ELSEVIER

Signal Processing: *Image Communication* 18 (2003) 67–89

SIGNAL PROCESSING: IMAGE COMMUNICATION

www.elsevier.com/locate/image

A modular approach to facial feature segmentation on real sequences

George N. Votsis*, Athanasios I. Drosopoulos, Stefanos D. Kollias

Department of Electrical and Computer Engineering, National Technical University of Athens, Iroon Polytechniou 9, Zografou 15773, Athens, Greece

Received 8 February 2002; received in revised form 30 July 2002; accepted 9 September 2002

Abstract

In this paper a modular approach of gradual confidence for facial feature extraction over real video frames is presented. The problem is being dealt under general imaging conditions and soft presumptions. The proposed methodology copes with large variations in the appearance of diverse subjects, as well as of the same subject in various instances within real video sequences. Areas of the face that statistically seem to be outstanding form an initial set of regions that are likely to include information about the features of interest. Enhancement of these regions produces closed objects, which reveal—through the use of a fuzzy system—a dominant angle, i.e. the facial rotation angle. The object set is restricted using the dominant angle. An exhaustive search is performed among all candidate objects, matching a pattern that models the relative position of the eyes and the mouth. Labeling of the winner features can be used to evaluate the features extracted and provide feedback in an iterative framework. A subset of the MPEG-4 facial definition or facial animation parameter set can be obtained. This gradual feature revelation is performed under optimization for each step, producing a posteriori knowledge about the face and leading to a step-by-step visualization of the features in search.

Keywords: Facial feature extraction; Optimal segmentation; Seed growing; Active contours; Dominant angle; Feature labeling

1. Introduction

Understanding what the visual content of human face reveals has been a very active field of research for almost three decades. Researchers through time have endeavoured in various approaches on coding the content of the face (as with the MPEG-4 standard) and on using it for identification or emotion recognition reasons [6,8]. Applications have involved a wide range that meets telecommunications and human-computer interaction. Most of the times, the key issue for all approaches has been the suitable selection of features, having as an objective either facial representation or facial evaluation.

The specific problem of automatic facial feature selection falls under the general problem of automatic image segmentation. Segmentation using sole traditional low-level image processing techniques often requires interactivity in a large degree in order to achieve the desired results. On the other hand, automation frequently involves models in order to guide the segmentation

*Corresponding author.
E-mail address: stefanos@cs.ntua.gr (G.N. Votsis).

0923-5965/02/$ - see front matter
PII: S0923-5965(02)00103-0

processes. Choosing the right model set for the—as generic as possible—representation of objects is difficult due to shape complexity and variability within and across a set of objects belonging to the same class. It has been reported that for fixed objects (and the facial features could well be considered as such) it is possible to partition the image into a globally consistent interpretation through the use of deformable templates, while using statistical shape models to enforce prior probabilities on global deformations within the same class [41]. Other recent work in image segmentation includes stochastic model-based approaches [11,25,34,52,55] morphological watershed-based region growing [43], energy diffusion [28], and graph partitioning [44]. Non-model-guided segmentation aims at separating homogeneous colour-texture regions [12], but generally do not satisfy semantic partitioning.

Robust and accurate facial analysis and feature extraction has always been a complex problem that has been dealt with by posing presumptions or restrictions with respect to facial rotation and orientation, occlusion, lighting conditions and scaling. These restrictions are being eventually revoked in the literature [36,3], since authors deal more and more with realistic environments, while keeping in mind pioneering works in the field [19,40,5,21,6].

Video approaches try to combine both spatial and temporal information in order to assess the segmentation performance [9,17,3]. The use of stereo vision and special processors [29] promises fast feature tracking, given that the initial position of the features is predefined. Hierarchical feature extraction uses the combination of prior statistics and linear discriminant and complementarily aggregates multiple Gabor jet representations for a whole sequence [42,31,37]. Simpler approaches on video use heuristics for the purpose of practical implementation [26]. Parallel processing employing motion and spatial cues are combined collaboratively to prove that fast segmentation can be achieved by using only standard and complementary techniques. In an earlier work of Black et al. [3], local parameterized models of image motion had been used for recovering and recognizing the non-rigid and articulated motion of human faces. Parametric flow models were used in order to provide a concise description of facial motion (intuitively related to facial feature motion) in terms of a small number of parameters. The work dealt with facial expression understanding, but feature extraction and tracking seemed to be the input in order to reach any expression estimate.

Model-based estimation of facial features utilizes matching of either generalized or subject-dependent textured models, but requires computationally consuming actions, such as model rendering with modified shape and position [16]. Investigation for the reduction of computational complexity has been attempted by regarding model-based facial feature extraction as an optimization problem [1].

Higher-level approaches attempt to exploit a large amount of evidence, related and reinforced by model knowledge through a probabilistic framework [53]. Under this framework, feature groupings that form meaningful entities using perceptual organization are examined under the assignment of probabilities for each grouping and their reinforcement using Bayesian reasoning techniques.

The work proposed in this paper aspires to deal with more realistic environments by considering general imaging conditions and developing more structured solutions. A gradual, robust analysis of facial features is presented, coping with large variations in the appearance of diverse subjects, as well as of the same subject in various instances within real video sequences. Soft a priori assumptions are made on the pose of the face or the general location of the features in it. The gradual revelation of information concerning the face is supported under the scope of optimization, energy or error minimization for each step, producing a posteriori knowledge about it and leading to a step-by-step visualization of the features in search. This comes in contrast with the basic perspective of other solutions proposed in literature [16,47,38,1,32,45], which use specific feature representation models or presume an upright position of the face. It merges, however, some of the tools already proposed in literature with new ones, under a unified hierarchical approach, each step of which is thought of as an optimization question.

The main focus of this work is on structured facial feature extraction on already detected and segmented facial images; its contribution may be summarized as follows:

- The problem is being dealt under general imaging conditions and soft presumptions. Other existing approaches that deal with video sequences assume that in the first frame the features have already been segmented.
- Rotation-independent feature extraction yields a fuzzy mechanism for automatically detecting in-plane facial rotation.
- The objective of each step of the structured procedure is considered as a separate optimization problem.
- Integrating attractive techniques proposed in literature with new ones provides more powerful tools towards the achievement of our goal.

The remainder of the paper is organized as follows. In Section 2, some of the adopted techniques are elaborated. Their effectiveness is presented and their drawbacks under circumstances are briefly discussed. The proposed facial feature segmentation method, formulated as a step-wise optimization problem, is thoroughly discussed in Section 3. In Section 4 a series of simulation results on real sequences containing faces is considered. Conclusions are presented in Section 5.

2. Adopted technologies

2.1. Facial feature segmentation by min–max analysis

Integral projections have played an important role in the long bibliography on face recognition and feature extraction, as a tool for estimating the position of features. This technique was introduced by Kanade in his pioneering work [22] on recognition of human faces. Projections are simple to implement, while being at the same time extremely effective in determining the position of features, provided the window on which they act is suitably located to avoid misleading interferences. Kanade performed the projection analysis on a binary picture obtained by applying a Laplacian operator on the gray-level picture and by thresholding the result at a proper level. Since then, authors have used the same tool from time to time. Brunelli and Poggio [5] have performed edge projection analysis by partitioning the edge map in terms of edge directions, horizontal and vertical. Areas pointed out by the integrals formed a good first estimate of where the templates were to be applied. Other researchers found out that more exact results can be obtained by applying the projection analysis on the intensity image, because of the smooth contours of most of the facial features [38,23,45,51,33]. In [38] rubber snake models are applied on the areas where the integral projections point out. This is performed in a sequential manner analogously to the template deformation on different epochs used by Yuille et al. [54]. In [45] the authors compute the y-projection of the topographic gray-level relief. They smooth the y-relief through average filtering. Significant minima are determined in the y-relief by checking the gradients of each minimum to its neighbor maxima. Each significant minimum is considered as a possible vertical position for facial features. For each y-candidate, smoothed local x-relief is calculated and the minima reached by this procedure are candidate feature points.

In the above-mentioned works, the authors exploit the a priori knowledge that faces within images are in an upright position, by labeling the significant minima of the projections as candidate estimates of specific features or feature groups, i.e. upper group (eyebrows, eyes), middle group (nostrils), lower group (mouth chin). In more realistic environments the hypothesis of facial upright position is usually not valid, a fact that eliminates any assumption on where a feature group lies within the image. Even so, all relevant papers have shown how powerful such a simple in concept tool can be.

2.2. Unsupervised color–texture segmentation

The problem of unsupervised segmentation is ill defined, because semantic objects do not usually correspond to homogeneous spatial regions in

color or texture. Recent work in image segmentation includes stochastic model-based approaches [25,52], morphological watershed-based region growing [40], energy diffusion [28] and graph partitioning [44].

In a recent work presented by Deng et al. [12], the goal is homogeneous color–texture region segmentation. According to this work, colors in the image are quantized to several representative classes, generating a class-map, which may be viewed as a special kind of texture composition. Spatial segmentation is performed directly on this class-map without considering the corresponding pixel color similarity.

Let Z be the set of all N data points in a class-map. Let $z=(x,y)$, $z\in Z$ and m be the mean,

$$m=\frac{1}{N}\sum_{z\in Z} z. \tag{1}$$

Suppose Z is classified into C classes, Z_i, $i=1,\ldots,C$. Let m_i be the mean of the N_i data points of class Z_i,

$$m_i=\frac{1}{N_i}\sum_{z\in Z_i} z. \tag{2}$$

Let also

$$S_{\mathrm{T}}=\sum_{z\in Z}\|z-m\|^2 \tag{3}$$

and

$$S_{\mathrm{W}}=\sum_{i=1}^{C}\sum_{z\in Z_i}\|z-m_i\|^2, \tag{4}$$

where S_{T} is the total variance and S_{W} is the variance of points belonging to the same class. Let us also define

$$J=(S_{\mathrm{T}}-S_{\mathrm{W}})/S_{\mathrm{W}}. \tag{5}$$

For an image with several homogeneous regions, color classes are well separated from each other and J takes large values. If all classes are uniformly distributed over the image, J takes small values. If we calculate J over each segmented region, we may define

$$\bar{J}=\frac{1}{N}\sum_{k} M_k J_k, \tag{6}$$

where J_k is calculated over neighborhood k, M_k is the number of points in neighborhood k and N is the total number of points in the class-map. According to Deng et al. [12], $\bar{J}$ is an expression of an energy function, which has proven to be a good criterion to be minimized over all possible ways of segmenting the image, given the number of regions. By constructing an image whose pixel values correspond to J values calculated over small windows centered at the pixels, the so-called J-images are built. The higher the J value is, the more likely that the corresponding pixel lies near a region boundary.

This new criterion for color–texture segmentation is complex enough to require offline processing for medium quality images. However, if applied in better-specified segmentation scenarios, the procedure may be significantly less computationally consuming.

2.3. Segmentation and active contours

In the general field of segmentation as well as in facial feature extraction, deformable models and active contours (snakes) have often been used [7,10,20,24,30,46,3,47,38]. For example, Yuille et al. [54] employ deformable templates to model facial features. The template-based approach allows for inclusion of object-specific knowledge in the model. Nevertheless, such methods require careful construction and parameterization of templates.

Snakes in general incorporate knowledge about a contour's smoothness and resistance to deformation. A regular estimate of a contour is obtained by defining image forces that pull on the snake model, while at the same time intrinsic contracting or inflating forces can be used in order to either shrink or expand the snake, respectively, towards directions that are irrelevant to the image content. One of the chief virtues of snake representations is that it is possible to specify a wide range of snake properties through its energy function, in analogy with physical systems. Controlling a snake causes it to evolve as to reduce its energy. By specifying an appropriate energy function, we can make a snake that will evolve to have particular properties.

One of the most common forms of the energy function representing a snake is the following:

$$\begin{aligned} E_{\text{snake}} &= E_{\text{int}} + E_{\text{ext}} \\ &= E_{\text{elastic}} + E_{\text{image}} \\ &= K_1 \sum_{i=1}^{N} (d(i, i-1))^2 + K_2 \sum_{i=1}^{N} I(x_i, y_i), \quad (7) \end{aligned}$$

where by $d(i, i-1)$ we denote the distance between two successive control points and by $I(x_i, y_i)$ we denote the intensity at pixel i. The constants K_1 and K_2 are arbitrarily selected and control the influence that each of the two factors has.

The internal or elastic energy is the part that depends on intrinsic properties of the snake, such as its length or curvature. The external or image energy generally depends on image structure and particular constraints the user wants to impose. Each of the energies corresponds to a related force, a fact that allows for a simple implementation of the dynamics of the snake. At each step, each control point moves by an amount proportional to the force acting on it. In the physical analogy, this is like making a light snake move through a viscous fluid—it should dissipate its energy without oscillating. These forces are used to move the control points and are computed based on gradient descent of each of the two energy terms in Eq. (7).

Speeding up snake evolution could be achieved through dynamic programming techniques, by calculating directly the configuration of the snake that will cause the internal forces to balance a given set of external forces. This allows bigger steps to be taken, and is more efficient overall, despite the extra computation involved.

In the seminal paper on snakes [24], snakes are regarded as a "power assist" for a human operator needing to measure structures in images. The operator would point the snake at, say, particular cells in a histological image, or at a road in a satellite image, and the snake would lock on to it and provide an accurate measure of its shape. In our case, as it will be described in the following, snakes will need to be initialized automatically in order to finally settle on the objects that will serve as potential features.

3. The proposed approach

3.1. System overview

According to our approach, primary facial features, such as the eyes and the mouth, are considered as major discontinuities on a segmented, arbitrarily rotated face. Due to arbitrary facial rotation, as well as to occlusion, labeling of those features requires careful selection among all possible discontinuities. Selection is based on criteria that give birth to a posteriori knowledge about the face, such as dominant angles of the candidate features and symmetry considerations.

Fig. 1 shows the proposed approach. The pipeline of processing steps mainly consists of three stages: (i) optimized segmentation and extraction of main facial features, (ii) estimation of the dominant facial angle based on the extracted features, and (iii) extracted feature labeling. Stage (i) generates an initial estimate of the facial segments (seeds), which it uses as input to an iterative optimization scheme that provides optimal seed estimates. The stage also includes a post-processing enhancement task (also expressed as an energy minimization problem), which assists in obtaining closed segments and removing some of the created artefacts. The following stage (ii) targets another issue, i.e., determination of the dominant facial angle, for in-plane rotation (pose) estimation. It takes as input the results (seeds) of stage (i), models them by active contours, and computes, using a fuzzy system, the dominant facial angle. Stage (iii) follows stage (ii), adopting the obtained results (i.e., the dominant angle), and uses knowledge about the facial position of mouth and eyes to label the extracted features and evaluate them. Evaluation is based on knowledge about the symmetries and relative positions of features in the facial area, making it possible to iterate on stages (ii) and (iii), so as to make the selection of dominant angle and feature labeling more accurate. MPEG-4 feature definition (FDP) and animation (FAP) parameters can be finally extracted from the proposed three-stage procedure.

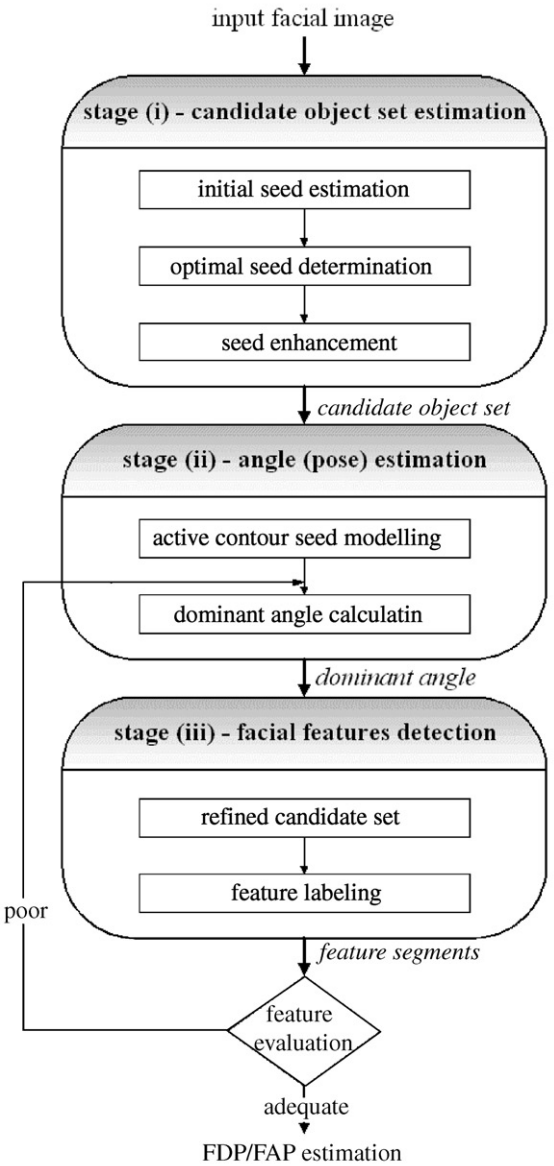

Fig. 1. Schema representing the method followed in our approach.

In the following, we describe the steps of our approach one by one.

3.2. Segmentation and seed determination

The input to this module is an image $\underline{X}$ of size $k \times l$ containing a face segmented from its background, i.e. a masked facial image. This image is assumed to depict only information that lies within the facial closed boundary, defined by the visible facial skin's perimeter. Typical examples are given in Fig. 2. Initial processing is based on pure statistical analysis. The assumption under which the module works is simple: all facial features in the input image are viewed as discontinuities upon a generally smooth surface. Of course, one has to deal with irregularities of this surface, such as illumination disparities depending on the direction of the light source in relevance with the object's position, occlusion cases, where other objects' parts hide part of the object of interest, as well as various rotation and orientation angles of the face which change the shape and the distribution of the features upon it. Occlusion is definitely a complex matter, which requires content knowledge. Such cases are being tackled through the existence of a symmetry control mechanism, which is discussed later on in this paper and uses knowledge of the already traced and appointed facial features.

Statistical analysis of facial images—as subsequently described—presents a significant advantage compared to other feature tracking approaches: it does not require knowledge of feature allocation upon a face in the first place, a fact that greatly disassociates it from head rotation and orientation variations. This means that significant characteristics of the face are first detected, even if at this stage this module "knows" nothing about their labeling.

In this context, combination of fundamental image processing tools was used to determine areas of the face that probably contain regions of interest. Edge detection using Sobel operators was applied on the masked facial image and the resulting vertical and horizontal integrals were calculated. Fig. 3 illustrates the results on some of the subjects of the used natural video database.

Under the same rationale, the inverse grayscale information of the masked face was both vertically and horizontally integrated, as seen in Fig. 4. Inversion was deemed useful both for comparison and for combination with the respective graphs depicted in Fig. 3. The projections in both cases

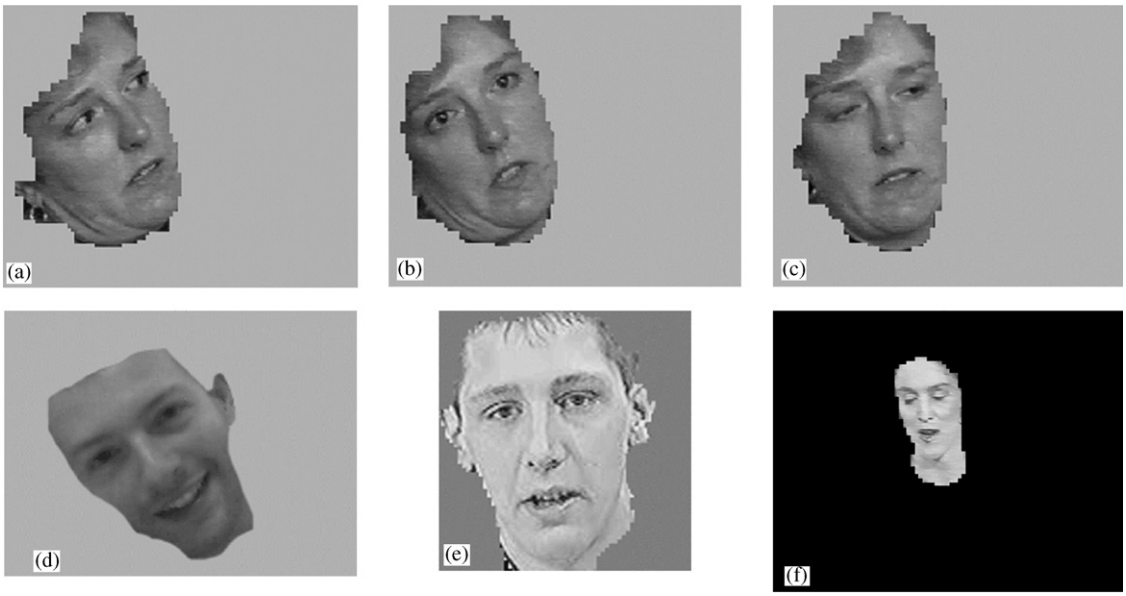

Fig. 2. Six representative input frames depicting subjects of the natural video database. The samples illustrate the variability in pose (rotation and orientation), lighting conditions and scaling.

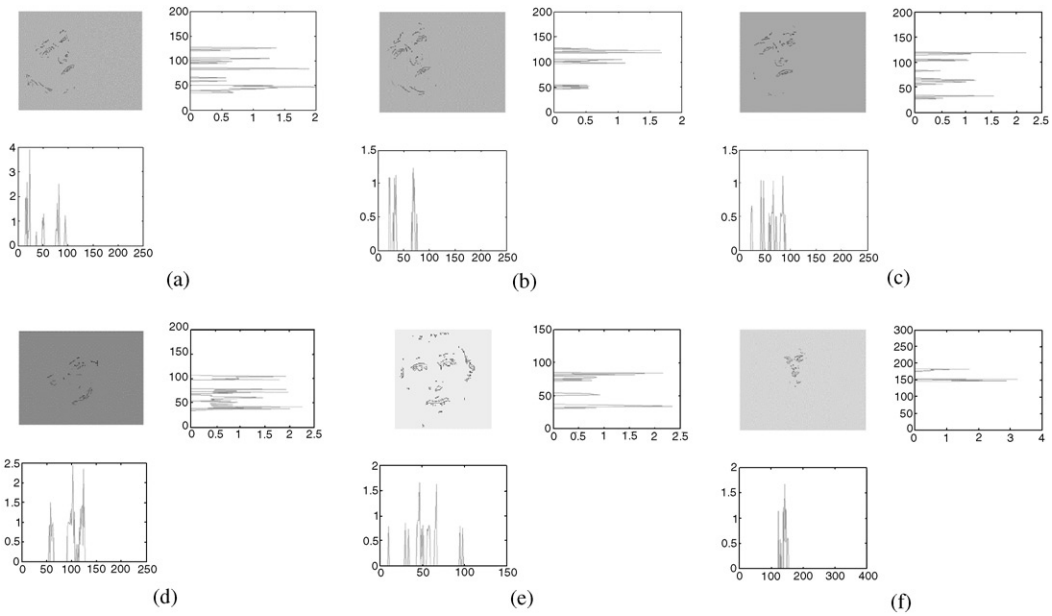

Fig. 3. Vertical and horizontal integral projections for edge image, in our six representative input examples.

point out—in a strict and a wider sense, respectively—the one-dimensional (1D) windows where one may detect the regions of interest.

Some of the information that the raw edge image includes is irrelevant to the characteristic features. So, the plots of Fig. 3 were weighted,

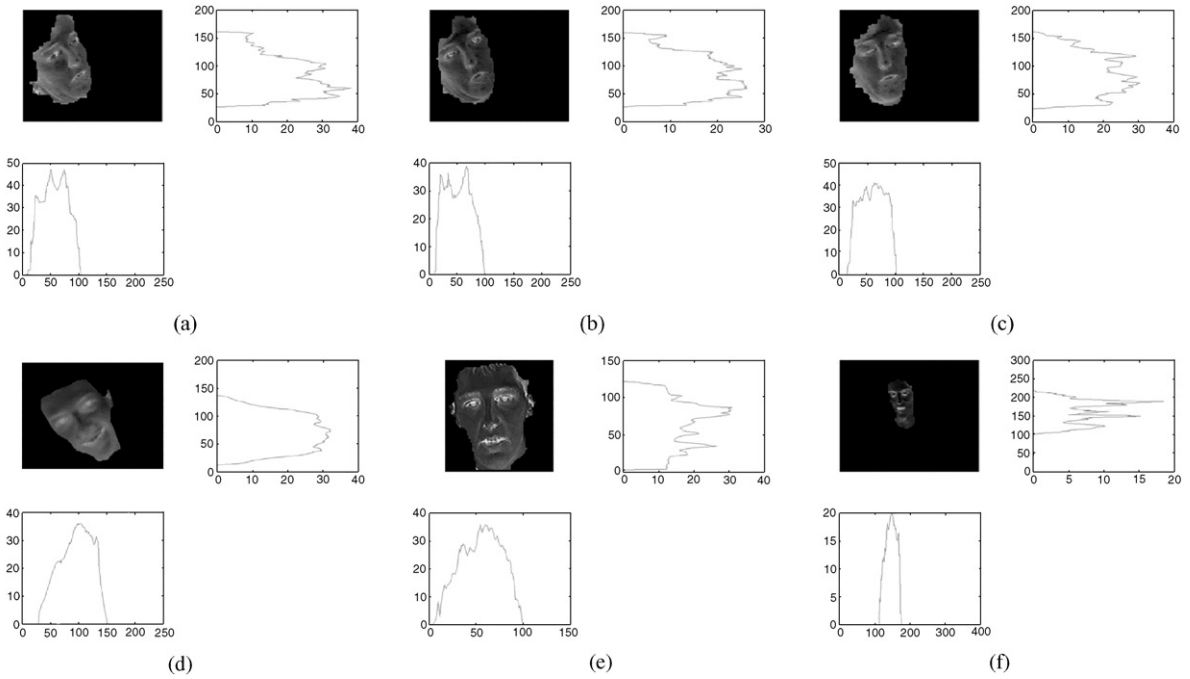

Fig. 4. Vertical and horizontal integral projections for inverse grayscale image, in our six representative input examples.

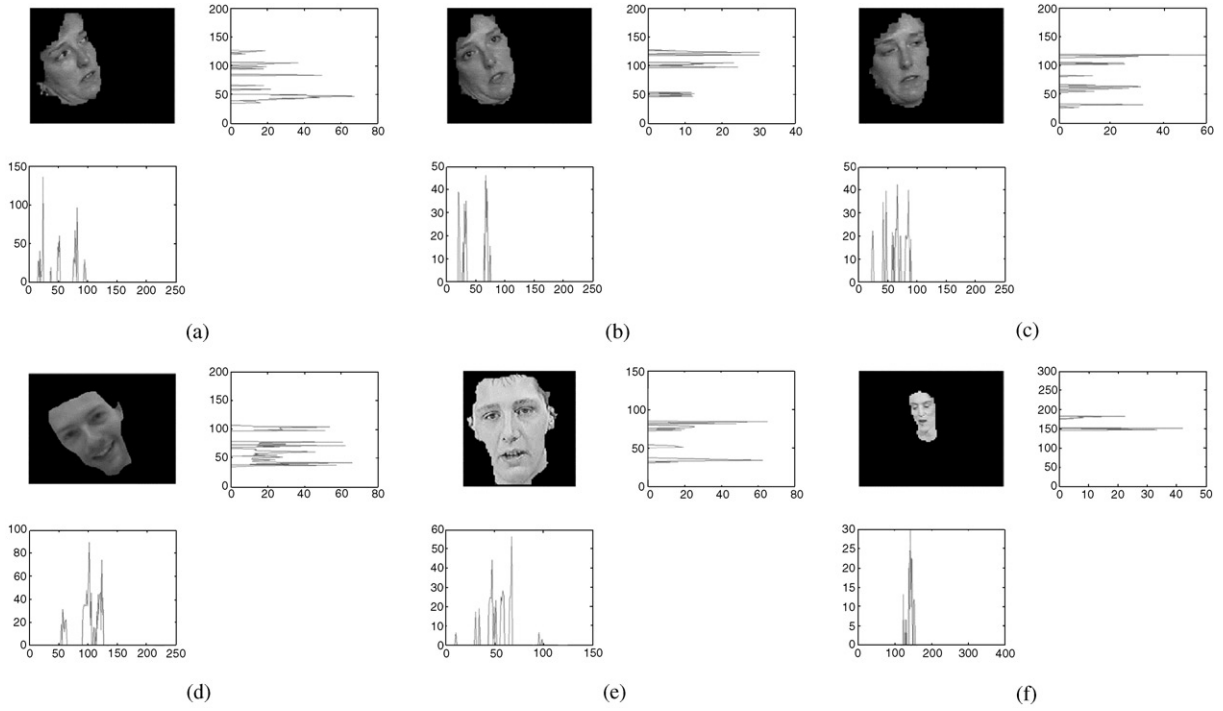

Fig. 5. Vertical and horizontal integral projections for weighted combination, in our six representative input examples.

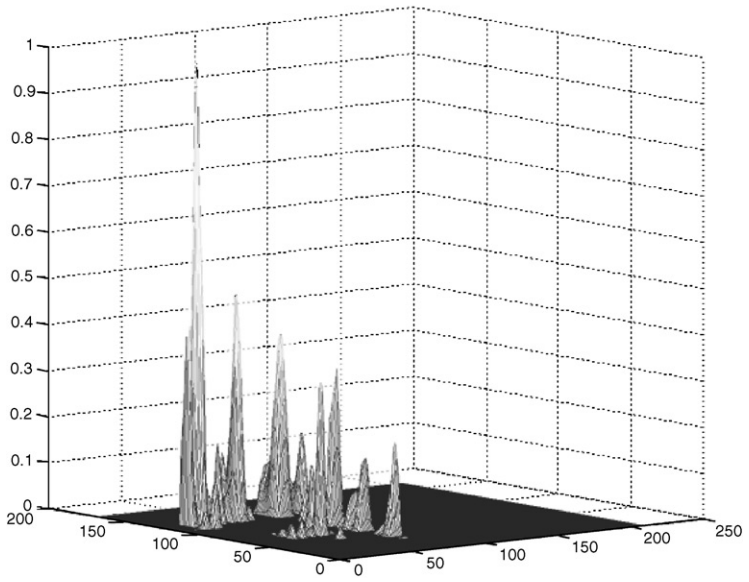

Fig. 6. The product of the vertical and horizontal weighted vectors yields a map M, where one may discern the valleys and the hills (pixel accumulations) that form different "objects".

using as envelopes the respective plots viewed in Fig. 4. This kind of modulation (implemented through multiplication of the respective curves) was adopted, because it refines the windows in both horizontal and vertical directions, also reducing the noise that has occurred by edge detection. Results are shown in Fig. 5.

The resulting weighted windows, $\underline{P_{\mathrm{h}}}$ and $\underline{P_{\mathrm{v}}}$ for the horizontal and vertical projections, respectively, provide 1D clues on where the regions of interest within the image lie. The next step postulates expansion of the windows in two dimensions, as well as thresholding and binarization of the output's values. To achieve this expansion, the product of the vertical and horizontal weighted vectors was created, yielding a map M, as the one depicted in Fig. 6.

$$M = \underline{P_{\mathrm{v}}^T} \cdot \underline{P_{\mathrm{h}}}. \tag{8}$$

On this map we may discern pixel accumulations, according to their M-values, forming different "objects".

In the following, and for each "object" we identify its respective "seed" as its closed subset with maximum area and minimum hue variance. An initial estimate of the seeds $\underline{M_0}$, is the binary map that comes from M, where $M(i,j) > 0$ for those areas that include discontinuities, as seen in Fig. 7. We assume that, based on the former steps, this initial estimate includes all possible seed candidates. The above concept is expressed in mathematical formalism as follows, where for simplicity, we consider $\underline{M_0}$ to contain only one object.

Let $\underline{S}$ be a vector containing the probability that each pixel

$$x_i \in \underline{X},\ \ i \in [1, \ldots, k \times l] : m_0^i = 1 \tag{9}$$

belongs to the seed or not,

$$\underline{S} = [s(x_1), s(x_2), \ldots, s(x_N)], \tag{10}$$

where m_0^i is the value of the ith element of $\underline{M_0}$, N is the number of the 1-valued pixels of $\underline{M_0}$ and $s(x_i) = p^i$ is the probability that pixel x_i belongs to the seed.

For a given image $\underline{X}$ (with a specific hue component, as in the examples given in Fig. 8) and an initial estimation $\underline{M_0}$ of the seed within it, our target is to find the probability vector $\underline{S}$, as well as a finer estimate $\underline{M_{\mathrm{f}}}$ of the seed. This may be mathematically expressed as a maximization

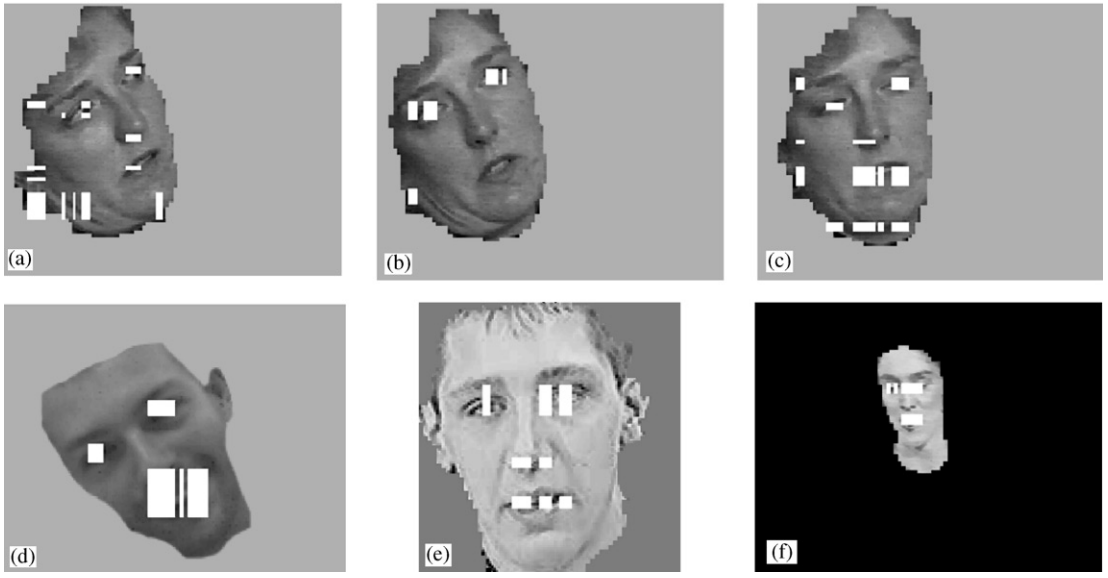

Fig. 7. Initial estimates $\underline{M}_0$ of the seeds in our six representative examples.

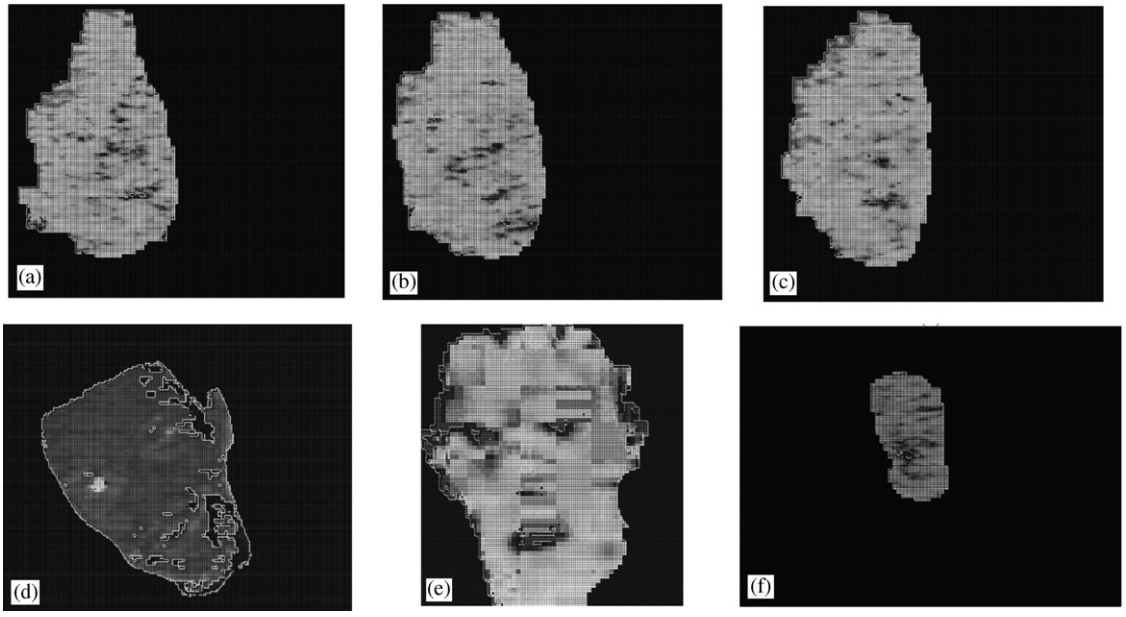

Fig. 8. Hue component (from the Hue-Saturation-Value representation) for our six input examples.

problem of a conditional likelihood function,

$$\{\hat{\underline{M}}_{\mathrm{f}}, \hat{\underline{S}}\} = \arg\max_{\underline{S}} L(\underline{M}_{\mathrm{f}}, \underline{S}/\underline{M}_0, \underline{X}), \tag{11}$$

where L denotes the log-likelihood function.

The probability density included in the above expression may be written, given the fact that the input image is given,

$$\begin{aligned}\Pr(\underline{M}_{\mathrm{f}}, \underline{S}/\underline{M}_0) &= \Pr(\underline{S}/\underline{M}_0)\Pr(\underline{M}_{\mathrm{f}}/\underline{M}_0, \underline{S}) \\ &= \Pr(\underline{S}/\underline{M}_0)\Pr(\underline{M}_{\mathrm{f}}/\underline{S}),\end{aligned} \tag{12}$$

where the fine estimate of the seed $\underline{M}_f$ depends explicitly on the probability vector $\underline{S}$. However, since we are interested in maximization of the likelihood function, the second factor of the above expression takes the value 1 only when $\underline{S} \geqslant \underline{T}$, where $\underline{T}$ is an appropriate threshold vector. This implies that the factor may be omitted, in the following sense:

$$\begin{aligned}\hat{\underline{S}} &= \arg\max_{\underline{S}}\{\log \Pr(\underline{S}/\underline{M}_0)\}\\ &= \arg\max_{\underline{S}}\{\log \Pr(\underline{M}_0/\underline{S}) - \log \Pr(\underline{M}_0)\\ &\qquad + \log \Pr(\underline{S})\}\\ &= \arg\min_{\underline{S}}\{-\log \Pr(\underline{M}_0/\underline{S}) - \log \Pr(\underline{S})\}. \end{aligned} \tag{13}$$

By assuming that for a given probability vector $\underline{S}$ there is only one initial estimate $\underline{M}_0$, there exists a matrix $\boldsymbol{R}$ that linearly relates the two vectors. This allows us to omit the first of the two terms under the condition of minimization, due to the fact that the related probability is 1 if the linear relation is fulfilled and 0 otherwise [13]. Such a simplification means that the only term we have to minimize for a given input is the second one.

On the one hand $\underline{S}$ is a vector random variable (RV), which is actually a function that depends on the hue component of the input image on a neighborhood of pixels, as well as on the area defined by that neighborhood. To be more precise, the function, which we want to minimize, is proportional to the hue variance of the neighborhood and inversely proportional to the neighborhood area. This function is an RV, since it has the following required properties [35]:

- Its domain includes the range of the variable x_i, which is also considered as an RV.
- It is a Baire function, since for every s, the set of x_i, R_s, such that $s(x_i) \leqslant s$ consists of the union and intersection of a countable number of intervals (on the x-axis).
- The events $s(x_i) = \pm\infty$ have zero probability.

On the other hand, the elements of $\underline{S}$ represent the probabilities that a picture element x_i is classified in one of two classes, i.e. to a seed or not. It is known, however, that pixels are strongly correlated to their spatial neighbors. This fact has as a consequence local correlation among neighboring elements of $\underline{S}$. This property characterizes an MRF [4,18].

These two observations allow us to use a Gibbs distribution in order to model $\Pr(\underline{S})$, since this kind of distribution is able to explicitly express MRFs in probabilistic frameworks,

$$\Pr(\underline{S}) = F\exp\left\{-a\sum_{n\in N} G_n(x)\right\}, \tag{14}$$

where F and a are normalization factors, n are the neighborhoods, N is their union, and $G_n(x)$ should be expressed in terms of hue variance minimization and area maximization within each neighborhood. We chose

$$G_n(x) = \sum_{i\in n}\{(H(x_i) - \bar{H}_n)^2 + (s(x_i) - m_0(x_i))^2\}, \tag{15}$$

where i is the index for each pixel, H is the hue component for the image $\underline{X}$ and $\overline{H_n}$ is the mean hue for neighborhood n. The first term of the sum models hue variance in neighborhood n. The second term is minimal in the case that neighborhood's probabilities remain as high as possible, meaning that the neighborhood remains as big as possible (maximum area).

From Eqs. (13)–(15), we may find

$$\begin{aligned}\hat{\underline{S}} &= \arg\max_{\underline{S}}\{\log \Pr(\underline{S})\}\\ &= \arg\min_{\underline{S}}\left\{\sum_{n\in N}\sum_{i\in n}\{(H(x_i) - \overline{H_n})^2\right.\\ &\qquad \left. + (s(x_i) - m_0(x_i))^2\}\right\}. \end{aligned} \tag{16}$$

Since both factors of the above expression are of squared form, (16) defines a convex function (see [27, p. 178]), a fact that implies that a global solution is present. Minimization of (16) is straightforward, if we calculate the derivative of the expression in (16),

$$\begin{aligned} D &= \frac{\partial\{\sum_{n\in N}\sum_{i\in n}\{(H(x_i) - \bar{H}_n)^2 + (s(x_i) - m_0(x_i))^2\}\}}{\partial s}\\ &= \sum_{n\in N}\sum_{i\in n}\left\{2(H(x_i) - \overline{H_n})\frac{\partial H(x_i)}{\partial s}\right.\\ &\qquad \left. + 2(s(x_i) - m_0(x_i))\right\}, \end{aligned} \tag{17}$$

where $\partial H(x_i)/\partial s$ is either zero in the case where the value of $\underline{M_f}$ at pixel x_i is the same from state j to state $j+1$, or $-H(x_i)$ in the case where this value changes. Through Eq. (17) we observe that there is indeed a linear relation between $\underline{S}$ and $\underline{M_0}$, as assumed before. Due to the neighborhood structure of our problem, the iterative conditional modes (ICM's) technique [13] has been used to estimate the optimal vector $\underline{\hat{S}}$. The respective optimal vector $\underline{\hat{M}_f}$ is estimated through $\underline{\hat{S}}$ with the selection of an appropriate threshold vector, as mentioned before.

Examples of the output are shown in Fig. 9.

3.3. Enhancement and seed growing

The output of the previous module provides a minimal estimation of the facial features' positions. The goal of the current module is to enhance the occurring pixel set. In determining the seeds of our objects of interest, the hue component of the input image was merely used to locally define those maximal areas where hue variance was minimal. As hue represents the chromatic component (H) of an image, a region-based enhancement for each one of the minimal objects tracked in the first module is applied. On the other hand, the value component (V) has a significant amount of information to offer in this stage. Of all the edge information that is gathered from V, only the lines that partially overlap with the seeds are kept. This kind of enhancement results in objects with holes, with no concave shape and which do not accurately describe human facial features. Moreover, the existence of noise is still evident. For these reasons, a series of morphological operators are used to provide further smoothing of the located regions; these include majority morphing (setting a pixel to 1 if five or more pixels in its 3×3 neighborhood are 1's) and filling of gaps (i.e., isolated interior pixels with a value of 0 surrounded by 1's) for the removal of noise and the determination of closed objects, as well as bridging, for retaining connectivity of objects that originally seem to be separate and extremely close to each other, but in fact comprise the same object. Finally, successive application of such operators results in superfluous perimeter enhancement, which calls for perimeter erosion. The kernel for all morphological operations is automatically adapted according to object size and complexity, with the scaling factor being empirically selected,

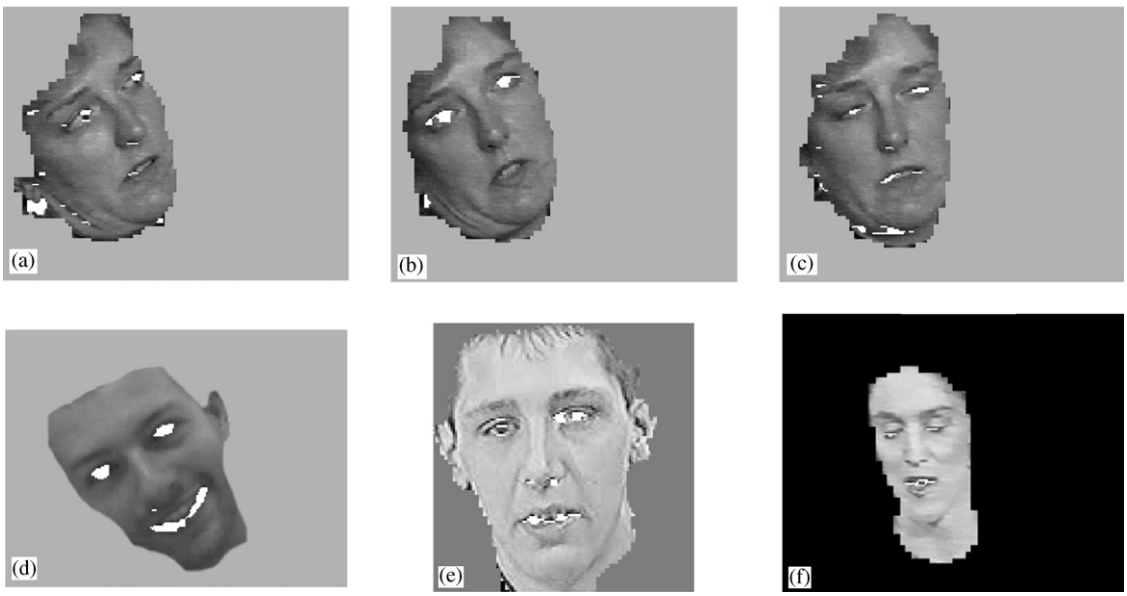

Fig. 9. Fine estimates $\underline{M_f}$ of the seeds in our six representative examples.

based on the number of practiced morphological operators.

The enhancement procedure is viewed as an energy minimization problem of the terms E_H, E_{VM}, where the first term refers to hue enhancement, while the second term is related to value and morphological enhancement, as described above in principle.

Hue enhancement is based on the idea of energy minimization proposed by Deng et al. [12], only that in our case there are only two classes (i.e. one class of interest) and the computational complexity is considerably reduced due to the fact that:

- calculations are restricted in neighborhoods that are well specified by the seeds,
- color reduction and quantization is easily achieved, since for each seed there is a dominant color.

Consequently, the first term of the enhancement energy takes the following form:

$$E_H = \frac{1}{N} \sum_{k \in W} M_k J_k, \tag{18}$$

where the summation is performed over neighborhoods k, i.e. windows W defined around the seeds determined in the previous module. N is the total number of points in the W-masked class-map. Another essential modification to the approach proposed by Deng et al. is that for each neighborhood, J_k is calculated over different classifications. To be more precise, the classes Z_i, $i = 1, 2$, as presented in (2), are different for each k and depend on the respective seed's dominant color's presence and absence. So, (18) is minimized over all possible extensions for the hue components, as well as for multiple extension window sizes, which are adapted for each seed by its area and structure.

The value–morphological enhancement term E_{VM} is defined as a seed/enhancement overlapping model,

$$E_{VM} = \frac{A_S A_{VM}}{A_O^2}, \tag{19}$$

where A_S is the area of the seed, A_{VM} is the area of the value–morphological enhancement and A_O is the overlapping area. As the common area between the two regions gets larger, E_{VM} gets lesser. Minimization of this ratio maximizes the overlap between the seed and the intersecting value/morphological enhancement object. No threshold is applied, due to the fact that the segmentation seed could be significantly smaller than the enhancement combination. Convergence of this term is based on the assumption of value connectivity for each feature in search.

The two enhancement procedures are independent. The final enhanced seeds are computed as the logical OR of the regions (objects) provided independently by the two minimization procedures.

Some results are shown in Fig. 10.

3.4. Object determination and dominant angle calculation

The processing steps that have been utilized up to this stage yield a set of objects that are candidates for the desired features within the input facial image. Let us denote by

$$S_c = \{O_c : (O_c \in F) \vee (O_c \notin F)\} \tag{20}$$

the *candidate object set*, where the *candidate object* O_c may or may not belong to the desired *feature set* F. Each object O_c may be modeled by an active contour (snake), which is able to describe its shape as well as its direction in the 2D space. The active contour model consists of a set of control points, connected by straight lines forming closed loops. For each O_c, an initial snake is estimated by the rectangle including the object in the image, as seen in Fig. 11. Automatic snake initialization is accomplished in this manner. Approximation of the actual object is achieved by combining two kinds of energy: the so-called *internal* or *elastic* and the *external* or *gradient* [24], as described earlier. Sequential minimization of these terms is applied, for a number of iterations that depends on the object's size and structural complexity. For our examples, the final active contour models for all objects O_c may be seen in Fig. 11.

As mentioned above, the active contour models are directional by nature. This means that for each O_c we may define an angle θ_i that reveals the object's orientation in the 2D space. The dominant angle θ_d is computed using a fuzzy subsystem. For

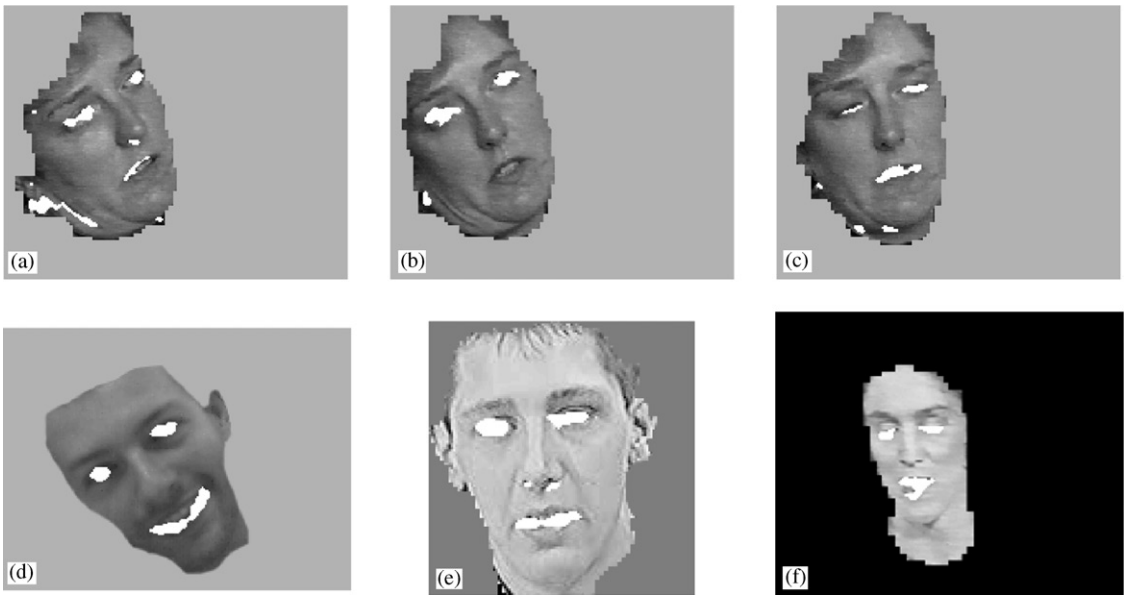

Fig. 10. Enhanced seeds for the examined input images.

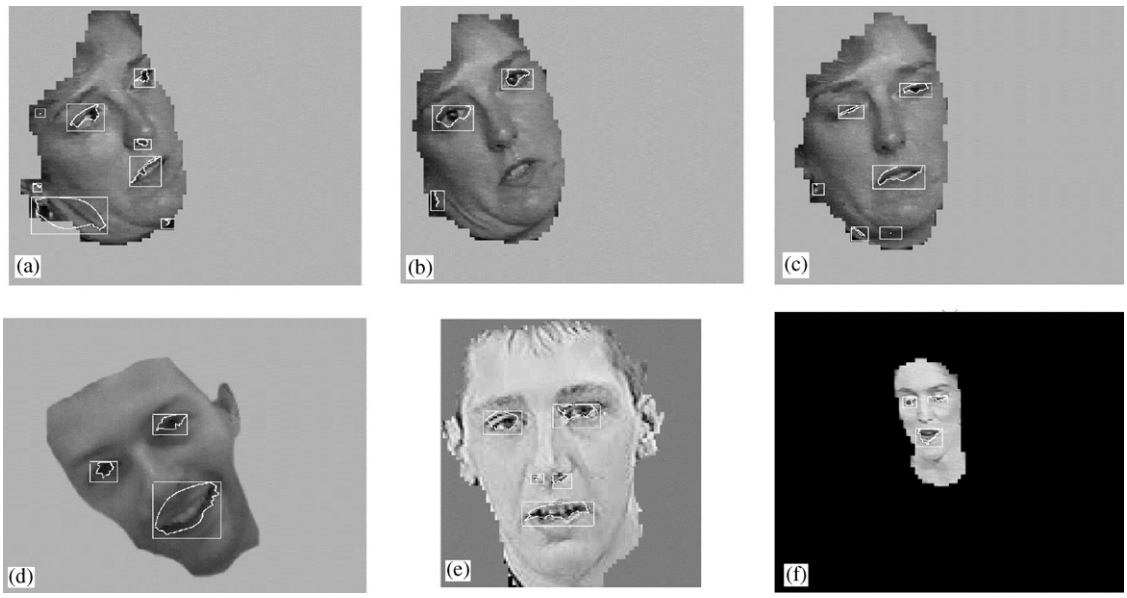

Fig. 11. Initializations and final approximations of active contours for all enhanced seeds within each input image.

each object O_c, its angle θ_i is assigned a triangular membership function, whose amplitude is proportional to the size of the object. The summation of all the membership functions over the θ axis usually shows an accumulation around a specific angle. So, the angle that obtains the maximum amplitude is the one we call θ_d. The same problem may be viewed as an error minimization issue,

$$E_{\text{ang}} = \sum_{i=1}^{N_O} [C - f_m(O_i)]F_O, \tag{21}$$

where C is a constant, N_O is the number of objects in the set S_c, f_m is the selected membership function for objects O_i and corresponding snake angles, while F_O is an outlier factor. This error reflects the selection of the fuzzy system (accumulative membership function maximization) including the outlier exclusion factor. For a given membership function f_m, it depends on the snake angle and the amplitude of f_m for the object O_i. We chose to use a triangular membership function (which could also be Gaussian) with amplitude that is linearly proportional to the object area, in order to express a degree of uncertainty around the angle estimate, which is automatically calculated for each candidate object. The scalar outlier factor included depends on the object's relative size.

The angle that minimizes (21) is used in order to rotate the segmented faces. Results on dominant angle estimation and on facial rotation are illustrated in Figs. 12 and 13, respectively.

3.5. Basic feature selection and feature labeling

The dominant angle serves as a criterion for further restriction of the set S_c. Using the standard deviation of the angle distribution that our fuzzy system yields, a range of values around θ_d is defined, which provides this restriction. Objects that lie within this angle window form the new set. Let us denote by

$$S'_c = \{O'_c : (O'_c \in F) \vee (O'_c \notin F)\} \tag{22}$$

the refined candidate set. Some results on the test images examined are shown in Fig. 14.

Up to this stage of the algorithm, we have made soft assumptions on the nature of the features and we have used no a priori knowledge about the specific or relative positions of the features in the facial area. Feature labeling is attempted at this point for the three basic facial features, i.e. the eyes and the mouth. These are considered to be sufficient to form the basis for further feature extraction, since they may provide a posteriori knowledge on the real position of other facial features, such as eyebrows, nose and nostrils, as well as transient features (nasolabial furrows, nose wrinkles, eyebrow wrinkles) [47], if these are visible. In this attempt, we employ vague knowledge of the geometric interdependence that these features have on a rotation-wise normalized face, as seen in Fig. 15.

The normalized triplet of candidates that is closer to the geometric pattern expressed in (23) and illustrated in Fig. 14 is considered to be the winning triplet,

$$E_{sym} = (|a| - |b|) \cdot |90^\circ - \phi| / \max(|a|, |b|). \tag{23}$$

This triplet is estimated by examining the number of possible triplets, which is

$$\binom{n}{3} = \frac{n!}{3! \cdot (n-3)!}, \tag{24}$$

where n is the dimension of S'_c.

The case, however, where all three features are not present, is possible. This could be due to occlusion, or failure of the algorithm to include all of them in S'_c, or even due to dimension insufficiency of the set S'_c ($n<3$). In the case where $n=2$, or generally in the case of couple candidates, when these lie in the same parallel line with the horizontal axis (after rotation normalization), even if small deviations are present, they may be labeled as the eyes. For $n=1$ and for unit candidates, there can be no sound conclusion on the feature's label. Moreover, when $n=1$, the question of how one feature imposes a dominant angle arises. Both in this case and in the case of only two detected features, we can automatically repeat the whole process from the stage of determining a new dominant angle and forth, by enhancing the contribution of any correctly labeled features or decreasing the respective contribution of solely detected ones. It should also be mentioned that other, e.g. semi-automatic, evaluation of the features can be used to lead to such iterative implementation of the procedure. Some results of feature labeling are shown in Fig. 16.

3.6. Feature definition or animation parameter estimation

Once the winner objects are labeled, a subset of the facial definition (FDP) or animation (FAP) parameter sets (as defined in the MPEG-4 standard) can be calculated. FDP extraction is

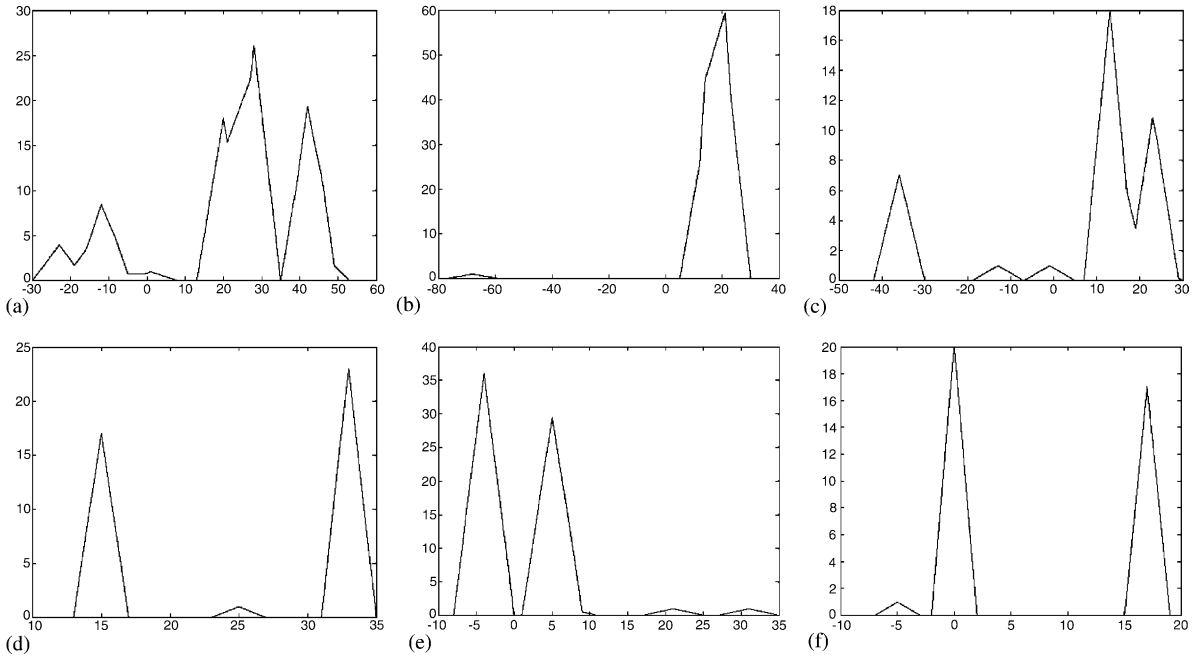

Fig. 12. Output of the fuzzy system that determines the dominant angle. Triangular membership functions for each object are added, forming accumulations around the dominant angle θ_d.

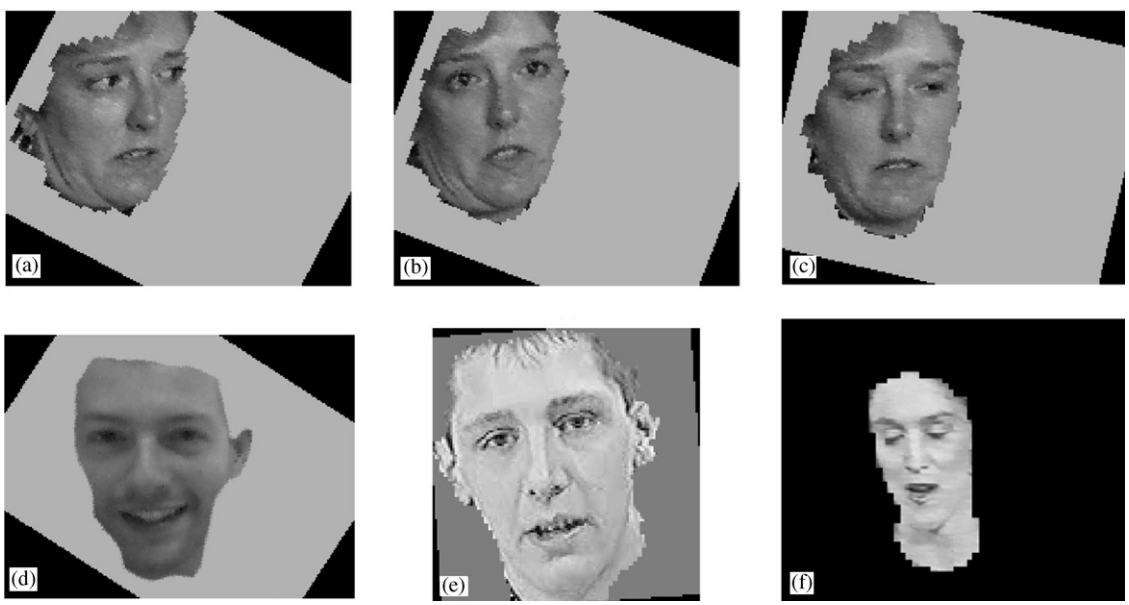

Fig. 13. Rotated input images, according to the automatically calculated dominant angles θ_d for each case.

based on the computation of minima and maxima according to the orientation of the dominant angle. After specifying the coordinates of the detected facial points we can measure their displacements between two successive frames. The measured values can be modeled through

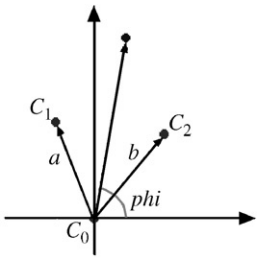

Fig. 14. Normalized geometric pattern for eyes and mouth.

FAPs and, e.g., fed to an MPEG-4 decoder. The labeled FDP points for some test frames may be seen in Fig. 17.

4. Simulation results

Each of the aforementioned optimization procedures was applied on real video frames, where the facial segmentation had already been performed [2,48].

The system was implemented on a PC with a Pentium III processor and for the simulations we used MATLAB language. All performance statistics are reported for non-optimized code. To support the system's efficiency, the code was run for over 100 cases coming from various natural sources of emotionally rich material: pictures, lab recordings, or even broadcast shows [8].

Figs. 3–9 show all the stages of the seed determination phase for a number of characteristic frames. The optimization of the initial estimates yields the fine maps $\underline{M_f}$ that may be viewed in Fig. 9 for our example images. Fig. 10 illustrates seed growing as a result of expansion energy minimization, as mathematically described in (18)–(19). In Fig. 11 snake initializations and final snake estimates may be seen. Figs. 12 and 13 report the estimation of the dominant angles, i.e. the facial rotation angles for each of the presented examples. Refined object sets S_c' and the normalized winner triplets are depicted in Fig. 15. The primary features are selected and labeled in Fig. 16. Finally, resulting feature points are calculated as in Fig. 17, and form a subset of the MPEG-4 FDP set.

Automatically obtained results were compared to manually extracted and labeled ones on the input material. In order to measure the consistency of the results, a series of metrics was applied for each feature as well as on the assemblage of all the features, in both cases (automatic and manual). To be more precise, the Euclidean distance was computed between the two cases for each feature point, for the centroid of each feature, as well as for the relative position of all centroids. Presence of artefact features or absence of existing ones was taken into account. These three levels of measures and the outlier factor were jointly employed so as to estimate the success or the failure of the procedure for all examined frames. Results show that only a 4% of the feature sets was totally mislabeled. It is deemed that this failure is mainly due to imprecise facial segmentation, meaning that paraphernalia not expected to be present in the input image actually were. For 7% of the input images the feature sets were not detected at all. In this case, the system proved its attribute of realizing its inability to come up with a confident solution, by presenting no solution at all. This was a result of the symmetry mechanism as described in the above. For the rest 89% of the input images, almost 9% of the features (mouth) were not detected while present, a fact that is again mainly due to the system's uncertainty on whether the missing feature actually belongs to the sought feature set. Only a 1% of the features were incorrectly placed, while the rest of the features were acceptably detected. The efficiency of the automatic in-plane rotation mechanism was assessed separately. Failure percentages for the estimation of the rotation angle go along with mislabeling percentages, i.e. 4%, the reason being the same. These results may be viewed in Table 1.

The focus of the proposed approach is on facial feature extraction from already detected and segmented facial images. However, problems can exist, especially when real-time simple, e.g. color-based, face detection methods are used, causing missing facial features and facial areas, imprecise external contours, disconnected regions, or inclusion of hair/background in the facial area. Missing facial areas and features can be evaluated in the feature labeling stage. Imprecise contours,

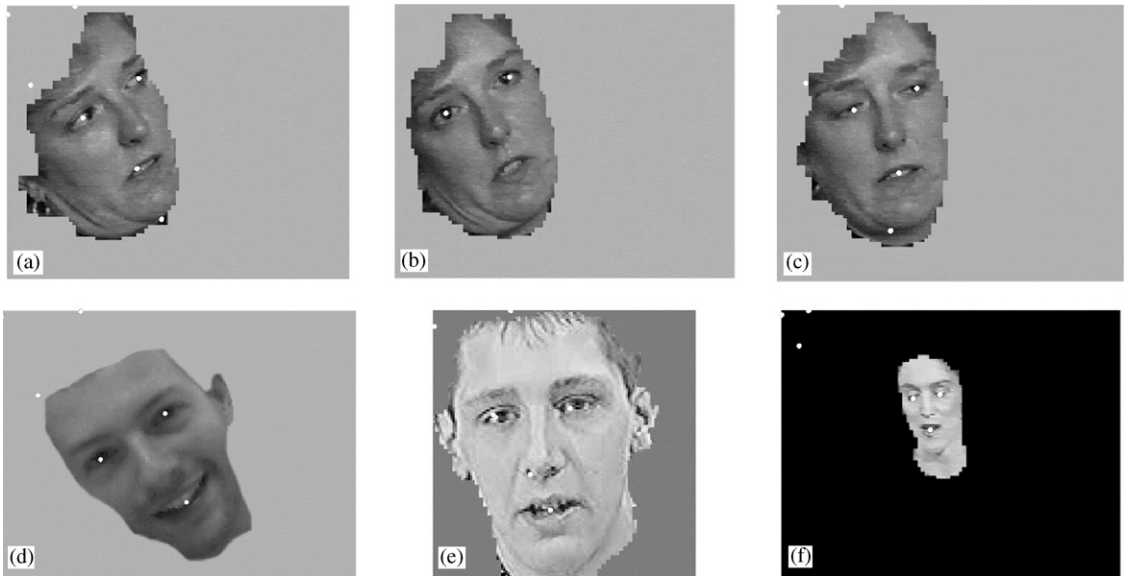

Fig. 15. The refined candidate object set S'_c. In the upper part of each image, one may observe the winner triplet (where present), rotated by θ_d.

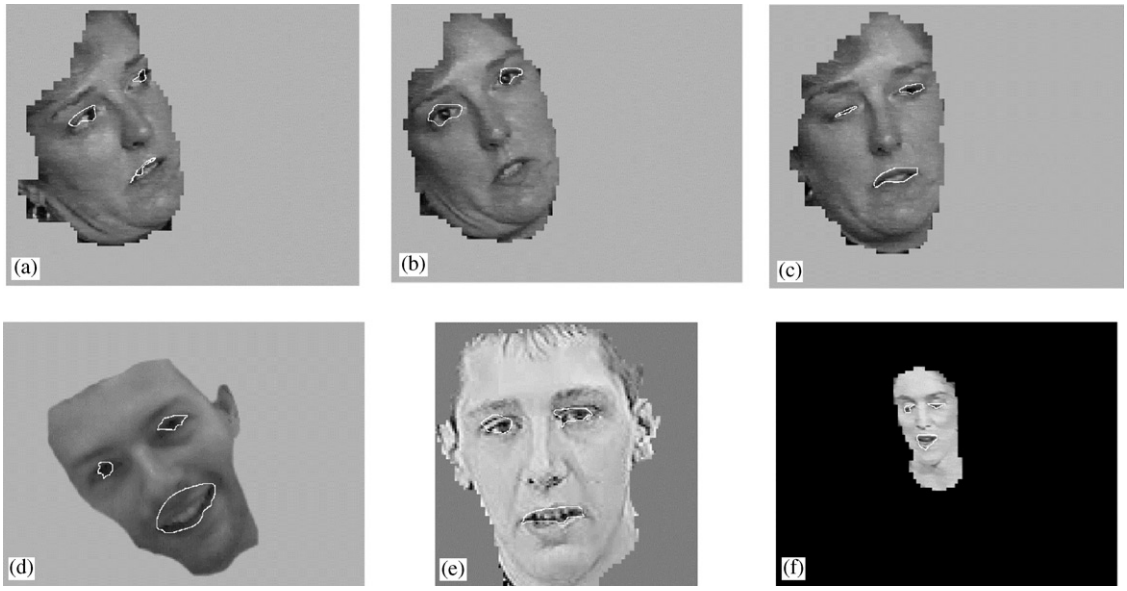

Fig. 16. Feature labeling.

background and hair can provide false feature generation. We tested the robustness of the proposed method to such cases. Fig. 18(a) shows a poorly segmented face, which was used as input to our method. Fig. 18(b) illustrates snake initializations and final snake estimates, showing a significant increase in the number of estimated seeds. Almost every part of the visible background

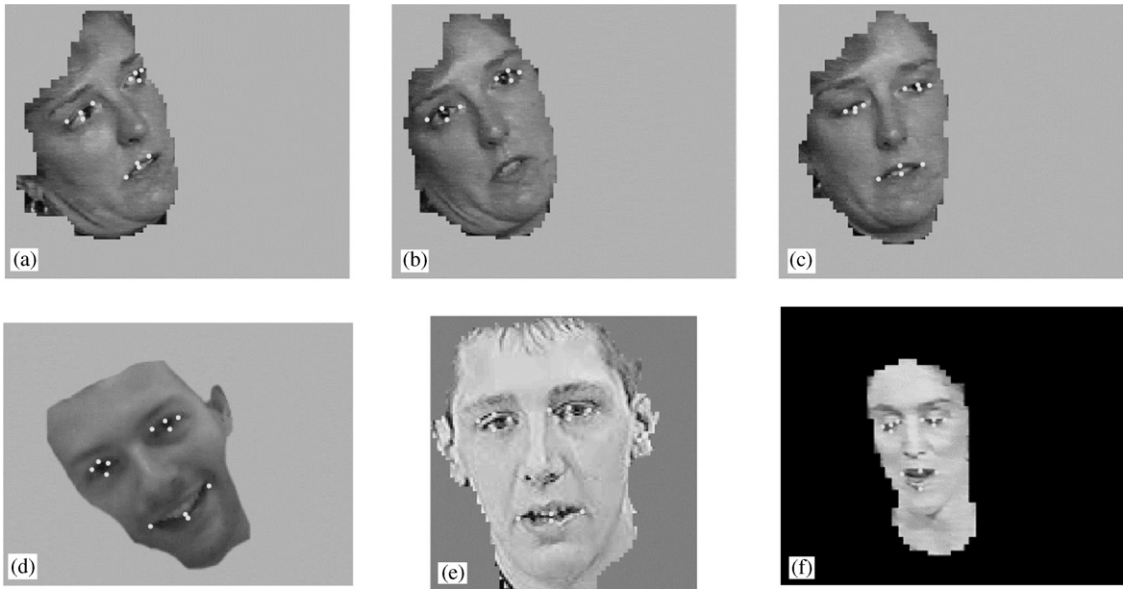

Fig. 17. The obtained feature point sets, as subsets of the MPEG-4 FDP set.

Table 1
Results in terms of success and failure percentages, concerning the three levels: feature sets, features and in-plane rotation angles

Feature sets		Features		Rotation angles	
Mislabeled (%)	4			Incorrect (%)	4
Not labeled (%)	7			Correct (%)	96
Labeled (%)	89	Mislabeled (%)	1		
		Not labeled (%)	9		
		Labeled (%)	90		

and of the person's hair is considered as a candidate facial feature. The primary features that were labeled from our approach are shown in Fig. 18(c): only the right eye and the mouth were successfully labeled.

We investigated two methods to overcome such poor initial face detection. The first one has been by iterating through stages (ii) and (iii) of the procedure shown in Fig. 1, gradually increasing the contribution of the two successfully detected features in stage (ii) computations. The second approach first computed an active contour of the poorly segmented input face [50], thus improving face detection, and then applied our 3-stage procedure to the modeled facial area. Fig. 18(d) shows the active contour of the face drawn in white line. Fig. 18(e) shows the facial area given as input to the 3-stage procedure. Fig. 18(f) shows the correct labeling that has been achieved by either of the two methods.

Next, for comparison purposes, we applied a low-level method for facial features extraction [15] on one of the images we used in our experiments. The method consists of the following steps. It first uses a vertical edge edge-detection. After thresholding, the gradient image is dilated and eroded (using an 8-connectivity kernel), so that blobs are created, that are candidate feature areas. The resulting blob image is then vertically split in two windows, and twin blobs are searched for in them, to locate the left and right eye. Spatial horizontal location and the area are the criteria used to select the twin blobs. Similarly, continuing the application of vertical edge detectors below the estimated eye locations, followed by thresholding and morphological operations, blobs that correspond to the nose and the mouth position can be obtained [15]. Fig. 19 shows the result obtained by this method. By comparing Figs. 19 and 16, it can be seen that the proposed method provided

Fig. 18. Test with a poorly detected initial image (a); estimated seeds (b); labeled features (c); active contour of the initial facial image (d); corresponding initial input image (e); and the labeled features (f).

much better estimation of the facial features, as well as their pose.

Finally, we provide some experimental results in the direction of MPEG-4 facial animation through the use of FDPs and FAPs. Based on [49,39], we have been using FDP estimation between neutral and slowly, but continuously varying facial states, to estimate FAPs. For our experiments on creating the animation of the face we used the face model developed in the context of the European Project ACTS MoMuSys, available at the website http://www.iso.ch/ittf. Using the FDPs provided by our method when applied to consecutive frames a video sequence (Figs. 2 and 17 (a–c)) and considering one of the frames as the initial neutral state, we generated the corresponding facial animations shown in Figs. 20(a, b).

Implementation of the 3-stage procedure of Fig. 1 took about 10–15 s per image. Each images' dimensions were approximately 150×200 pixels. The procedure could be significantly accelerated in the case that:

- code optimisation was supported,
- parallelisation was operated, where applicable,

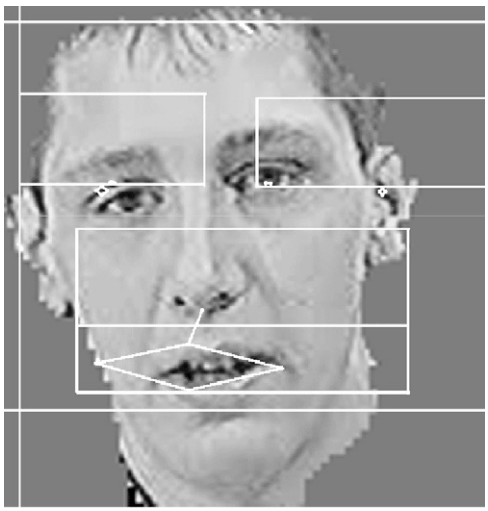

Fig. 19. Comparison with a low-level facial method.

- the code was converted to C and graphic representations were omitted and
- special hardware was used.

All these reasons indicate that even real time processing could be achieved, taking under consideration a standard video sequence frame rate

Fig. 20. MPEG-4 facial animation based on estimated FAPs: neutral position (a); and animation corresponding to image 17(c) (b).

(i.e. PAL, NTSC). However, this is still a postulation and needs to be further explored.

5. Conclusions

In the current work, a gradual confidence approach concerning facial feature extraction over real video frames is presented. The proposed methodology copes with large variations in the appearance of diverse subjects, as well as of the same subject in various instances within real video sequences. In this sense, the problem of feature extraction is being dealt with under general imaging conditions. The system extracts the areas of the face that statistically seem to be outstanding and forms an initial set of regions that are likely to include information about the features of interest. It then enhances their content, producing closed objects, which generally include the desired features. The system determines the dominant angle over all objects of the object set (which is thought of being the facial rotation angle), using a relevant fuzzy system. The object set is restricted using the dominant angle. An exhaustive search is performed seeking for an anthropomorphic pattern that suits that of the eyes and the mouth among all candidate objects. These features are finally labeled. As a consequence, the associated feature points, which constitute a subset of the MPEG-4 facial definition parameter set, as well as corresponding facial animation parameters can be extracted. The gradual revelation of information concerning the face is supported under the scope of optimization for each step, producing a posteriori knowledge about it and leading to a step-by-step visualization of the features in search. By expressing all stages of the proposed method as energy minimization problems, a block component method (BCM), which iteratively implements all three stages, can be created. Convergence of such a scheme is an interesting research topic that is currently under investigation.

In our approach, primary facial features, such as the eyes and the mouth, are being consistently located. Future work involves extensions towards reliable, hierarchical extraction of secondary and transient facial features, as well as their robust, real time tracking over video sequences. We are currently investigating the use of the proposed approach for generating emotionally rich HCI, where a workstation analyzes its user's speech and facial gestures to recognize his or her emotional state and behave analogously [14].

References

[1] J. Ahlberg, H. Li, Representing and compressing facial animation parameters using facial action basis functions, IEEE Trans. Circuits Systems Video Technol. 93 (3) (April 1999) 405–411.

[2] Y. Avrithis, N. Tsapatsoulis, S. Kollias, Broadcast news parsing using visual cues: a robust face detection approach, in: IEEE International Conference on Multimedia and Expo, New York City, NY, USA, July 2000.

[3] M. Black, Y. Yacoob, A. Jepson, D. Fleet, Learning parameterized models of image motion, in: Proceedings of the IEEE CVPR, 1997, pp. 561–567.
[4] C. Bouman, M. Shapiro, A multiscale random field model for Bayesian image segmentation, IEEE Trans. Image Process. 3 (2) (March 1994) 162–177.
[5] R. Brunelli, T. Poggio, Face Recognition: Features versus Templates, IEEE Trans. Pattern Anal. Machine Intell. 15 (10) (October 1993) 1042–1052.
[6] R. Chellappa, C.L. Wilson, S. Sirohey, Human and machine recognition of faces: a survey, Proc. IEEE 83 (5) (May 1995) 705–740.
[7] L.D. Cohen, On Active Contour Models and Baloons, Proc. CVGIP: Image Understanding 53 (2) (March 1991) 211–218.
[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human computer interaction, IEEE Signal Process. Mag. 18 (1) (January 2001) 32–80.
[9] T.J. Darrell, A.P. Pentland, Recognition of space-time gestures using distributed representation, Technical Report No. 197, M.I.T. Media Laboratory, Vision and Modeling Group, 1993.
[10] A. DelBimbo, P. Pala, Visual image retrieval by elastic matching of user scetches, IEEE Trans. Pattern Anal. Machine Intell. 19 (2) (February 1997) 121–132.
[11] Y. Delignon, A. Marzouki, W. Pieczynski, Estimation of generalized mixtures and its application in image segmentation, IEEE Trans. Image Process. 6 (10) (1997) 1364–1376.
[12] Y. Deng, B.S. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Trans. Pattern Anal. Machine Intell. 23 (8) (August 2001) 800–810.
[13] A.D. Doulamis, N.D. Doulamis, S.D. Kollias, On-line retrainable neural networks: improving the performance of neural networks in image analysis problems, IEEE Trans. Neural Networks 11 (1) (January 2000) 137–155.
[14] EC IST Project "ERMIS", www.image.ntua.gr/ermis.
[15] EC Training and Mobility Research (TMR) Project "PHYSTA", www.image.ntua.gr/physta.
[16] P. Eisert, B. Girod, Model-based estimation of facial expression parameters from image sequences, in: Proceedings of the IEEE International Conference on Image Processing, Washington, DC, 1997.
[17] I.A. Essa, A.P. Pentland, Coding, analysis, interpretation and recognition of facial expressions, IEEE Trans. Pattern Anal. Machine Intell. 19 (7) (July 1997) 757–763.
[18] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Machine Intell. PAMI-6 (1984) 721–741.
[19] L.D. Harmon, M.K. Khan, R. Lasch, P.F. Ramig, Machine identification of human faces, Pattern Recognition 13 (2) (1981) 97–110.
[20] A.K. Jain, Y. Zhong, S. Lakshmanan, Object matching using deformable templates, IEEE Trans. Pattern Anal. Machine Intell. 18 (3) (March 1996) 267–278.
[21] X. Jia, M.S. Nixon, Extending the feature vector for automatic face recognition, IEEE Trans. Pattern Anal. Machine Intell. 17 (12) (December 1995) 1167–1176.
[22] T. Kanade, Picture processing by computer complex and recognition of human faces, Technical Report, Department of Information Science, Kyoto University, 1973.
[23] K. Karpouzis, G. Votsis, N. Tsapatsoulis, S. Kollias, Compact 3D model generation based on 2D views of human faces: application to face recognition, Machine Graphics Vision 7 (1–2) (1998) 75–85.
[24] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, Int. J. Comput. Vision 1 (4) (1988) 321–331.
[25] D.A. Langan, J.W. Modestino, J. Zhang, Cluster validation for unsupervised stochastic model-based image segmentation, IEEE Trans. Image Process. 7 (2) (1997) 180–244.
[26] B.K. Low, M.K. Ibrahim, A fast and accurate algorithm for facial feature segmentation, in: Proceedings of the IEEE International Conference on Image Processing, 1997.
[27] D.J. Luenberger, Liner and Nonlinear Programming, Addison-Wesley, Reading, MA, 1984.
[28] W.Y. Ma, B.S. Manjunath, Edge flow: a framework of boundary detection and image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 744–749.
[29] Y. Matsumoto, A. Zelinsky, Real-time stereo face tracking system for visual human interfaces, in: Proceedings of the IEEE Workshop on Real-Time Analysis and Tracking of Face and Gesture in Real-Time Systems, Kerkyra, Greece, September 1999.
[30] T. McInerney, D. Terzopoulos, Topologically adaptable snakes, in: Proceedings of the International Conference on Computer Vision, Cambridge, MA, 1995, pp. 840–845.
[31] S.J. McKenna, S. Gong, R.P. Würtz, J. Tanner, D. Banin, Tracking facial feature points with gabor wavelets and shape models, in: G. Borgefors, G. Chollet, J. Biguen (Eds.), International Conference on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, Springer, Berlin, 1997.
[32] D. Metaxas, Deformable model and HMM-based tracking, analysis and recognition of gestures and faces, in: Proceedings of IEEE Workshop on Real-Time Analysis and Tracking of Face and Gesture in Real-Time Systems, Kerkyra, Greece, September 1999.
[33] A. Nikolaidis, I. Pitas, Facial feature extracion and pose determination, in: Proceedings of the NOBLESSE Workshop on Nonlinear Model Based Image Analysis, Glasgow, Scotland, 1998.
[34] D.K. Panjwani, G. Healey, Markov random field models for unsupervised segmentation of textured color images, IEEE Trans. Pattern Anal. Machine Intell. 17 (10) (October 1995) 939–954.
[35] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd Edition, McGraw-Hill, Singapore, 1991, Chapter 5, pp. 86–123.

[36] A. Pentland, B. Moghaddam, T. Starner, View-based and modular eigenspaces for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, 1994.
[37] M. Pötzsch, N. Krüger, C. von der Malsburg, Improving object recognition by transforming gabor filter responses, Network: Comput. Neural Systems 7 (2) (May 1996) 341–347.
[38] P. Radeva, E. Marti, Facial features segmentation by model-based snakes, in: International Conference on Computer Analysis and Image Processing, Prague, 1995.
[39] A. Raouzaiou, N. Tsapatsoulis, K. Karpouzis, S. Kollias, Parameterized facial expression synthesis based on MPEG-4, Eurasip J. Appl. Signal Proccess., Vol. 2002, No. 10, October 2002.
[40] A. Samal, P.A. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, Pattern Recognition 25 (1) (1992) 65–76.
[41] S. Sclaroff, L. Liu, Deformable shape detection and description via model-based region grouping, IEEE Trans. Pattern Anal. Machine Intell. 23 (5) (May 2001) 475–489.
[42] A.W. Senior, Recognizing faces in broadcast video, in: Proceedings of IEEE Workshop on Real-Time Analysis and Tracking of Face and Gesture in Real-Time Systems, Kerkyra, Greece, September 1999.
[43] L. Shafarenko, M. Petrou, J. Kittler, Automatic watershed-segmentation of randomly textured color images, IEEE Trans. Image Process. 6 (11) (1997) 1530–1544.
[44] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Machine Intell. 22 (8) (2000) 888–905.
[45] K. Sobottka, I. Pitas, A novel method for automatic face segmentation, facial feature extraction and tracking, Signal Processing: Image Communication 12 (1998) 263–281.
[46] L.H. Staib, J.S. Duncan, Boundary finding with parametrically deformable models, IEEE Trans. Pattern Anal. Machine Intell. 14 (11) (November 1992) 1061–1075.
[47] Y. Tian, T. Kanade, J.F. Cohn, Multi-state based facial feature tracking and detection, Technical Report CMU-RI-TR-99-18, Robotics Institute, Carnegie Mellon University, August 1999.
[48] N. Tsapatsoulis, Y. Avrithis, S. Kollias, Facial image indexing in multimedia databases, Pattern Anal. Appl. 4 (2/3) (2001) 93–107.
[49] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, E. Douglas-Cowie, Emotion recognition and synthesis based on MPEG-4 FAPs, in: I. Pandzic, R. Forchheimer (Eds.), MPEG-4 Facial Animation, Wiley, UK, 2002.
[50] G. Tsechpenakis, Y. Xirouhakis, A. Delopoulos, A multiresolution approach for main mobile object localization in video sequences, in: International Workshop on Very Low Bitrate Video Coding (VLBV01), Athens, October 2001.
[51] S. Tsekeridou, I. Pitas, Facial feature extraction in frontal views using biometric analogies, in: Proceedings of the IX European Signal Processing Conference, Rhodes, Greece, Vol. 1, 1998, pp. 315–318.
[52] J.P. Wang, Stochastic relaxation on partitions with connected components and its application to image segmentation, IEEE Trans. Pattern Anal. Machine Intell. 20 (6) (1998) 619–636.
[53] K.C. Yow, R. Cipolla, A probabilistic framework for perceptual grouping of features for human face detection, in: Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Vermont, USA, 1996.
[54] A.L. Yuille, D.S. Cohen, P.W. Hallinan, Feature extraction from faces using deformable templates, Int. J. Comput. Vision 8 (2) (August 1992) 99–111.
[55] S.C. Zhu, A. Yuille, Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation, IEEE Trans. Pattern Anal. Machine Intell. 18 (9) (September 1996) 884–900.