the facial area. It is not difficult to see that (23)–(25) satisfy (16)–(22).

Tables 9 and 10 illustrate mask fusion examples for the left eye and mouth where some of the masks are problematic. Validation tags refer to the corresponding mask validation tag while D_{bp} is quoted as an indication of the sequence resolution. For illustration purposes, the feature points extracted from the final masks are presented verifying the precise extraction of the features and feature points, based on the mask fusion process.

5.3. Eye, eyebrow, and mouth mask confidence estimation

Confidence values are needed for expression analysis and are thus propagated from mask extraction to the corresponding FPs, FAPs, and the expression evaluation stage. Their role is to indicate the confidence that a given feature has been correctly extracted and therefore the measure by which expression analysis should rely on a specific feature. To estimate confidence, we have used extracted feature resemblance to mean anthropometry data from [33]. Since data for eyebrow sizes was not available in the literature, confidence values were expanded to rely also on information such as facial feature size constancy and face symmetry.

Confidence values can be attached to each final mask and are denoted as $C^e, C^b, C^m \in [0,1]$. Confidence values vary between 0 and 1 with the latter indicating the best case. For the nose, no confidence value is estimated and is always assumed that $C^n = 1$. Those values are generated through a set of criteria, which complement final validation tags V_f used for fusion; these criteria relate to

(1) size constancy over time, producing C^b_{med}, C^e_{med};

(2) face symmetry, producing C^e_s;

(3) and anthropometric measurement conformance, producing C^e_1, C^e_2, C^m_1.

These values are calculated as follows.

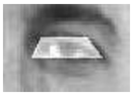
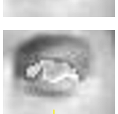
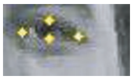
(1) With the exception of mouth, facial feature width is mostly constant even in intense expressions. Measured width for eyebrows w^b_i and each of the eyes $w^{e_L}_i$, $w^{e_R}_i$ is examined in each frame i the median value $\widetilde{w}^x$ over the last 10 frame period for feature x is calculated. In each frame, similarity between w^x_i and $\widetilde{w}^x$ on the last 10 frames is used as an estimate for C^b_{med} for the eyebrows and C^e_{med} for the eyes:

$$C^x_{\text{med},i} = 1 - \left| w^x_i - \text{med}\left(w^x_j,\ j = i - 10 \ldots i\right) \right| \left(w^x_i\right)^{-1}. \tag{26}$$

(2) $C^e_s \in [0,1]$ denotes shape similarity between the left and right upper eyelid; exploiting the symmetry of the face, we estimate the resemblance between the upper parts of left and right eyelids. Let us define $\mathbf{X}^L$, $\mathbf{X}^R$ as matrices containing the horizontal coordinates of the left and right upper eyelid boundaries; a value C_s indicating their similarity can be calculated as a two-dimensional correlation coefficient between the two vectors,

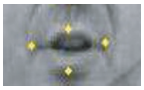
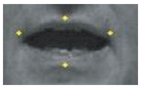
$$C^e_s = \frac{\sum_n \left(\left(\mathbf{X}^L_n - \left\langle \mathbf{X}^L \right\rangle\right)\left(\mathbf{X}^R_n - \left\langle \mathbf{X}^R \right\rangle\right)\right)}{\sqrt{\left(\sum_n \left(\mathbf{X}^L_n - \left\langle \mathbf{X}^L \right\rangle\right)^2\right)\left(\sum_n \left(\mathbf{X}^R_n - \left\langle \mathbf{X}^R \right\rangle\right)^2\right)}}. \tag{27}$$

Table 9: Examples of mask fusion on the left eye with corresponding validation tags and detected feature points.

Sequence-frame	kk-1002	V_f^e	kk-1998	V_f^e	rd-12259	V_f^e	al-27	V_f^e
Mask								
$\mathbf{M}_{\mathrm{nn}}^e$								
$\mathbf{M}_1^e$		0.825		0.813		0.823		0.839
$\mathbf{M}_2^e$		0.782		0.581		0.763		0.810
$\mathbf{M}_3^e$		0.866		0.733		0.716		0.787
$\mathbf{M}_4^e$		0.883		0.917		0.826		0.872
$\mathbf{M}_f^e$								
FPs								
	D_{bp}: 58 px		D_{bp}: 58 px		D_{bp}: 96 px		D_{bp}: 36 px	

D_{bp} denotes bipupil breadth in pixels and is quoted as an image resolution indicative.

Table 10: Examples of mouth mask fusion with corresponding validation tags and detected feature points.

Sequence-frame	kk-1014	V_f^m	rd-1113	V_f^m
Mask				
$\mathbf{M}_1^m$		0.820		0.538
$\mathbf{M}_2^m$		0.868		0.752
$\mathbf{M}_3^m$		0.828		0.821
$\mathbf{M}_f^m$				
FPs				

(3) C^e, C^m, C^b are calculated using measurements based on anthropometry from [33]. Table 11 summarizes estimation of C_1^e, C_2^e, C_1^m.

Confidence values for features are estimated by averaging on the previously defined criteria and final mask validation tags as follows:

$$\begin{aligned} C^e &= \langle V_f^e, C_1^e, C_2^e, C_s^e, C_{\mathrm{med}}^e \rangle, \\ C^m &= \langle V_f^m, C_1^m \rangle, \\ C^b &= C_{\mathrm{med}}^b. \end{aligned} \tag{28}$$

6. EXPRESSION ANALYSIS

An overview of the expression recognition process is shown in Figure 1. At first, 19 feature points (FPs) are calculated from the corresponding feature masks. Those FPs have to be compared with the FPs of the neutral frame, so as to measure movement and estimate FAPs. FAPs are then used to evaluate expression profiles, providing the recognized expression.

6.1. From masks to feature points

Left-, right-, top-, and bottom-most coordinates of the final masks $\mathbf{M}_f^{e_L}$, $\mathbf{M}_f^{e_R}$, $\mathbf{M}_f^m$, left right and top coordinates of $\mathbf{M}_f^{b_L}$, $\mathbf{M}_f^{b_R}$, as well as nose coordinates, are used to define the 19 feature points (FPs) shown in Table 12, Figures 18 and 9(a). Feature point x is then assigned with confidence C_x^{FP} by inheriting the confidence level (C^e, C^m, C^b, C^n) of the final mask from which it derives.

Table 11: Anthropometric evaluation [33] for eye and mouth location and size.

Description	Confidence measure
Bientocanthus breadth	D_7^a
Biectocanthus breadth	D_5^a
Bicheilion breadth	D_{10}^a
$(D_5^a - D_7^a)/2$	D_{ew}^a
Eye position/eye distance	$C_1^e = 1 - \lvert D_5^{an} - D_5^n \rvert / D_5^{an}$
Eye width	$C_2^e = 1 - \lvert D_{ew}^{an} - D_{ew}^n \rvert / D_{ew}^{an}$
Mouth	$C_1^m = 1 - \lvert D_{10}^{an} - D_{10}^n \rvert / D_{10}^{an}$

$D_i^{an} = D_x^a/D_7^a$: a denotes that distance i derives from [33]; n denotes that value is normalized by division with D_7^a.

Table 12: Feature points.

FP no.	MPEG-4 FP [6]	FP name
01	4.5	Outer point of left eyebrow
02	4.3	Middle point of left eyebrow
03	4.1	Inner point of left eyebrow
04	4.6	Outer point of right eyebrow
05	4.4	Middle point of right eyebrow
06	4.2	Inner point of right eyebrow
07	3.7	Outer point of left eye
08	3.11	Inner point of left eye
09	3.13	Upper point of left eyelid
10	3.9	Lower point of left eyelid
11	3.12	Outer point of right eye
12	3.8	Inner point of right eye
13	3.14	Upper point of right eyelid
14	3.10	Lower point of right eyelid
15	9.15	Nose point
16	8.3	Left corner of mouth
17	8.4	Right corner of mouth
18	8.1	Upper point of mouth
19	8.2	Lower point of mouth

6.2. From FP to FAP estimation

A 25-dimensional distance vector (D_v) is created containing vertical and horizontal distances between 19 extracted FPs, as shown in Figure 9(b). Distances are not measured in pixels, but in normalized scale-invariant MPEG-4 units, that is, ENS, MNS, MW, IRISD, and ES [6]. Unit bases are measured directly from FP distances on the neutral image; for example, ES is calculated as $|\mathrm{FP}_9, \mathrm{FP}_{13}|$.

The distance vector is created once for the neutral-expression image (D_v^n) and for each of the subsequent frames (D_v). FAPs are calculated by comparing D_v^n and D_v. Each FAP depends on one or more elements of D_v thus some FAPs are over defined; the purpose of calculating a FAP from more distances than necessary is to increase estimation robustness which is accomplished by considering the confidence levels of each distance element. Elements in D_v are calculated by measuring the FP distances illustrated in Figure 9(b). Uncertainty in FP coordinates should reflect to corresponding FAPs; therefore, distances needed to calculate an FAP are weighted according to the confidence of the corresponding FP from which they derive.

Table 13: Example of FAPs and related distances.

MPEG4 FAP	Description	Distance number
F_3	open_jaw	11
F_4	lower_top_midlip	3
F_5	raise_bottom_midlip	4
$F_6 + F_7$	widening_mouth	14
$F_{19} + F_{21}$	close_left_eye	12
$F_{20} + F_{22}$	close_right_eye	13
F_{31}	raise_left_inner_eyebrow	5,16
F_{32}	raise_right_inner_eyebrow	6,17
F_{33}	raise_left_medium_eyebrow	18,9
F_{34}	raise_right_medium_eyebrow	19,10
F_{35}	raise_left_outer_eyebrow	7,1
F_{36}	raise_right_outer_eyebrow	8,2
F_{37}	squeeze_left_eyebrow	24
F_{38}	squeeze_right_eyebrow	25
$F_{37} + F_{38}$	squeeze_eyebrows	15
F_{59}	raise_left_outer_cornerlip	22
F_{60}	raise_right_outer_cornerlip	23

A value C_i^{FAP} indicating the confidence of FAP i is estimated as $C_i^{\mathrm{FAP}} = \langle C_{\mathrm{Y}}^{\mathrm{FP}} \rangle$, Y:set of FPs used to estimate FAP i. Correspondences between FAPs and corresponding distance vector elements are illustrated in Table 13.

6.3. Facial expression recognition and human computer interaction

In our former research on expression recognition, a ru le-based system was created, characterising a user's emotional state in terms of the six universal, or archetypal, expressions (joy, surprise, fear, anger, disgust, sadness). We have created rules 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 [7]. This dataset contains several images for every one of the six archetypal expressions, which, however, are rather exaggerated. As a result, rules extracted from this dataset do not perform well if used in real human-computer interaction environments. Psychological studies describing the use of quadrants of emotion's wheel (see Figure 19) [52] instead of the six archetypal expressions provide a more appropriate tool in such interactions. Therefore, creation of rules describing the first three quadrants—no emotion is lying in the fourth quadrant—is necessary.

To accomplish this, facial muscle movements were translated into FAPs while each expression's FAPs on every quadrant were experimentally verified through analysis of prototype datasets. Next, the variation range of each FAP was computed by analysing real interactions and corresponding video sequences as well as by animating synthesized exam-

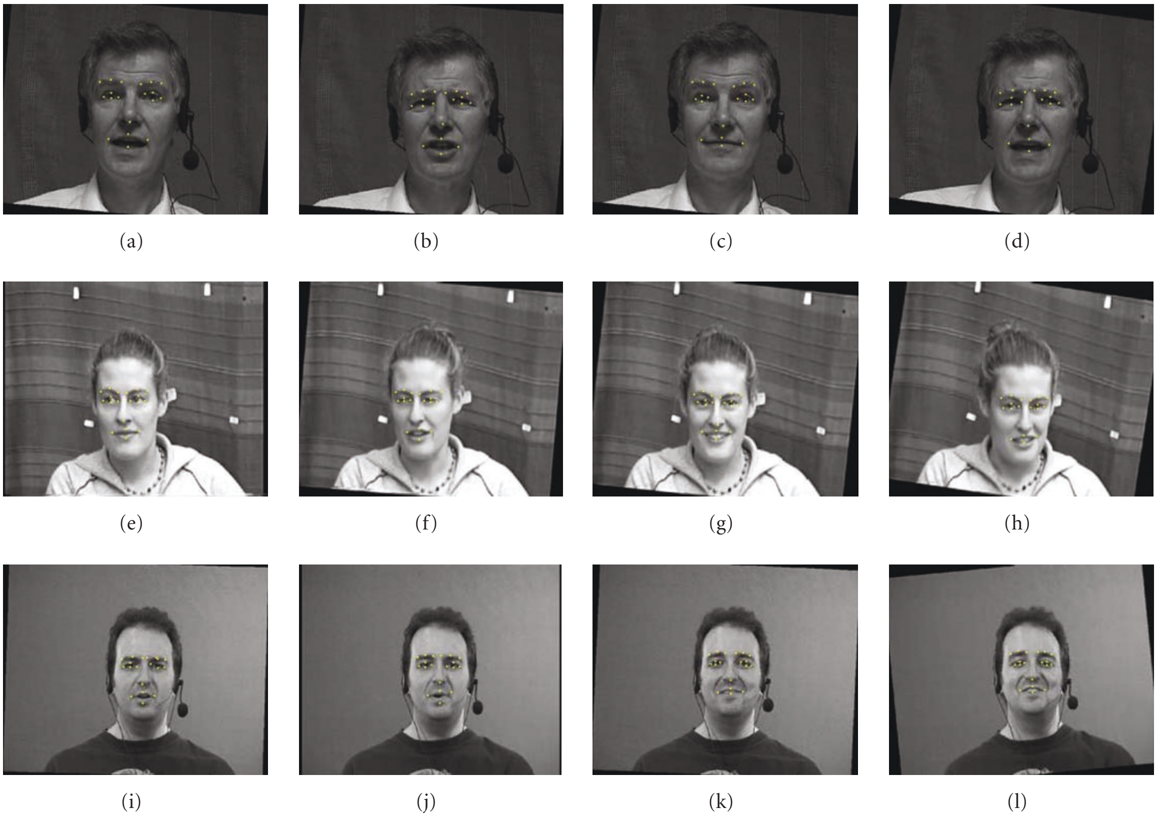


FIGURE 18: The 19 detected feature points. Automatic head-pose recovery has been performed.

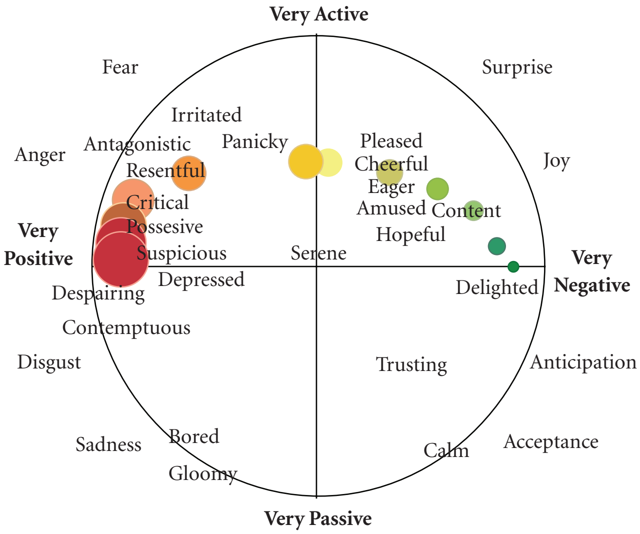


FIGURE 19: The activation-emotion space.

ples. Table 14 illustrates three examples of rules that were created based on the developed methodology.

In order to use these rules in a system dealing with the continuous activation-emotion space and fuzzy representation, we transformed the rules replacing the range of variation with the terms *high, medium, low* after having normalized the corresponding partitions. The full set of rules can be found in [53].

In the process of exploiting the knowledge contained in the fuzzy rule base and the information extracted from each frame in the form of FAP measurements, with the aim to analyze and classify facial expressions, a series of issues have to be tackled.

(i) FAP activation degrees need to be considered in the estimation of the overall result.

(ii) The case of FAPs that cannot be estimated, or equivalently are estimated with a low degree of confidence, needs to be considered,

$$\text{if } x_1, x_2, \ldots, x_n, \quad \text{then } y. \tag{29}$$

The conventional approach to the evaluation of fuzzy rules of the form described in (29) is as follows [54]:

$$y = t(x_1, x_2, \ldots, x_n), \tag{30}$$

where t is a fuzzy t-norm, such as the minimum

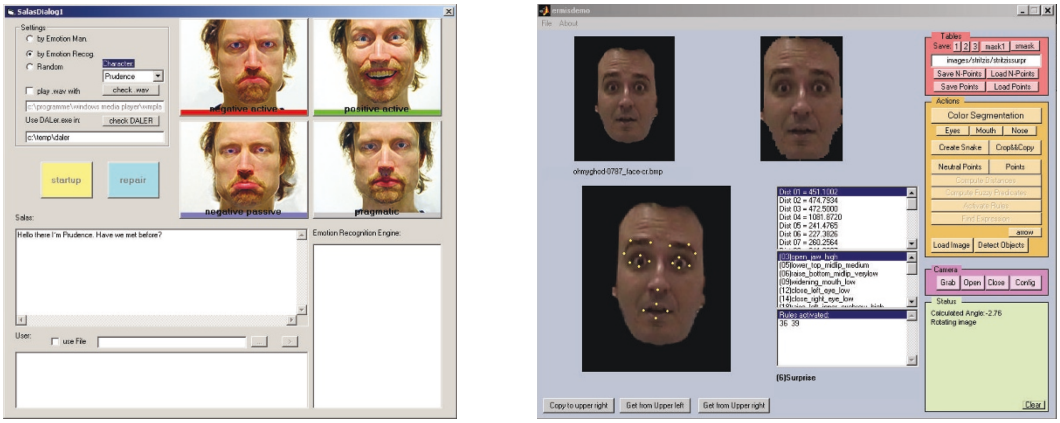
$$t(x_1, x_2, \ldots, x_n) = \min(x_1, x_2, \ldots, x_n), \tag{31}$$

the algebraic product

$$t(x_1, x_2, \ldots, x_n) = x_1 \cdot x_2 \cdot \ldots \cdot x_n, \tag{32}$$

Table 14: Rules with FAP range of variation in MPEG-4 units.

Rule	Quadrant
$F_6 \in [160, 240]$, $F_7 \in [160, 240]$, $F_{12} \in [260, 340]$, $F_{13} \in [260, 340]$, $F_{19} \in [-449, -325]$, $F_{20} \in [-426, -302]$, $F_{21} \in [325, 449]$, $F_{22} \in [302, 426]$, $F_{33} \in [70, 130]$, $F_{34} \in [70, 130]$, $F_{41} \in [130, 170]$, $F_{42} \in [130, 170]$, $F_{53} \in [160, 240]$, $F_{54} \in [160, 240]$	(++)
$F_{16} \in [45, 155]$, $F_{18} \in [45, 155]$, $F_{19} \in [-330, -200]$, $F_{20} \in [-330, -200]$, $F_{31} \in [-200, -80]$, $F_{32} \in [-194, -74]$, $F_{33} \in [-190, -70]$, $F_{34} \in [-190, -70]$, $F_{37} \in [65, 135]$, $F_{38} \in [65, 135]$	(−+)
$F_3 \in [400, 560]$, $F_5 \in [-240, -160]$, $F_{19} \in [-630, -570]$, $F_{20} \in [-630, -570]$, $F_{21} \in [-630, -570]$, $F_{22} \in [-630, -570]$, $F_{31} \in [460, 540]$, $F_{32} \in [460, 540]$, $F_{33} \in [360, 440]$, $F_{34} \in [360, 440]$, $F_{35} \in [260, 340]$, $F_{36} \in [260, 340]$, $F_{37} \in [60, 140]$, $F_{38} \in [60, 140]$	(−+)



c:\programme\windows media player\wmpla

c:\temp\daler

Hello there I'm Prudence. Have we met before?

(a)

images/stritzis/stritzissurpr

ohmyghod-0787_face-cr.bmp

Dist 01 = 451.1002
Dist 02 = 474.7934
Dist 03 = 472.5000
Dist 04 = 1081.8720
Dist 05 = 241.4765
Dist 06 = 227.3826
Dist 07 = 260.2564

(03)open_jaw_high
(05)lower_top_midlip_medium
(06)raise_bottom_midlip_verylow
(09)widening_mouth_low
(12)close_left_eye_low
(14)close_right_eye_low

Rules activated:
36 39

(6)Surprise

Calculated Angle:-2.76
Rotating image

(b)

Figure 20: (a) SALAS interaction interface. (b) Facial expression analysis interface.

the bounded sum

$$t(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \cdots + x_n + 1 - n, \qquad (33)$$

and so on. Another well-known approach in rule evaluation is described in [55] and utilizes a weighted sum instead of a t-norm in order to combine information from different rule antecedents:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n. \qquad (34)$$

Both approaches are well studied and established in the field of fuzzy automatic control. Still, they are not adequate for the case of facial expression estimation: their main disadvantage is that they assume that all antecedents are known, that is, that all features are measured successfully and precisely. In the case of facial expression estimation, FAPs may well be estimated with a very low confidence, or not estimated at all, due to low video quality, occlusion, noise, and so on. Thus, a more flexible rule evaluation scheme is required, that is able to incorporate such uncertainty as well. Moreover, the second one of the conventional approaches, due to the summation form, has the disadvantage of possibly providing a highly activated output even in the case that an important antecedent is known to be missing; obviously, it is not suitable for the case examined in this paper, where the non-activation of an FAP automatically implies that the expression profiles that require it are not activated either. For this reason, in this work we have used a flexible rule evaluation scheme [56], which is in fact a generalization of the t-norm-based conventional approach. In this approach and in the t-norm operation described in (30), antecedents with lower values affect most the resulting value of y, while antecedents with values close to one have trivial and negligible affect on the value of y. Having that in mind, we can demand that only antecedents that are known with a high confidence will be allowed to have low values in that operation. Then, the activation level of a rule with this approach can be interpreted in a possibilistic manner, that is, it can be interpreted as the degree to which the corresponding output is possible, according to the available information; in the literature, this possibilistic degree is referred to as plausibility. The confidence is determined by the confidence values of the utilized inputs, that is, by the confidence values of the rule antecedents, as follows:

$$y^c = \frac{x_1^c + x_2^c + \cdots + x_n^c}{n}. \qquad (35)$$

7. EXPERIMENTAL RESULTS

7.1. Test data generation: the SALAS-emotion induction framework

Our test data have been produced using the SALAS testbed application developed within the ERMIS and HUMAINE projects, which is an extension of one of the highlights of AI research in the 1960s, Weizenbaum's ELIZA [57]. The ELIZA framework simulates a Rogerian therapy, during which clients talk about their problems to a listener that provides responses that induces further interaction without passing any comment or judgment.

Recording is an integral part of this challenge. With the requirement of both audio and visual inputs, the need to compromise between demands of psychology and signal processing is imminent. If one is too cautious about the recording quality, subjects may feel restrained and are unlikely to show the everyday, relaxed emotionality that would cover most of the emotion representation space. On the other hand, visual and audio analysis algorithms cannot be expected to cope with totally unconstrained head and hand movement, subdued lighting, and mood music. Major issues may also arise from the different requirements of the individual modalities: while head mounted microphones might suit analysis of speech, they can have devastating consequences for visual analysis. Eventually arrangements were developed to ensure that on the visual side, the face was usually almost frontal and well and evenly lit to the human eye; that it was always easy for a human listener to make out what was being said; and that the setting allowed most human participants to relax and express emotion within a reasonable time.

The implementation of SALAS is mainly a software application designed to let a user work through various emotional states. It contains four "personalities" shown in Figure 20(a) that listen to the user and respond to what he/she says, based on the different emotional characteristics that each of the "personalities" possesses. The user controls the emotional tone of the interaction by choosing which "personality" they will interact with, while still being able to change the tone at any time by choosing a different personality to talk to.

The initial recording took place with 20 subjects generating approximately 200 minutes of data. The second set of recordings comprised 4 subjects recording two sessions each, generating 160 minutes of data, providing a total of 360 minutes of data from English speakers; both sets are balanced for gender, 50/50 male/female. These sets provided the input to facial feature extraction and expression recognition system of this paper.

7.2. Facial feature extraction results

Facial feature extraction can be seen as a subcategory of image segmentation, that is, image segmentation into facial features. According to Zhang [58] segmentation algorithms can be evaluated analytically or empirically. Analytical methods directly treat the algorithms themselves by considering the principles, requirements, utilities, complexity, and so forth of algorithms; while these methods can provide an algorithm evaluation which is independent from the implementation itself or the arrangement and choice of input data, very few properties of the algorithm can be obtained or is practical to obtain through analytical study. On the other hand, empirical methods can be divided in two categories: empirical goodness methods, which use a specific "goodness"; measure to evaluate the performance of algorithms, and empirical discrepancy methods which measure the discrepancy between the automatic algorithm result and an ideally labeled image. Zhang 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 does not necessarily require expert annotation, it is clear that 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 interobserver variability. A way to compensate for the latter is Williams' Index (WI) [59], 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 [60]. 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:

$$\mathrm{WI} = \frac{(1/n)\sum_{j=1}^{n}\left(1/D_{0,j}\right)}{\left(2/n(n-1)\right)\sum_{j}\sum_{j':j'>j}\left(1/D_{j,j'}\right)}, \tag{36}$$

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

$$D_{j,j'} = \frac{1}{D_{\mathrm{bp}}}\left|\left|M_j^x \veebar M_{j'}^x\right|\right|, \tag{37}$$

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

From a dataset of about 50 000 frames, 250 frames were selected at random and the 19 FPs were manually selected from two observers on each one. WI was calculated using (36) for each feature and for each frame separately. At a value of 0, the computer mask is infinitely far from the observer mask. When *WI* is larger than 1, the computer generated mask disagrees less with the observers than the observers disagree with each other. Distribution of the average *WI* calculated over the two eyes and mouth for each frame is shown in Figure 21, while Figure 22 depicts the average *WI* calculated on the two eyebrows. Table 15 summarizes the results.

For the eyes and mouth, *WI* has been calculated for 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; $\langle \mathrm{WI}_x \rangle$ denotes the average *WI* for mask x calculated over all test frames.

Column 7 of Table 15 shows the percentage of frames where the mask fusion resulted in an improvement of the *WI*, while columns 8 and 9 display the average *WI* in the frames

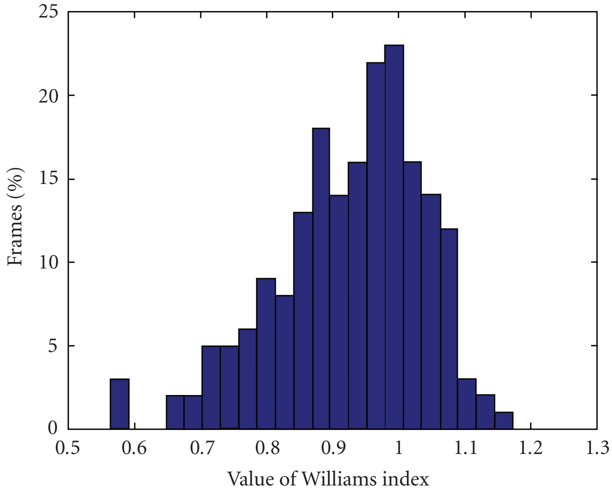


Figure 21: Williams index distribution (average on eyes and mouth).

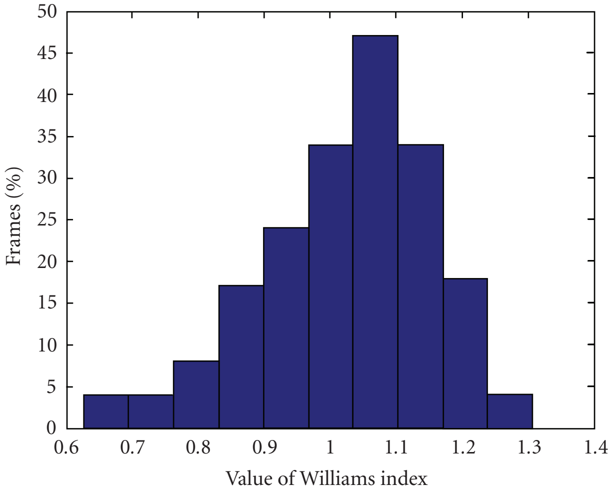


Figure 22: Williams index distribution (average on left and right eyebrows).

where the fusion result was better and worse from the single mask, respectively. One may be tempted to deduct from this table that some feature detectors perform better than the combined mask result; it may seem so when considering the average values, but this is not the case when examining each frame: different methods perform better for different input images and the average results that seem to favor some methods over the others are dependent on the selection of the input frames. This may be also justified especially by looking at the result variation between the left and right eyes for the same mask, as well as from the values of column 8: the average *WI* on frames where eye mask 4 performed better than the fused result is still a bit lower than the total average *WI*, thus it may seem that this mask performs better for the specific test but the improvement is not significant; this means that even when considering the same sequence, the average values may be slightly better for one mask but relying solely on this one mask, the system will have no safeguard to refer to when the algorithm resulting in this mask performs poorly. The latter is demonstrated in column 10 where one can see that when using the fused masks, the worst cases will be on average better than the worse case of the mask with the best mean *WI*. Nevertheless, the aim of this work is not to find the best feature extractor, but to combine them intelligently with respect to the input video. What can be deducted about the different masks is that looking at the value differences between column 8 and column 9, one can conclude that for example eye mask 0 performs better in "very difficult" test frames, where the total average *WI* has a value of 0.69.

7.3. Expression analysis results

Since the ERMIS dataset was created by engaging participants to emotional dialogue, facial expressions in these video sequences were not acted and extreme, but are mostly naturalistic. We evaluated sequences totalling about 30 000 frames. Expression analysis results were tested against manual multimodal annotation from experts [61] and the results are presented in Table 16.

In order to produce the facial expression analysis results, we utilized the neurofuzzy network presented in [53]. The architecture of this network was able to exploit not only FAPs values produced by tracking the feature points and their distances, but also the confidence measures associated with each intermediate result. Since we are dealing with video sequences depicting human-computer interaction, expressivity, head movement and rotation are usually unconstrained. As a result, exact feature point localization is not always possible due to changing lighting conditions, such as varying shadow artifacts introduced by the eyebrow protrusion or the nose. It is given that the contribution of the algorithm presented here lies not only in the fact that it performs stable feature point localization, but more importantly in the fusion process and the confidence measure that it produces for each mask, as well as the fused result. The confidence measure is utilized by the neurofuzzy network to reduce the importance of a set of FAP measurements in a frame where confidence is low, thereby catering for better network training and adaptation, since the network is trained with examples that perform better. The significance of this approach is proven with the increase in performance shown in [53], as well as in the second column of Table 16, where the possibilistic approach [56], which also utilizes the confidence measure, also outperforms a "naive" fuzzy rule implementation based only on FAP values.

In addition to this, Column 5 in Table 15 indicates that the fusion step almost always improves the performance of the individual masks, in the sense that it produces a final result which agrees more with the expert annotators than in the case of the single masks (higher Williams Index value, which produces a ratio of the fused mask over the single masks > 1). The robustness of the feature extraction process, when combined with the provision of confidence measures, is shown in the videos at http://www.image.ece.ntua.gr/ijivp. These videos contain the results from the feature extraction process per frame, and the estimated quadrant which con-

Table 15: Result summary.

Algorithm[1]	Mask #	$\langle \mathrm{WI}_x \rangle$	$\langle \mathrm{WI}_f \rangle$	$\frac{\langle \mathrm{WI} \rangle_f}{\langle \mathrm{WI} \rangle_x}$	σ^2	% of frames where $\mathrm{WI}_f > \mathrm{WI}_x$	$\langle \mathrm{WI} \rangle$ in frames where $\mathrm{WI}_f < \mathrm{WI}_x$	$\langle \mathrm{WI} \rangle$ in frames where $\mathrm{WI}_x < \mathrm{WI}_f$	$\langle \mathrm{WI} \rangle$ in 5% worst frames[4]
1	2	3	4	5	6	7	8	9	10
Left Eye									
NN[2]	0	0.677	—	1.287	0.103	74.2	0.697	0.885	0.351
Section 4.1.1	1	0.701	—	1.216	0.056	78.8	0.731	0.868	0.414
Section 4.1.2	2	0.821	0.838	1.029	0.027	82.4	0.770	0.887	0.459
Section 4.1.3	3	0.741	—	1.131	0.057	76.2	0.811	0.847	0.265
Section 4.1.4	4	0.870	—	0.979	0.026	44.3	0.812	0.867	0.427
	f	0.838	—	1.000	—	—	—	—	0.475
Right Eye									
NN[2]	0	0.800	—	1.093	0.020	75.2	0.672	0.946	0.411
Section 4.1.1	1	0.718	—	1.243	0.084	81.4	0.674	0.929	0.352
Section 4.1.2	2	0.774	0.875	1.140	0.021	58.2	0.836	0.883	0.396
Section 4.1.3	3	0.650	—	1.346	0.028	84.5	0.632	0.920	0.305
Section 4.1.4	4	0.893	—	0.982	0.02	48.4	0.778	0.996	0.418
	f	0.875	—	1.000	—	—	—	—	0.429
Mouth									
Section 4.4.1	1	0.763	—	1.051	0.046	59.2	0.752	0.772	0.288
Section 4.4.2	2	0.823	0.780	0.963	0.038	44.8	0.721	0.852	0.345
Section 4.4.3	3	0.570	—	1.446	0.204	96.9	0.510	0.793	0.220
	f	0.780	—	1.000	—	—	—	—	0.359
Eyebrows[3]									
Left		1.034	—	—	—	—	—	—	—
Right		1.013	—	—	—	—	—	—	—

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

$\langle \bullet \rangle$ denotes the average over all features in all frames, $\langle \bullet \rangle_f$ denotes the average of the final masks over all frames while $\langle \bullet \rangle_x$ denotes the average of mask x over all frames.

[1] Refer to indicated subsection number

[2] NN denotes $\mathbf{M}^e_{\mathrm{nn}}$, the eye mask derived directly from the neural network output

[3] Using eyebrow mask $\mathbf{M}^b_{E_2}$, prior to thinning

[4] $\langle \mathrm{WI} \rangle$ in the 5% of total frames with the lowest *WI*.

Table 16: Comparison of results between manual and two automatic expression analysis approaches.

Naive fuzzy rules	Possibilistic approach	Annotator disagreement
65.1%	78.4%	20.01%

tains the observed facial expression. Even though feature localization may be inaccurate or even fail in specific frames, this fact is identified by a low-confidence measure, effectively instructing the expression analysis algorithm to ignore these features and try to estimate the facial expression on the remaining results.

As a general rule, the last column of Table 16 indicates that the human experts that classify the frames to generate the ground truth make contrasting evaluations once every five frames; this fact is clearly indicative of the ambiguity of the observed emotions in a naturalistic environment. It is also worth underlining that this system achieves a 78% classification rate while operating based solely on expert knowledge provided by humans in the form of fuzzy rules, without weights for the rule antecedents. Allowing for the specification of antecedence importance as well as for rule optimization through machine learning is expected to provide for even further enhancement of the achieved results.

8. CONCLUSIONS

In this work we have presented a method to automatically locate 19 facial feature points that are used in combination with the MPEG-4 facial model for expression estimation. A robust method for locating these features has been presented which also extracts a confidence estimate depicting a "goodness" measure of each detected point, which is used by the expression recognition stage; the provision of this measure enables the expression recognition process to discard falsely located features, thus enhancing performance in recognizing both universal (basic) emotion labels, as well as intermediate expressions based on a dimensional representation. Our algorithm can perform well under a large variation of facial image quality, color, and resolution.

Since the proposed method only handles roll facial rotation, an extension to be considered is the incorporation of a facial model. Recently, a lot of work has been done in facial feature detection and fitting of facial models [62]. While these techniques can detect facial features, but not extract their precise boundary, they can extend our work by accurately predicting the face position in each frame. Thus, feature candidate areas would be defined with greater precision allowing the system to work even under large head rotation and feature occlusion.

REFERENCES

[1] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, no. 9, pp. 52–55, 1968.

[2] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[3] P. Ekman and W. V. Friesen, *Facial Action Coding Systems: A Technique for the Measurement of Facial Movement*, Consulting Psychologist Press, Palo Alto, Calif, USA, 1978.

[4] C. Tomasi and T. Kanade, "Detection and tracking of point features," Tech. Rep. CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, Pa, USA, April 1991.

[5] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[6] A. M. Tekalp and J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication*, vol. 15, no. 4, pp. 387–421, 2000.

[7] 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, 2002.

[8] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[9] A. Lanitis, C. J. Taylor, T. F. Cootes, and T. Ahmed, "Automatic interpretation of human faces and hand gestures using flexible models," in *Proceedings of the 1st International Workshop on Automatic Face and Gesture Recognition (FG '95)*, pp. 98–103, Zurich, Switzerland, September 1995.

[10] Y. Yacoob and L. S. Devis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 636–642, 1996.

[11] C. L. Lisetti and D. E. Rumelhart, "Facial expression recognition using a neural network," in *Proceedings of the 11th International Florida Artificial Intelligence Research Society Conference*, pp. 328–332, AAAI Press, Sanibel Island, Fla, USA, May 1998.

[12] S. Kaiser and T. Wehrle, "Automated coding of facial behavior in human-computer interactions with facs," *Journal of Nonverbal Behavior*, vol. 16, no. 2, pp. 67–84, 1992.

[13] G J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *Proceedings of the 5th European Conference on Computer Vision (ECCV '98)*, vol. 2, pp. 581–595, Freiburg, Germany, June 1998.

[14] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facial expression," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition (FG '98)*, pp. 396–401, Nara, Japan, April 1998.

[15] M. J. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *International Journal of Computer Vision*, vol. 25, no. 1, pp. 23–48, 1997.

[16] K.-M. Lam and H. Yan, "An analytic-to-holistic approach for face recognition based on a single frontal view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 7, pp. 673–686, 1998.

[17] H. Gu, G.-D. Su, and C. Du, "Feature points extraction from face images," in *Proceedings of the Image and Vision Computing Conference (IVCNZ '03)*, pp. 154–158, Palmerston North, New Zealand, November 2003.

[18] S.-H. Leung, S.-L. Wang, and W.-H. Lau, "Lip image segmentation using fuzzy clustering incorporating an elliptic shape function," *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 51–62, 2004.

[19] N. Sarris, N. Grammalidis, and M. G. Strintzis, "FAP extraction using three-dimensional motion estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 10, pp. 865–876, 2002.

[20] Y. Tian, T. Kanade, and J. F. Cohn, "Robust lip tracking by combining shape, color and motion," in *Proceedings of the 4th Asian Conference on Computer Vision (ACCV '00)*, pp. 1040–1045, Taipei, Taiwan, January 2000.

[21] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FG '04)*, pp. 517–522, Seoul, South Korea, May 2004.

[22] D. DeCarlo and D. Metaxas, "The integration of optical flow and deformable models with applications to human face shape and motion estimation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pp. 231–238, San Francisco, Calif, USA, June 1996.

[23] M. Pantic and L. J. M. Rothkrantz, "Expert system for automatic analysis of facial expressions," *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

[24] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[25] C.-L. Huang and Y.-M. Huang, "Facial expression recognition using model-based feature extraction and action parameters classification," *Journal of Visual Communication and Image Representation*, vol. 8, no. 3, pp. 278–290, 1997.

[26] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[27] H. Hong, H. Neven, and C. von der Malsburg, "Online facial expression recognition based on personalized galleries," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition (FG '98)*, pp. 354–359, Nara, Japan, April 1998.

[28] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

[29] S. J. McKenna, Y. Raja, and S. Gong, "Tracking colour objects using adaptive mixture models," *Image and Vision Computing*, vol. 17, no. 3-4, pp. 225–231, 1999.

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

[31] HUMAINE IST, "Human-Machine Interaction Network on Emotion," 2004–2007, http://www.emotion-research.net/.

[32] ISTFACE, "MPEG-4 Facial Animation System—Version 3.3.1 Gabriel Abrantes," (Developed in the context of the European Project ACTS MoMuSys 97-98 Instituto Superior Tecnico).

[33] J. W. Young, "Head and Face Anthropometry of Adult U.S. Civilians," FAA Civil Aeromedical Institute, 1963–1993, (final report 1993).

[34] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[35] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proceedings of the 6th IEEE International Conference on Computer Vision (ICCV '98)*, pp. 555–562, Bombay, India, January 1998.

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 511–518, Kauai, Hawaii, USA, December 2001.

[37] I. Fasel, B. Fortenberry, and J. Movellan, "A generative framework for real time object detection and classification," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 182–210, 2005.

[38] R. Fransens, J. De Prins, and L. van Gool, "SVM-based nonparametric discriminant analysis, an application to face detection," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, vol. 2, pp. 1289–1296, Nice, France, October 2003.

[39] S. Kollias and D. Anastassiou, "An adaptive least squares algorithm for the efficient training of artificial neural networks," *IEEE Transactions on Circuits and Systems*, vol. 36, no. 8, pp. 1092–1101, 1989.

[40] M. H. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, 1994.

[41] L. Yin and A. Basu, "Generating realistic facial expressions with wrinkles for model-based coding," *Computer Vision and Image Understanding*, vol. 84, no. 2, pp. 201–240, 2001.

[42] M. J. Lyons, M. Haehnel, and N. Tetsutani, "The mouthesizer: a facial gesture musical interface," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*, p. 230, Los Angeles, Calif, USA, August 2001.

[43] S. Arca, P. Campadelli, and R. Lanzarotti, "An automatic feature-based face recognition system," in *Proceedings of the 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '04)*, Lisboa, Portugal, April 2004.

[44] K.-M. Lam and H. Yan, "Locating and extracting the eye in human face images," *Pattern Recognition*, vol. 29, no. 5, pp. 771–779, 1996.

[45] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[46] D. O. Gorodnichy, "On importance of nose for face tracking," in *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG '02)*, pp. 181–186, Washington, DC, USA, May 2002.

[47] S. C. Aung, R. C. K. Ngim, and S. T. Lee, "Evaluation of the laser scanner as a surface measuring tool and its accuracy compared with direct facial anthropometric measurements," *British Journal of Plastic Surgery*, vol. 48, no. 8, pp. 551–558, 1995.

[48] L. Vincent, "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[49] L. Vincent, "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[50] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems*, G. Tesauro, D. Touretzky, and T. Leen, Eds., vol. 7, pp. 231–238, The MIT Press, Cambridge, Mass, USA, 1995.

[51] V. Tresp, "Committee machines," in *Handbook for Neural Network Signal Processing*, Y. H. Hu and J.-N. Hwang, Eds., CRC Press, Boca Raton, Fla, USA, 2001.

[52] C. M. Whissel, "The dictionary of affect in language," in *Emotion: Theory, Research and Experience. The Measurement of Emotions*, R. Plutchnik and H. Kellerman, Eds., vol. 4, pp. 113–131, Academic Press, New York, NY, USA, 1989.

[53] S. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, T. P. Mailis, K. Karpouzis, and S. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," *Neural Networks*, vol. 18, no. 4, pp. 423–435, 2005.

[54] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice-Hall, Upper Saddle River, NJ, USA, 1995.

[55] M. A. Lee and H. Takagi, "Integrating design stages of fuzzy systems using genetic algorithms," in *Proceedings of the 2nd IEEE International Conference on Fuzzy Systems (FUZZY'93)*, pp. 612–617, San Francisco, Calif, USA, March-April 1993.

[56] M. Wallace and S. Kollias, "Possibilistic evaluation of extended fuzzy rules in the presence of uncertainty," in *Proceedings of the 14th IEEE International Conference on Fuzzy Systems (FUZZ '05)*, pp. 815–820, Reno, Nev, USA, May 2005.

[57] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[58] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[59] G. W. Williams, "Comparing the joint agreement of several raters with another rater," *Biometrics*, vol. 32, no. 3, pp. 619–627, 1976.

[60] V. Chalana and Y. Kim, "A methodology for evaluation of boundary detection algorithms on medical images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 5, pp. 642–652, 1997.

[61] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'Feeltrace': an instrument for recording perceived emotion in real time," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 19–24, Belfast, Northern Ireland, September 2000.

[62] D. Cristinacce and T. F. Cootes, "A comparison of shape constrained facial feature detectors," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FG '04)*, pp. 375–380, Seoul, South Korea, May 2004.