Pattern Recognition 43 (2010) 1763–1775

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Real-time 2D+3D facial action and expression recognition

Filareti Tsalakanidou *, Sotiris Malassiotis

Informatics and Telematics Institute, Centre for Research and Technology Hellas, 1st km Thermi-Panorama Road, Thermi 57001, P.O. Box 60361, Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 21 July 2009
Received in revised form
11 November 2009
Accepted 10 December 2009

Keywords:
3D face tracking
Facial action unit detection
Facial expression classification

ABSTRACT

This paper presents a completely automated facial action and facial expression recognition system using 2D+3D images recorded in real-time by a structured light sensor. It is based on local feature tracking and rule-based classification of geometric, appearance and surface curvature measurements. Several experiments conducted under relatively non-controlled conditions demonstrate the accuracy and robustness of the approach.

1. Introduction

Next generation computing systems are expected to interact with users in a way that emulates face to face encounters. Face to face communication relies significantly on the implicit and non-verbal signals expressed through body and head posture, hand gestures and facial expressions for determining the spoken message in an non-ambiguous way [1]. Facial expressions in particular are considered to be one of the most powerful and immediate means for humans to communicate their emotions, intentions and opinions to each other and this is why much effort has been devoted to their study by cognitive scientists and lately computer vision researchers [2,3].

Several approaches have been reported towards automatic facial expression recognition from 2D static images or video sequences [3]. In all of these works, after the face has been detected, facial features that are relevant to the display of expressions are extracted and classified into a predefined set of facial actions or furthermore to emotion related expressions. The majority of facial expression recognition research is limited to the six basic emotions, i.e. happiness, sadness, anger, fear, surprise and disgust, the display of which is widely assumed to be universal. However, there is a growing number of approaches, which instead try to detect a set of facial muscle movements known as Facial Action Units, which are more subtle but their combinations may describe effectively any facial expression [2].

Facial features used for expression recognition may be roughly classified into geometric, e.g. distances between facial points [4], appearance-based such as Gabor filter responses [5] or holistic such as optical flow fields [6]. Classification methods can be roughly divided into static and dynamic ones. Static classifiers use feature vectors related to a single frame to perform classification. In the case of image sequences, this frame corresponds to the peak of the depicted expression. Probabilistic as well as rule-based techniques are popular [4,7]. Temporal classifiers on the other hand try to capture the temporal pattern in the sequence of feature vectors over subsequent frames [8,9].

A problem with existing techniques is that the subtle skin deformations that characterize facial expressions are difficult to capture by a 2D camera. Moreover, 2D techniques are prone to illumination changes and pose variations that affect the perceived geometry and appearance of facial features. To handle problems caused by pose variations, some researchers proposed the use of multiple views of the face [7], deformable 3D models fitted on 2D images [10] or 3D images.

Although the advantages of using 3D facial images are self-evident, very few works have examined 3D facial expression recognition. Works that use a single 3D image are [11–14]. Wang et al. [11] employ a surface labelling approach based on the distribution of principal curvature descriptors defined over different face regions. Expression recognition is subsequently based on the distribution of the above labels over the face. Tang and Huang [12] and similarly Soyel and Demirel [13] rely on 3D Euclidean distances between manually annotated feature points. In [14], feature localization is addressed using an elastically deformable model, which establishes point correspondence between facial surfaces, while face and facial expression recognition is based on bilinear models that effectively decouple identity and facial expression.

Works that use a sequence of range images are [15–18]. Huang et al. [15] fit a deformable face model on sequences of 3D face

* Corresponding author. Tel.: +30 2310 464160; fax: +30 2310 464164.
E-mail addresses: filareti@iti.gr (F. Tsalakanidou), malasiot@iti.gr (S. Malassiotis).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.12.009

scans depicting facial expressions, but do not perform recognition of facial expressions. Chang et al. [16] learn a generalized expression manifold from range image sequences, which is subsequently exploited for recognizing temporal expression patterns. They use 2D feature tracking and a coarse 3D mesh model. In [17], a spatio-temporal approach is adopted based on 3D surface descriptors [11] and 2D hidden Markov models. Good recognition rates are reported, however the proposed method relies on semi-automatic face tracking and computationally expensive curvature estimation. Most similar to our approach is the work of Liebelt et al. [18] that also uses 2D+3D image sequences. However, our approach relies on local features for tracking and classification while [18] relies largely on texture information and classification is based on AAM model parameters that do not decouple expression from identity.

In this paper, we address the problem of facial expression recognition by a combination of 2D and 3D image streams, which allows us to achieve real-time, accurate, pose and illumination invariant recognition of facial actions and facial expressions. We employ a model-based feature tracker applied to sequences of 3D range images and corresponding grayscale images recorded by a novel real-time 3D sensor [19]. To achieve real-time performance we do not rely on 3D shape registration algorithms, which require a dense face mesh, but instead on feature based 3D pose estimation followed by iterative tracking of 81 facial points using local appearance and surface geometry information. Special trackers are developed for important facial features such as the mouth and the eyebrows that account for the non-linearity of the appearance of these features. A set of measurements (geometric, appearance and curvature-based) is subsequently extracted, which effectively model changes in the shape of facial features and their geometrical arrangement as well as deformations of the face surface caused by wrinkles or furrows. We use these measurements to recognize four facial expressions and 11 facial action units using a rule-based approach. Temporal information is also exploited for detecting action units and facial expression activation periods. Finally, the efficiency of the 3D face analyzer is evaluated in a database with more than 50 subjects and 800 sequences. A preliminary version of this work has appeared in [20].

To the best of our knowledge this is the first fully automatic real-time 3D facial expression recognition system. Additional contributions of this paper are:

- We present a novel 2D+3D facial feature tracker that relies on local information only and is thus computationally efficient and robust to illumination variations unlike [15,18]. We also propose techniques for achieving robustness with noisy or incomplete data.
- The system is completely independent of the user's facial appearance and geometry and no per-user calibration is required, in contrast with techniques such as [18] where the generic model used includes identity information that eventually affects classification performance. The proposed approach can also be easily extended to include a large range of facial action units. On the contrary, techniques such as [16,17] rely on pre-recorded sequences of each facial action unit to be available so that the temporal expression classifier or manifold can be trained.
- Apart from 3D Euclidean distances used in [12,13], we introduce a set of meaningful 3D surface measurements such as wrinkling, stretching, etc., which enable us to detect subtle facial deformations around the nose and the mouth. Unlike existing works that have been tested with moderate expressions, we also handle extreme deformations of the mouth.

The paper is organized as follows. The face tracker and the local facial feature detectors are described in Section 2. A set of geometric and surface deformation measurements is presented in Section 3, while a rule-based approach for facial expression and facial action unit classification in both static images and image sequences is outlined in Section 4. The performance of the proposed system is evaluated in Section 5. Finally, conclusions are drawn in Section 6.

2. Face and facial feature tracking

The first and most important step towards automatic recognition of facial expressions is accurate detection of the position of the face and prominent facial features such as the eyes, eyebrows, mouth, etc. In this section, we present a novel 3D face tracker based on the well-known Active Shape Model (ASM) technique [21], which was extended to handle 3D data and also cope with measurement uncertainty and missing data.

The ASM is a point distribution model (PDM) accompanied by a local appearance pattern for every point, which effectively models the shape of a class of objects, faces in our case. Point and local appearance distributions are obtained using a set of annotated training images. Any shape can then be expressed as the sum of a mean shape and a linear combination of basis shapes computed during training. Although ASMs have been demonstrated less accurate than Active Appearance Models (AAM), they have the advantage of robustness to illumination variations (using local gradient search) and are very efficient.

Our approach employs 2D and 3D facial information in the form of pairs of depth and associated grayscale images recorded by a novel 3D sensor based on NIR structured light [19] (see Section 5). Pixel values of depth images represent the distance of the corresponding point from the camera plane. Using the one-to-one pixel correspondence of depth and grayscale images as well as camera projection parameters, we can directly associate every image point with its 3D coordinates and a texture value.

The shape $\mathbf{s}$ of the face is represented as a sequence of $n=81$ points corresponding to salient facial features as can be seen in Fig. 1. The PDM is then expressed as

$$\mathbf{s}=\tilde{\mathbf{s}}+\sum_{i=1}^{m} a_i\mathbf{s}_i=\tilde{\mathbf{s}}+\mathbf{a}\cdot\mathbf{S} \tag{1}$$

where $\mathbf{s}=\{x_1,y_1,z_1,\ldots,x_n,y_n,z_n\}$ is the vector of n landmark coordinates, $\mathbf{s}_i$ are the basis shapes computed by applying principal component analysis to a set of manually annotated training examples, which are aligned to a common coordinate frame (called model coordinate frame), $\tilde{\mathbf{s}}$ is the mean shape computed in the same space and $\mathbf{a}$ is a vector of shape parameters.

Note that image alignment involves 3D rotation and translation of original image pairs so that all faces have a frontal orientation and be at the same distance from the camera plane as well as linear interpolation of missing depth values as proposed in [22].

The local appearance model for each landmark L_i is computed from image gradient information gathered in all 2D training images along a line passing through $\mathbf{p}_i$, the projection of L_i in the 2D image plane. This line is chosen to be perpendicular to the boundary of the shape of the feature that L_i belongs to (e.g. eyebrow, mouth, etc.). A set of shape boundaries is defined in terms of connectivity information between landmarks as illustrated in Fig. 1. Let us assume that L_i is connected to L_k and L_m. Then the normal at $\mathbf{p}_i$ is equal to $\mathbf{n}_i=(\mathbf{u}_{ki}+\mathbf{u}_{im})/2$, where $\mathbf{u}_{ki}$ and $\mathbf{u}_{im}$ are unit vectors perpendicular to segments defined by $\mathbf{p}_i$, $\mathbf{p}_k$ and $\mathbf{p}_i$, $\mathbf{p}_m$, respectively. Note that since all boundary curves have

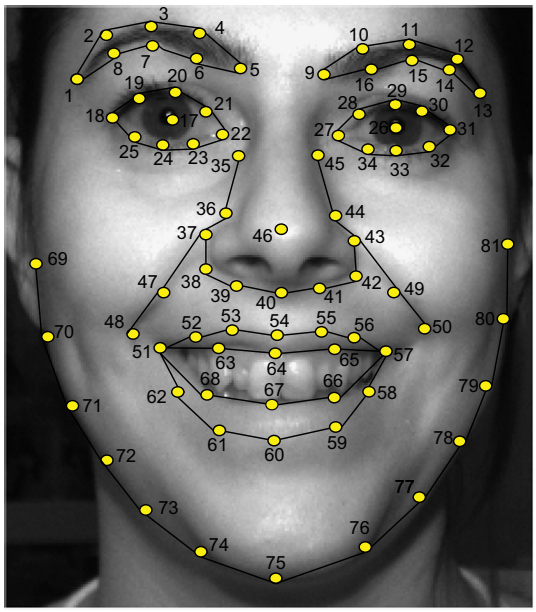

Fig. 1. The 81 landmarks and corresponding segments of the ASM.

been defined clockwise, the direction of $\mathbf{n}_i$ (and $\mathbf{u}_{ki}$, $\mathbf{u}_{im}$) is always from the inside to the outside of the specific feature.

Based on the estimated normal direction, we then define a set of $2 \cdot m_q + 1$ pixels $\mathbf{q}_j$ along $\mathbf{n}_i$, where $\mathbf{q}_j = j \cdot \mathbf{n}_i + \mathbf{p}_i, j = -m_q, \ldots, m_q$. Obviously $\mathbf{q}_0 = \mathbf{p}_i$. For each pixel $\mathbf{q}_j$, we compute a gradient measurement

$$g_j = \sum_{k=1}^{m_g} z_k \cdot (c_{j+k} - c_{j-k}) \tag{2}$$

where c_j is the intensity of $\mathbf{q}_j$, m_g the Gaussian kernel width and z_k the kernel weights. We set $m_g = 3$. The estimated gradient values represent the local gradient profile $\mathbf{g} = [g_j]$ of $\mathbf{p}_i$.

After computing the gradient profiles of L_i in all training images, we can build a local model of gradient changes associated with this landmark assuming a unimodal Gaussian distribution. The same procedure is applied for every landmark thus obtaining n local appearance models.

Using Eq. (1), we can represent the shape of any face in the model coordinate frame. To express the same shape in the real-world coordinate frame we use

$$\mathbf{x} = \mathbf{R} \cdot \mathbf{s} + \mathbf{T} = \mathbf{R} \cdot (\tilde{\mathbf{s}} + \mathbf{a} \cdot \mathbf{S}) + \mathbf{T} \tag{3}$$

where $\mathbf{R}$ is the 3D rotation matrix and $\mathbf{T}$ the 3D translation vector that rigidly align the model coordinate frame with the real-world coordinate frame and $\mathbf{x}$ represents the landmark coordinates in the real-world coordinate frame. By projecting $\mathbf{x}$ in the image plane, we obtain the corresponding 2D shape $\mathbf{v} = P(\mathbf{x})$, where P represents a camera projection function that models the imaging process. $\mathbf{v}$ represents the landmark positions in the 2D image.

To estimate the landmark positions in a new pair of 2D and 3D images the following steps are taken:

1. Let $\mathbf{R}$ be the 3D rotation matrix and $\mathbf{T}$ the 3D translation vector that rigidly align the model with the face. A first estimate of these is obtained using the 3D face detection and pose estimation technique proposed in [23]. First, the face is roughly detected in the input 3D image using global moment descriptors and a priori knowledge of the geometry and relevant dimensions of the head and other body parts. Then, the face position is localized using a knowledge-based 3D technique that allows us to detect the ridge of the nose with high accuracy. Finally, the pose of the face is reliably estimated based on the detected nose ridge and the inherent bilateral symmetry of the face. A detailed description of the aforementioned algorithm can be found in [23]. Based on the estimated face pose and position, we obtain an initial rigid transformation $(\mathbf{R}, \mathbf{T})$. The shape parameters $\mathbf{a}$ are initialized to zero, i.e. we start from the mean face shape $\tilde{\mathbf{s}}$.
2. The current shape $\mathbf{s}$ is transformed to the real-world coordinate frame using the rigid transformation $(\mathbf{R}, \mathbf{T})$ and is subsequently projected on the 2D camera plane through P. A local search is then performed around each projected landmark position to find the point that best matches the local appearance model. To do this, first we compute the normal vector at the specific location as described above. Then, we define a set of candidate pixels along this line and compute a local gradient vector for each of them exactly as in the case of training images. Similarity between extracted gradient profiles and the corresponding local appearance model is measured using the Mahalanobis distance. The point associated with the lowest distance is selected. The same procedure is applied for all landmarks and a set of new landmark positions is estimated in the 2D image. These are subsequently back-projected in the 3D space using the inverse projection function P^{-1} and the z values of the corresponding pixels of the depth image. Thus a new 3D shape $\mathbf{x}$ is defined in the real-world coordinate frame. Moreover, each landmark is associated with a weight set to be the reciprocal of the computed Mahalanobis distance. In case the corresponding z value of a point is undefined, the median depth value in the neighborhood of this pixel is used. If no depth is defined in the greater area of this pixel, then a zero weight is assigned to this landmark, so that it is neglected in model estimation.
3. A new rigid transformation $(\mathbf{R}, \mathbf{T})$ aligning the new shape $\mathbf{x}$ with the current template $\mathbf{s}$ is estimated using Horn's quaternion method [24]. A new rectified shape $\mathbf{y} = \mathbf{R}^{-1} \cdot (\mathbf{x} - \mathbf{T})$ is computed in the model coordinate frame.
4. A new set of parameters $\mathbf{a}$ is estimated by minimizing $\|\tilde{\mathbf{y}} - \tilde{\mathbf{s}} - \mathbf{a} \cdot \mathbf{S}\|^2 + \lambda \cdot \|\mathbf{a}\|^2$, where the second term is a regularization constraint. A weighted least squares approach is adopted, where each landmark point is weighted proportionally to its strength. We also exclude points that may be occluded, for example points on the side of the face or nose, which may be easily determined using the estimated face orientation. Once a new set of parameters $\mathbf{a}$ is estimated, a new shape $\mathbf{s}$ is synthesized using Eq. (1).
5. Steps 2–5 are repeated until convergence of the fitting error $e = \|\mathbf{y} - \mathbf{s}\|$ or until a number of iterations is reached. Then a new real-world shape $\mathbf{x}$ is computed from $\mathbf{s}$ using Eq. (3).

For each subsequent frame, initialization is performed based on the previous frame, i.e. we start from step 2 using $\mathbf{R}$, $\mathbf{T}$ and $\mathbf{s}$ estimated in the previous frame. If the model has not converged, we re-initialize the tracker, i.e. we start from step 1 and repeat face detection, pose estimation and model fitting. For faster convergence we use a multi-resolution scheme with three layers.

In [18], 2D AAMs are extended to 2D+3D AAMs by introducing a correction step based on fitting a 3D shape model on 3D stereo data. This correction is then used to enhance the result of 2D face tracking. The main advantage of our method compared to [18] is that it is significantly faster and uses a single model, which effectively combines 3D shape information with local appearance data. Moreover, AAMs largely rely on appearance (texture) information thus being sensitive to illumination, occlusions and appearance changes.

The proposed tracker achieves small localization errors per landmark, however there are cases where localization of individual features such as the eyebrows and the mouth is not accurate enough for our purpose as can be seen in Figs. 3 and 4. This is due to the inadequacy of the linearity assumption in the PDM, but also due to the unimodal distribution chosen for local appearance variations (e.g. appearance of teeth when opening the mouth). Instead of resorting to non-linear modelling techniques, we propose a set of dedicated local facial feature detectors, presented in the following.

2.1. Local eyebrows detector

The global face tracker gives generally a good estimate of the eyebrows position but may sometimes fail to accurately localize their boundaries. An improved estimate is obtained using a local 3D ASM with 16 landmarks corresponding to the eyebrow boundaries (points 1–16 in Fig. 1), which is initialized using the eyebrows estimation provided by the global tracker. Such an approach offers increased localization accuracy, however it may also fail in cases that the eyebrow hair is light-colored or sparse or the eyebrow itself is very thin. This may be attributed to the fact that the simple Gaussian model used for local gradient patterns may not be adequate for modelling gradient changes in the eyebrow area. Instead of resorting to more complex statistical models (e.g. bimodal Gaussian distributions), we employ a simpler model fitting technique based on area intensity differences.

For each candidate landmark position, we define two rectangle areas A_1 and A_2 lying on the positive and negative side of the axis defined by the normal in this position (see Fig. 2). A_1 and A_2 are aligned with the normal and their dimensions are chosen to be 5×5 pixels. The average intensities S_1 and S_2 inside A_1 and A_2 are then computed. Eyebrow landmarks lie in the boundary between dark (eyebrow hair) and light-colored (skin) areas thus candidate points should maximize S_1-S_2. In addition to this criterion, we ask that S_1-$S_2 > T_1$ and $S_2 < T_2$. The first condition implies that the landmark point should lie in an area of adequate gradient change. The second is used to overcome the problem of shadows, which results in selecting a candidate point that lies in the border of shadowed and non-shadowed skin areas instead of lying in the border of eyebrow and skin areas. T_1 and T_2 are experimentally chosen with respect to the average brightness inside the face area and the values of S_1 and S_2 obtained from training images.

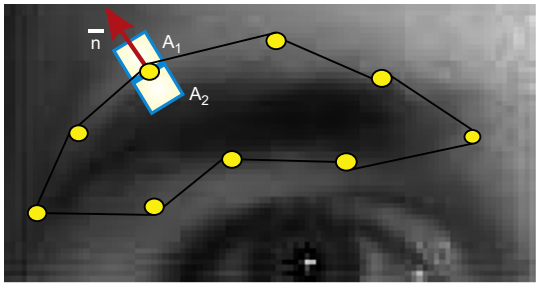

Fig. 2. Eyebrow boundary localization based on area intensity differences (Section 2.1).

The proposed local 3D ASM is initialized using the eyebrows estimation provided by the global 3D ASM and is fitted in the input image using steps 2–5 above. Note that in this case, landmark points are weighted proportionally to the corresponding intensity difference S_1-S_2. Since a bad initial estimation may prevent the local model from converging, we perform several local fittings with slightly perturbed initial positions and choose the one minimizing the fit error.

As can be seen in Fig. 3, the proposed local eyebrow detector enhances significantly the estimation provided by the global ASM especially in cases where the eyebrows are raised or lowered.

2.2. Local mouth detector

Lip boundary localization is also problematic due to the unimodal Gaussian distribution assumption used for the representation of local mouth appearance patterns, which is not suitable for landmarks lying in the inner lip boundaries, since their local gradient patterns are significantly affected by whether the mouth is open or closed. Fig. 4 presents some examples of incorrect estimation of mouth landmarks. It can be seen that the problem is more intense when the mouth is open and the teeth are visible, since in this case the boundary between the teeth and the dark area of the mouth cavity is erroneously recognized as a lip boundary.

To overcome this problem, we propose a two-step approach for localizing lip boundaries. First, a two-class support vector machine classifier with an RBF kernel is used to decide whether the mouth is open or closed. Then an open or closed mouth local 3D ASM is fitted on the face to localize the position of outer and inner lip boundaries.

Mouth classification is based on a 16-dimensional feature vector computed from local 3D geometric and 2D appearance measurements over the area defined by the current fit. Given the initial estimation of lip boundaries, we define four regions in the mouth area: upper lip, lower lip, between lips area and whole mouth area (union of the previous three). The attribute vector includes features such as mean intensity, intensity variance,

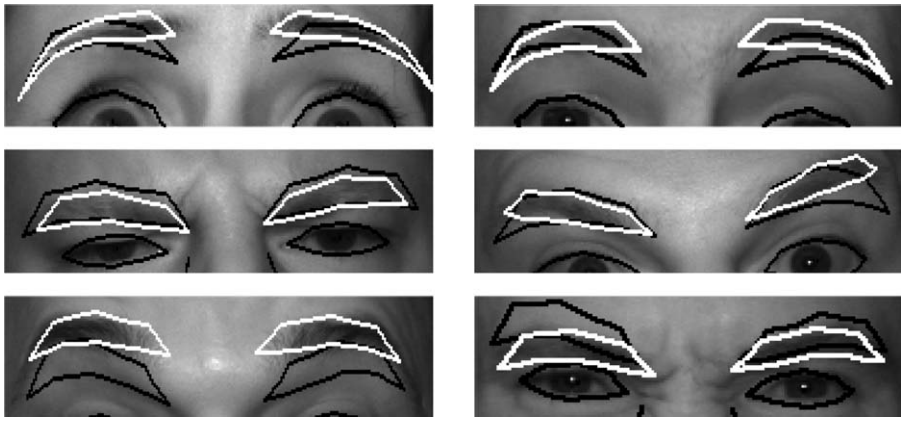

Fig. 3. Examples of eyebrow boundary localization using the global 3D ASM model (black line) and the proposed local detector (white line).

Fig. 4. Examples of lip boundary localization using the global 3D ASM model (black line) and the proposed local detector (white line).

Fig. 5. Examples of facial feature tracking results using the proposed global tracker and local feature detectors.

intensity gradient and variance of intensity gradient of mouth area, mean depth and depth variance of the mouth area, mean intensity and intensity variance of the area between the lips, mouth opening (3D distance between the upper and lower inner lip boundaries), percentage of dark pixels and percentage of white pixels in the mouth area (corresponding to mouth cavity and teeth, respectively), etc.

The classifier was trained with approximately 240 pairs of images (120 faces with open mouth and 120 faces with closed mouth), which were manually annotated to determine the position of lip boundary points. A correct classification rate of 98% was achieved in a test set of 200 images with various facial expressions, where the initial position of mouth landmarks was determined automatically by fitting the global ASM.

After the mouth is classified as open or closed, the corresponding mouth model (18 landmarks corresponding to points 51–68 of the global model) is fitted on the face. Model fitting is based on image gradient profiles. However, we do not only consider points along the normal but also points in a narrow zone aligned with the normal. Examples of improved mouth localization are depicted in Fig. 4.

2.3. Combining global and local feature position estimates

To incorporate the information provided by the local feature detectors into the global model, the fitting algorithm presented in Section 2 is modified as follows: after step 2, the parts of shape **x** corresponding to eyebrows and mouth are replaced with the improved estimates. Then we continue with step 3. Using the proposed 2D+3D ASM and dedicated local detectors very good localization accuracy may be achieved even under moderate face poses as can be seen in Fig. 5.

3. Extraction of facial feature measurements

To encode facial movements, we adopt the Facial Action Coding System (FACS) developed by Ekman and Friesen [2], where facial appearance changes are described in terms of 44 facial action units (AUs), each of which is related to the contraction of one or more facial muscles. Detection of a subset of these AUs is achieved by combining 23 geometric, appearance and surface deformation measurements denoted as M_1-M_{23}.

3.1. Geometric measurements

Geometric measurements are computed using the estimated positions of the 81 landmarks L_i. The 20 measurements used are presented in Table 1. Note that all such measurements are in 3D and are thus invariant to pose and distance from the camera plane.

Table 1
Geometric facial measurements.

	Measurement name	Measurement
M_1	Inner eyebrow displacement	$d_{5,22}$, $d_{9,27}$
M_2	Outer eyebrow displacement	$d_{7,17}$, $d_{15,26}$
M_3	Inner eyebrow corners dist.	$d_{5,9}$
M_4	Eyebrow from nose root dist.	$d_{5,35}$, $d_{9,45}$
M_5	Eye opening	$d_{20,24}$, $d_{29,33}$
M_6	Eye shape	$d_{20,24}/d_{18,22}$
M_7	Nose length	$(d_{35,36}+d_{45,44})/2$
M_8	Nose width	$d_{36,44}$
M_9	Cheek lines angle	$a(\varepsilon_{37,48}, \varepsilon_{43,50})$
M_{10}	Upper lip boundary shape	$a(\varepsilon_{51,57}, \mathbf{t}_{63})$
M_{11}	Lower lip boundary length	$l_{51,68,67,66,57}$
M_{12}	Lower lip boundary shape	$a(\varepsilon_{51,57}, \mathbf{t}_{68})$
M_{13}	Mouth corners dist.	$d_{51,57}$
M_{14}	Mouth opening	$d_{64,67}$
M_{15}	Mouth shape	$d_{64,67}/d_{51,57}$
M_{16}	Nose–mouth corners angle	$a(\varepsilon_{38,51}, \varepsilon_{42,57})$
M_{17}	Mouth corners to eye dist.	$d_{17,51}$, $d_{26,57}$
M_{18}	Mouth corners to nose dist.	$d_{51,40}$, $d_{57,40}$
M_{19}	Upper lip to nose dist.	$d_{54,40}$
M_{20}	Lower lip to nose dist.	$d_{67,40}$

$d_{i,j}$ is the 3D Euclidean distance between landmarks L_i and L_j. $\varepsilon_{i,j}$ is the 2D line defined by the projections of L_i and L_j in the 2D image plane. l_{ijk} is the length of the 3D curve defined by L_i, L_j, L_k, $\mathbf{t}_i$ is the tangent vector computed in L_i. $a(\varepsilon_a, \varepsilon_b)$ is the angle between 2D lines ε_a, ε_b. Measurements M_{10} and M_{12} represent the shape of the inner boundaries of the upper and lower lips and are a measure of the concavity or convexity of these curves. Depending on whether the lower (upper) lip has a $\smile$, $\frown$ or – shape, M_{12} (M_{10}) has a positive, negative or zero value, respectively.

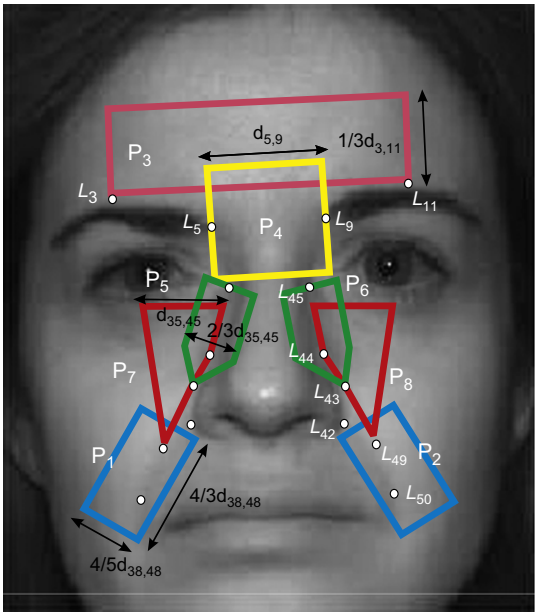

Fig. 6. Polygon areas P_k defined in the face surface for detecting the presence of wrinkles. White dots correspond to facial landmarks L_i. $d_{i,j}$ is the distance between landmarks L_i and L_j.

3.2. Surface deformation measurements

Surface deformation measurements are associated with wrinkles appearing on the skin due to muscle contractions. These include cheek wrinkling (M_{21}), forehead wrinkling (M_{22}) and nose wrinkling (M_{23}). The approximate position of these wrinkles may be easily determined using the estimated landmark positions. Wrinkling measurements are subsequently obtained using both image gradient and surface curvature descriptors.

Intensity gradient is obtained by applying a derivative of Gaussian filter on illumination rectified image patches. Curvature descriptors are based on surface gradients, which are robustly obtained by locally fitting a quadratic surface patch (9×9 pixels) after median filtering the depth image. In the following, we describe in detail the computation of M_{21}, M_{22} and M_{23}.

3.2.1. Cheek wrinkling (M_{21})

To detect the presence of wrinkles in the cheeks area, we define two rectangular boxes P_1 and P_2 enclosing the left and right cheek line (see Fig. 6). For each box, we compute the average intensity gradient perpendicular to segments defined by the 2D projections of landmarks L_{38}, L_{48} and L_{42}, L_{50} respectively. We also compute the Gaussian and mean curvature using the corresponding 3D image.

When cheek wrinkles appear in the face, then the mean of the absolute values of the Gaussian curvature K (M^2_{21}) and the mean curvature H (M^3_{21}) increase significantly due to cheek raising and deepening of the nasolabial furrow. The ratio of maximum to mean intensity gradient (M^1_{21}) also increases, especially when smiling is very intense. In case of subtle smiles, lip corners and cheeks are gently pulled up thus cheek lines are not accentuated and image gradient does not change significantly. However, curvature changes are still detectable.

Similar changes may be observed when someone crinkles his nose. In this case, wrinkles may also appear in the nasolabial furrow, however cheek lines are not pulled up obliquely (as is the case for smile) but rather vertically.

3.2.2. Forehead wrinkling (M_{22})

Measurement of forehead wrinkling is based on edge detection inside an oblong area P_3 on the forehead, which is defined using the middle points of the upper eyebrow boundary segments, i.e. landmarks L_3 and L_{11} (see Fig. 6). To detect the presence of edges, a Canny edge detector is used. The appearance of wrinkles in the forehead, usually caused by eyebrow raising, results in significant increment of the percentage of pixels inside P_3 that correspond to edge points (M_{22}).

3.2.3. Nose wrinkling (M_{23})

When someone crinkles up his nose, usually to express disgust or displeasure, wrinkles appear along the lateral nose boundaries and in the glabella. To detect the presence of such wrinkles, we define polygons P_4-P_8 in the face area (see Fig. 6) and compute a set of measurements including intensity gradient changes in the 2D image and surface curvature measurements (Gaussian, mean and principal curvatures) in the corresponding 3D image.

In the glabella area we define the quadrangle P_4. When the face has a neutral expression, then only a few edges appear inside P_4. When the nose wrinkles (and the eyebrows are lowered), a bulge is produced between the eyebrows and the nasal root and numerous horizontal and vertical edges appear inside this rectangle. Using a Canny edge detector, we compute the ratio of the number of pixels indicating edge points to all pixels in P_4 (M^1_{23}). When a person crinkles up her nose, we usually observe a 15% or more increase of this ratio.

Pentagons P_5 and P_6 are defined on the lateral sides of the nose. When the nose wrinkles, then skin folds appear in this area, which result in changes of local curvature values. More specifically, we usually observe an increment of the mean of the absolute values of both K (M^2_{23}) and H (M^3_{23}) as well as an increment of the mean of the maximum principal curvature k_1 (M^4_{23}) in $P_5 \cup P_6$.

Pentagons P_7 and P_8 are located on the cheeks surface near the nose boundary. Experiments have shown that, in most cases, nose wrinkling is associated with a significant increment in the

Table 2
Face surface deformation measurements over areas P_1-P_8 shown in Fig. 6.

	Measurement	Computed in
M_{21}^1	Mean intensity gradient	$P_1 \cup P_2$
M_{21}^2	Mean *abs*(*K*)	$P_1 \cup P_2$
M_{21}^3	Mean *abs*(*H*)	$P_1 \cup P_2$
M_{22}	Edge density	P_3
M_{23}^1	Edge density	P_4
M_{23}^2	Mean *abs*(*K*)	$P_5 \cup P_6$
M_{23}^3	Mean *abs*(*H*)	$P_5 \cup P_6$
M_{23}^4	Mean k_1	$P_5 \cup P_6$
M_{23}^5	Mean *abs*(*K*)	$P_7 \cup P_8$
M_{23}^6	Mean *abs*(*H*)	$P_7 \cup P_8$
M_{23}^7	Mean k_1	$P_7 \cup P_8$
M_{23}^8	Mean k_2	$P_7 \cup P_8$

K, *H*, k_1, k_2 denote Gaussian, mean, maximum principal and minimum principal curvature, respectively.

absolute values of *K* and *H* and the mean value of k_1 as well as a decrement of the mean of the minimum principal curvature k_2 in $P_7 \cup P_8$. Similar changes may be observed in case of smiling faces. People laughing or smiling intensely raise their cheeks too, although in a different way compared to the way cheeks are raised when someone crinkles his nose. Nose wrinkling measurements denoted as $M_{23}^1-M_{23}^8$ are summarized in Table 2.

4. Facial action and facial expression classification

In this section, we present a classification scheme, which uses changes in facial measurement values to detect action units and recognize basic expressions in 2D+3D image sequences. We examine two scenarios. In the first scenario, classification is performed at every frame and independently of previous frames. In the second scenario, we exploit temporal information to detect an action unit or expression event after it has been concluded.

A rule-based approach is adopted in both cases based on comparing facial measurements extracted from a new image to measurements obtained from a reference (neutral face) image of the same subject. Given a test video sequence, we assume that in the first 5–10 frames the human subject has a neutral expression and we extract a measurement vector from each of these frames. Then we compute the median of each measurement M_i and form a reference measurement vector $\{R_i\}$ corresponding to the neutral face.

To recognize the facial expression or action unit appearing on a new frame, first we localize the positions of the 81 landmarks as described in Section 2. Then, we extract the set of facial measurements presented in Section 3 and finally we classify the depicted facial expression or action unit using a set of rules that compare these measurements to the reference measurement vector.

4.1. Rules for facial action unit and facial expression recognition

Rules used for facial action unit or facial expression recognition have been defined based on [2]. For each action unit or facial expression a list of associated appearance changes was determined and was subsequently translated in changes of facial measurement values. AU1, for example, describes raising of the inner portion of the eyebrow. This action unit not only increases the distance between the eyebrows and the eyes (or the eyebrows and the nasal root) but also sometimes causes horizontal skin wrinkles to appear in the center of the forehead and thus affects the values of measurements M_1, M_4 and M_{22}. The rules used for recognizing a subset of 11 important action units (AU1, AU2, AU4, AU5, AU7, AU9, AU12, AU15, AU25, AU26, AU27 [2]) and four facial expressions (happy, sad, surprise, disgust) are presented in Tables 3 and 4, respectively. In order to classify the observed changes, we first transform these changes into a set of parameters, which describe the increase or decrease in the value of a facial measurement M_i with respect to the corresponding value in a neutral expression R_i (see caption in Table 3).

Rules for facial expression recognition are more complex since they take into account measurements computed on different parts of the face. Moreover, the same facial expression may be manifested in many ways thus one rule may include several sub-rules. For example, happy (smiling) expressions are associated with lip corners being raised obliquely, lower lip getting a $\smile$ shape, wrinkles appearing on the cheeks and eyelids narrowing. All changes however may not manifest themselves at the same time. To encode different expressions of happy, two rules have been defined.

The first rule is used to describe cases when smiling is intense, causing lip corners to rise significantly and cheek wrinkles to appear or become more intense if already present. This can be translated in the following changes in facial measurement values: the length of the lower lip line (M_{11}) increases, the concavity of the lower lip line (M_{12}) has a positive value, the cheek lines angle (M_9) and the angle of nose–mouth corner lines (M_{16}) increase and the mouth corners to eyes distance (M_{17}) decreases. Finally, cheek wrinkling measurements (M_{21}) also increase.

The second rule refers to cases where the lip corners are not raised obliquely but rather along the horizontal axis. Thus, cheek wrinkling (M_{21}) is detectable, but the lower lip does not have the characteristic $\smile$ shape, even though it is elongated (M_{11} and M_{13} increase). This usually happens when the human subject is not very happy, but nevertheless tries to smile.

4.2. Facial action unit and facial expression recognition using temporal dynamics

In the second scenario, the temporal variation of facial measurements is analyzed to detect action units or facial expressions. We assume that periods of facial activity, called action unit or facial expression events in the sequel, are proceeded and followed by periods of no facial activity, i.e. periods where the subject has a neutral expression. For example in case AU1 is activated in a time interval $[t_1, t_2]$, the values of M_1, M_4 and M_{22} will increase near t_1, reach a peak value between t_1 and t_2 and decrease near t_2. We call this temporal pattern a facial measurement event. We assume that a measurement event starts when the value of the corresponding measurement M_i starts to increase/decrease compared to the reference value R_i and ends when this value becomes again equal to R_i. In the between time, the value of M_i reaches a peak (maximum/minimum) value. Classification is based on this peak value.

Detection of measurement events over time is based on the ratio $Q_i = (M_i\text{-}R_i)/R_i$, which represents the increment or decrement of M_i compared to R_i. The beginning of a potential new event is signalled by $|Q_i|$ becoming greater than a threshold T_n and the end of the event by $|Q_i|$ becoming less than T_n and remaining so for at least 1 s. The value of T_n depends on measurement noise but for the majority of measurements was set to 5%. There is a case that a detected event is caused by an erroneous measurement. This case can be easily detected by testing duration and rejecting brief events (e.g. less than 0.5 s). We also neglect events if their peak measurement value is less than 10%.

Table 3
Rules for recognizing facial action units.

AU1	Raises the inner eyebrow part IF inc(M_1) > 10 OR inc(M_4) > 10 OR inc(M_{22}) > 30 THEN AU1 = true
AU2	Raises the outer eyebrow part IF inc(M_2) > 12 THEN AU2 = true
AU4	Lowers the eyebrows IF (dec(M_1) > 10 OR dec(M_4) > 10) AND (dec(M_3) > 10 OR inc(M^1_{23}) > 15) THEN AU4 = true
AU5	Raises the upper eyelid, widens the eye opening IF inc(M_5) > 12 AND inc(M_6) > 10 THEN AU5 = true
AU7	Raises the lower eyelid, narrows the eye opening IF dec(M_5) > 10 AND dec(M_6) > 10 THEN AU7 = true
AU9	Wrinkles the nose IF dec(M_4) > 10 AND (inc(M^1_{23}) > 15 OR dec(M_3) > 10) AND (($NW_1 \geq 2$ AND $NW_2 \geq 3$) OR ($NW_1 \geq 2$ AND (dec(M_7) > 10 OR inc(M_8) > 10)) OR ($NW_2 \geq 3$ AND (dec(M_7) > 10 OR inc(M_8) > 10))) THEN AU9 = true $NW_1 = H(\text{inc}(M^2_{23}) > 20) + H(\text{inc}(M^3_{23}) > 20) + H(\text{inc}(M^4_{23}) > 20)$ $NW_2 = H(\text{inc}(M^5_{23}) > 30) + H(\text{inc}(M^6_{23}) > 20) + H(\text{inc}(M^7_{23}) > 30) + H(\text{dec}(M^8_{23}) > 10)$
AU12	Pulls lip corners upwards obliquely IF inc(M_{11}) > 5 AND inc(M_{12}) AND $M_{12} > 5^\circ$ AND dec(M_{17}) > 5 AND inc(M_{16}) > 8 THEN AU12 = true
AU15	Presses lip corners downwards IF $M_{14} = 0$ AND dec(M_{12}) AND $M_{12} < -5^\circ$ AND inc(M_{18}) > 8 AND (NOT dec(M_{19}) > 15) THEN AU15 = true
AU25	Parts the lips slightly IF inc(M_{14}) AND $M_{14} > 0.3$ cm AND $M_{14} < 1$ cm AND inc(M_{20}) AND (NOT inc(M_{20}) > 10) THEN AU25 = true
AU26	Parts the lips, parts the jaws IF $M_{14} \geq 1$ cm AND inc(M_{20}) > 10 AND (NOT inc(M_{20}) > 80) AND (NOT dec(M_{19}) > 10) THEN AU26 = true
AU27	Stretches the mouth and pulls the lower jaw downwards IF $M_{14} \geq 1$ cm AND inc(M_{20}) > 80 THEN AU27 = true

M_i is the value of measurement i computed in the current frame and R_i is the corresponding reference measurement. inc(M_i) > a (dec(M_i) > a) denotes an increment (decrement) of more than a% in the value of M_i compared to R_i. inc(M_i)/dec(M_i) denotes that the value of M_i has increased/decreased. $H(h_1)$ equals 1 if the hypothesis h_1 is correct and 0 otherwise. Threshold values were determined experimentally.

Table 4
Rules for recognizing facial expressions (see caption in Table 3).

E1	Disgust IF AU9 = true OR (dec(M_{12}) AND $M_{12} < -5^\circ$ AND dec(M_{19}) > 15 AND $CL \geq 2$) THEN E1 = true $CL = H(\text{inc}(M^1_{21}) > 10) + H(\text{inc}(M^2_{21}) > 20) + H(\text{inc}(M^3_{21}) > 15)$
E2	Happy IF inc(M_{11}) > 10 AND inc(M_{12}) AND $M_{12} > 5^\circ$ AND (inc(M_9) > 8 OR inc(M_{16}) > 8) AND dec(M_{17}) > 10 AND inc(M^1_{21}) > 10 AND inc(M^2_{21}) > 30 AND inc(M^3_{21}) > 20 AND AU9 = false THEN E2 = true IF inc(M_{11}) > 5 AND inc(M_{13}) > 5 AND $M_{14} = 0$ AND (inc(M_9) > 8 OR inc(M_{16}) > 8) AND inc(M^1_{21}) > 10 AND inc(M^2_{21}) > 30 AND inc(M^3_{21}) > 20 AND AU9 = false THEN E2 = true
E3	Sad IF dec(M_{12}) AND $M_{12} < -4^\circ$ AND inc(M_{18}) > 8 AND (inc(M_4) > 10 OR dec(M_3) > 10 OR dec(M_5) > 10) AND (NOT dec(M_{19}) > 15) AND AU9 = false THEN E3 = true
E4	Surprise IF inc(M_1) > 10 AND inc(M_2) > 10 AND inc(M_5) > 15 AND $M_{15} > 0.25$ AND $M_{14} \geq 1$ cm AND inc(M_{20}) > 15 THEN E4 = true IF inc(M_1) > 15 AND inc(M_2) > 15 AND inc(M_5) > 15 THEN E4 = true

Since each measurement is processed independently, concurrent (overlapping) events may be detected. These are subsequently combined to recognize facial action units and facial expressions. More specifically, at each frame we examine whether we have one or more concluded measurement events and create a list of the rules associated with these measurements. Then for each rule we examine whether all associated measurements are present. If they are, then we have a potential action unit or facial expression event that is verified using the peak values of associated measurement events. If the result is true, then an action unit or facial expression event is detected in the interval bounded by the corresponding measurements duration. The same procedure is followed for all rules in the list.

Fig. 7 illustrates examples of detecting nose wrinkling (AU9) and surprise events.

5. Experimental results

In this section, we present the results obtained from the experimental evaluation of the system components. To this end a new 2D+3D image database was recorded using the prototype 3D sensor presented in [19]. The sensor is based on color coded light implemented in the invisible near infrared spectrum and is capable of quasi-synchronous acquisition of 3D and grayscale images. The resolution of generated images is 582 × 782 pixels, while the accuracy of depth data is better than 0.3 mm for objects standing at a mean distance of 60 cm in a working volume of $50 \times 50 \times 50\,\text{cm}^3$.

The database consists of 832 sequences of 52 participants, 12 female and 40 male, 24 to 40 years old. In each sequence, the human subject displays a single action unit (11 in total) or mimics a facial expression (happy, sad, disgust, surprise, neutral) 2–4 times. Facial action periods last approximately 5–10 s and are proceeded and followed by short neutral state periods. The duration of each recording is about 30–40 s and the frame rate is about 5 fps. Facial action and neutral face periods were manually identified in each of these sequences by an expert and an appropriate tag was assigned to each frame. Examples of recorded image pairs and image sequences are illustrated in Figs. 8 and 9.

First, we evaluate the performance of the 3D face tracker presented in Section 2. To train the global model as well as the local detectors we used a set of 400 image pairs depicting an

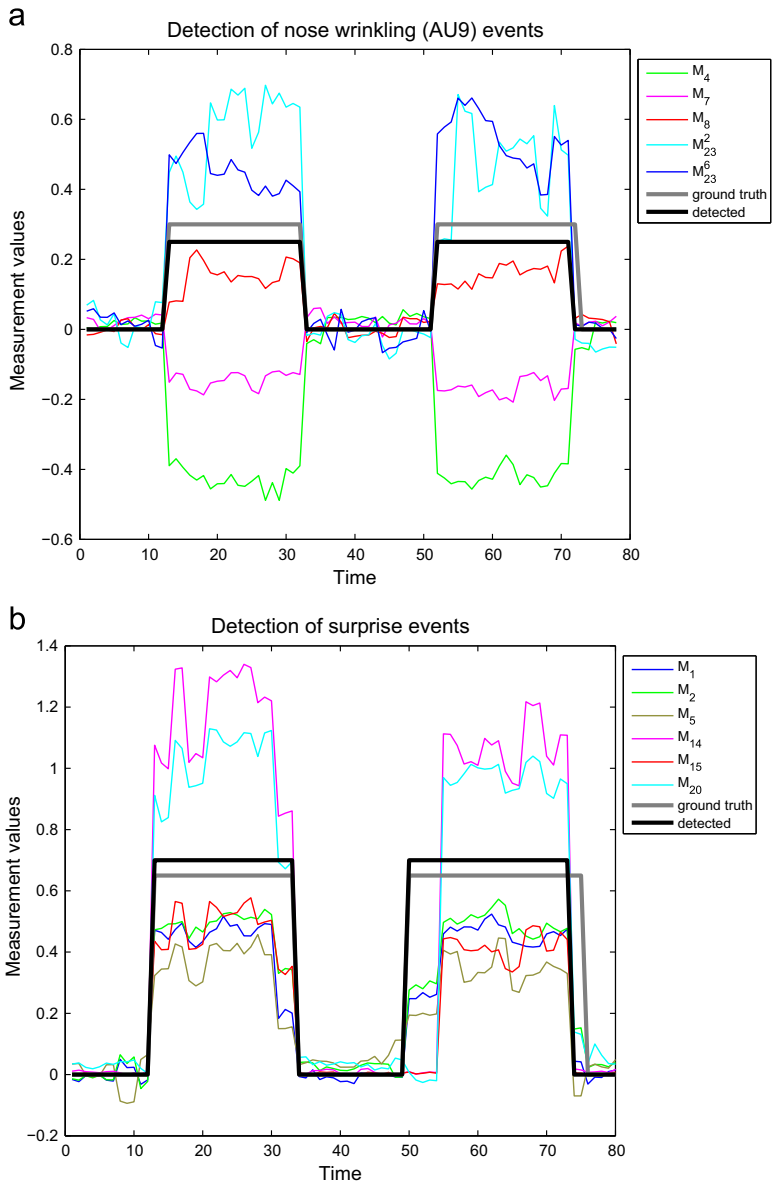

Fig. 7. Examples of temporal event detection: (a) AU9 and (b) surprise. The $(M_i\text{-}R_i)/R_i$ values of associated measurements M_i are shown. For M_{14} and M_{15} the $M_i\text{-}R_i$ values are shown instead. The values of M_{14} (mouth opening) were normalized by division with a constant. The bold gray line indicates the ground-truth activation periods and the bold black line indicates activation periods detected by the proposed method.

action unit or facial expression at its peak. To test the face tracker, we use another set of 600 images, where we manually mark the positions of facial landmarks. The estimated feature positions are compared against their ground-truth positions. Using the proposed face tracker, we achieve a mean localization error of 5.35 pixels and standard deviation 2.2, when the mean face dimensions are 280×370 pixels. On the contrary using the global detector only, the corresponding error is 7.8 pixels. We also compare the 2D+3D tracker against a 2D only ASM with the same 81 landmarks. In this case, we obtain a localization error of 10.2 pixels, which is mainly attributed to erroneous estimation of open mouth landmarks.

Next, we evaluate the performance of the facial action unit detector and the facial expression classifier under the first classification scenario. The first 10 frames of each sequence are used to extract the reference measurement vector. In each of the remaining frames, first we localize the positions of the 81 facial landmarks, next we extract a facial measurement vector and finally we (a) classify the user's facial expression or (b) detect a set of action units based on the rule-based approach presented in Section 4. Using this procedure we assign to each frame a single facial expression tag and one or more action unit tags. These tags are subsequently compared against the ground truth.

The action unit detector was tested in $52 \times 11 = 572$ test sequences, i.e. 52 sequences per action unit (one per subject). The evaluation results are illustrated in Fig. 10. The mean detection rate is 83.6%. The lowest detection rate is observed for AU27 (mouth stretched). The latter can be explained by the fact that when the mouth is widely open, pixels in the mouth area have undetermined depth values thus leading to erroneous estimates of lip boundaries. More specifically, the estimation of the lower lip boundary provided by the local mouth detector is usually placed closer to the upper lip than it actually is thus resulting to the detection of AU26 instead of AU27. A relative low detection rate is also observed for AU15 (lip corners pressed down). This is mainly

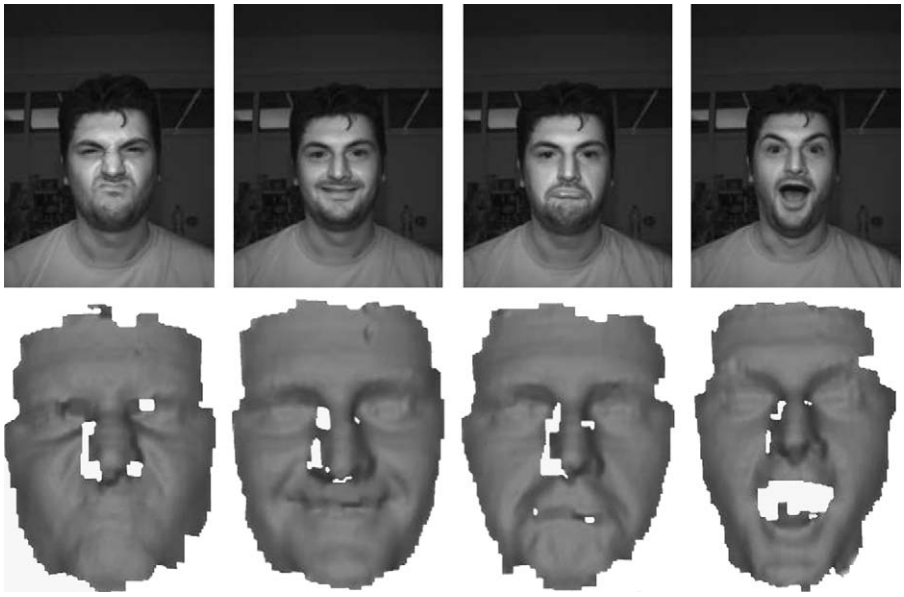

Fig. 8. Examples of grayscale images and corresponding 3D models of the facial expression database. The latter are generated from the recorded 3D images.

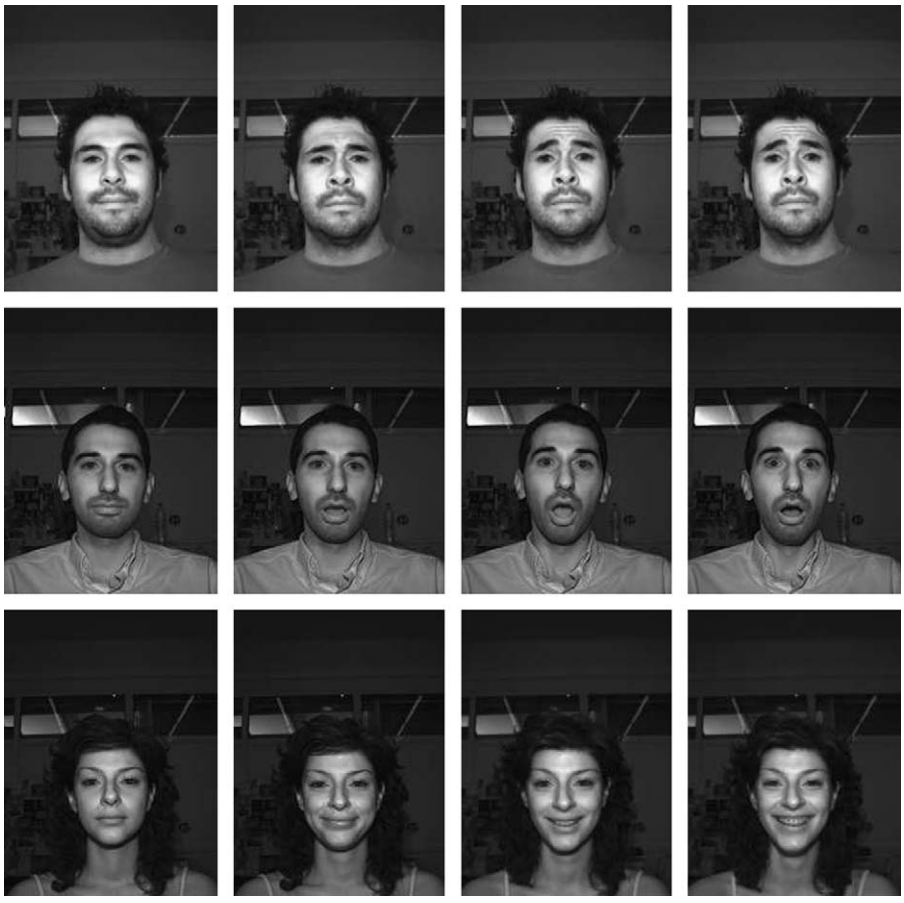

Fig. 9. Examples of recorded image sequences. First row: male subject displaying AU1. Second row: male subject expressing surprise. Third row: female subject expressing happiness.

due to the fact that most subjects displayed the specific action unit very subtly, thus no significant changes could be detected in the lip boundary shape and convexity.

Next, we evaluate the proposed facial expression recognition technique again under the first classification scenario. As already explained, our system is able to recognize facial expressions related to happiness, sadness, surprise and disgust. Facial expressions of anger and fear can also be detected though less reliably. For the evaluation of the facial expression classifier, we used $52 \times 5 = 260$ test sequences, i.e. five sequences per subject

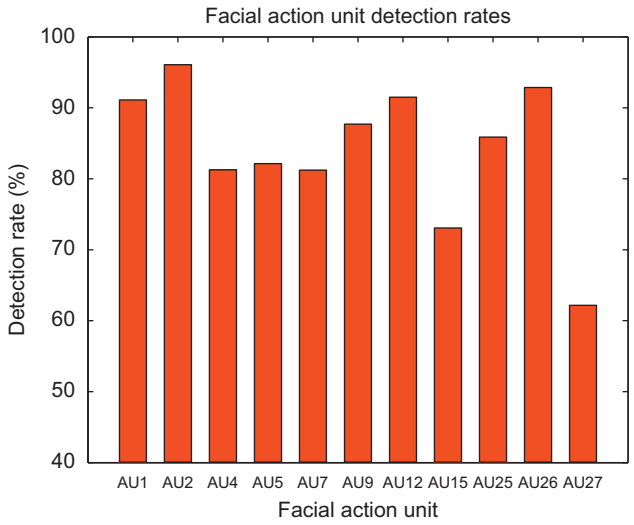

Fig. 10. Facial action unit detection rates under the per frame detection scenario.

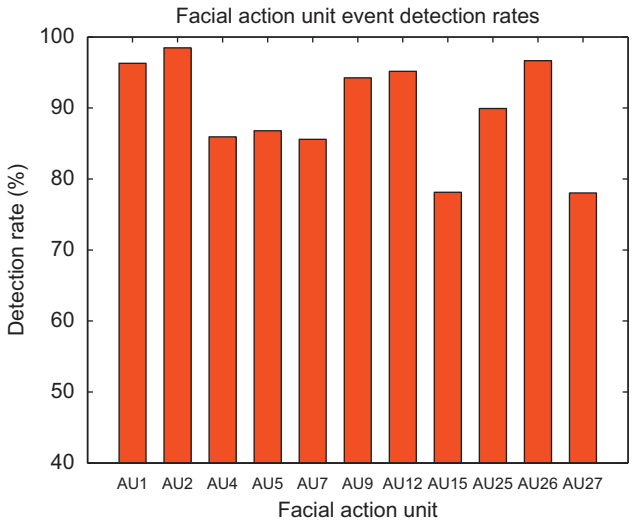

Fig. 11. Correct detection rates for action unit event recognition.

Table 5
Facial expression recognition rates (%) under the per frame classification scenario.

True/classified	Neutral	Disgust	Happy	Sad	Surprise
Neutral	**95.42**	0.92	1.87	1.31	0.48
Disgust	5.58	**82.63**	5.92	5.76	0.11
Happy	5.69	1.47	**90.84**	0.27	1.73
Sad	23.25	2.40	0.38	**73.97**	0.00
Surprise	2.53	0.75	9.12	0.00	**87.60**

Table 6
Recognition rates for facial expression events (%).

True event/classified	Neutral	Disgust	Happy	Sad	Surprise	Other
Disgust	1.25	**89.37**	2.50	1.87	0.00	5.01
Happy	1.23	0.00	**95.06**	0.00	0.00	3.71
Sad	11.95	0.63	0.00	**79.24**	0.00	8.18
Surprise	0.61	0.00	1.83	0.00	**93.90**	3.66

(four expressions + neutral). The evaluation results are presented as a confusion matrix in Table 5. The element (i,j) of this table represents the percentage of sequence frames depicting expression i, which were assigned emotion label j. The average expression recognition rate is 85%. The highest misclassification error is reported for sad, which 1 out of 4 times is classified as neutral. This can be attributed to the fact that most subjects expressed sadness only by slightly pressing lip corners down and it is directly associated with the relative low detection rate observed for AU15.

The same image sequences are also used for evaluating the proposed temporal event classifier. As in the case of the first classification scenario, the first 10 frames of each sequence are used to compute the reference measurement vector. For the remaining frames, we use the technique described in Section 4.2 and detect a set of events corresponding to action unit or facial expression activation periods. To evaluate the performance of the proposed temporal event detector, we compare the detected events against the ground truth activation periods based on the following assumption: we assume that a ground truth event is correctly identified if the overlapping between this event and a detected event with the same label is more than 80%. If the overlapping is below the aforementioned threshold or an event with another label is identified in the same time segment (e.g. disgust instead of happy) then the ground truth event is considered not detected. Fig. 11 presents the correct detection rates for different action unit events. A mean detection rate of 89.50% is achieved.

The results of temporal facial expression recognition are presented in Table 6. The element (i,j) of this table represents the percentage of ground truth events depicting expression i, which were recognized as events of expression j. We consider that a ground truth event e_g of expression i is assigned a label j, when there is an overlapping of more than 80% between the ground truth event and a detected event of expression j. If no event is detected during the activation period of e_g or more than 80% of the frames in this period are identified as neutral then this event is assigned a neutral label. In any other case, e.g. if there is an overlapping of less than 80% with a detected event or more than one events of different expressions were detected in this time period, e_g is classified as "other".

Table 7
Facial expression recognition rates (%) obtained for the proposed 2D+3D classifier and the 2D appearance-based classifier under the per frame classification scenario.

	2D+3D	2D
Neutral	95.42	83.60
Disgust	82.63	70.72
Happy	90.84	81.21
Sad	73.97	61.85
Surprise	87.60	79.75

Next, we evaluate the benefits obtained from the use of 3D facial data by comparing the proposed facial expression recognition system against a system based exclusively on 2D images. The latter consists of a 2D facial feature tracker based on 2D ASMs (a model for the whole face, one for the eyebrows and one for the mouth exactly as in the case of the proposed 3D face tracker) and a facial expression classifier based on Gabor filters and linear discriminant analysis. Given a 2D image frame, first we localize the position of the 81 facial landmarks using the global 2D ASM and the local feature detectors. Over each landmark position we compute a local brightness measurement vector by applying a set of Gabor filters and we create a concatenated feature vector for the whole face. The feature vector is then projected in an LDA

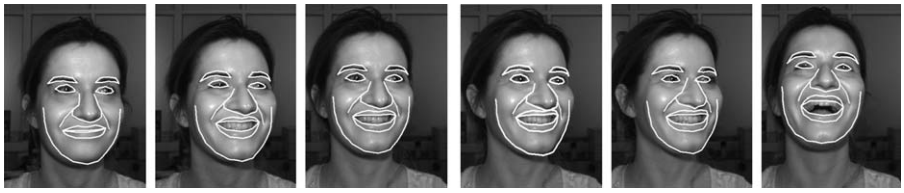

Fig. 12. Example of image sequence showing a happy expression under pose variations. White lines correspond to tracking results.

subspace giving rise to a discriminant feature vector, which is finally classified in one of the five emotion classes by means of the K-nearest neighbors technique. This technique was tested in the same set of sequences used for the evaluation of the proposed 3D system. Table 7 compares its performance against that of the 3D system under the first classification scenario. It is clear that use of 3D face geometry information significantly aids facial expression recognition. Examining the results we found out that this improvement can be equally attributed not only to increased localization accuracy (Gabor features partly compensate for mislocalization) but also to 3D features such as nose or cheek wrinkling curvature measurements and their identity invariant calculation. For this reason, we believe that our approach could be better in terms of recognition accuracy compared to others, e.g. [18], that use 3D information only for tracking and rely on 2D features or implicit 3D features (e.g. model shape parameters) for recognition. The use of explicit 3D facial measurements may facilitate detection of specific action units, e.g. AU9, substantially.

The 3D system significantly outperforms the 2D one if there is a lot of head movement as demonstrated by the following experiment. We have recorded eight image sequences showing four human subjects expressing happiness and surprise (two sequences per subject) while rotating their heads up to 30° to the left and right. An example of a recorded sequence is illustrated in Fig. 12. Facial feature tracking results are also displayed. We have followed the evaluation procedure described above for the first classification scenario and obtained a 87% recognition rate for happy and 84% for surprise. In case of larger poses where almost half of the face is occluded, the recognition rate may drop significantly. But this is mainly due to the failure of the facial feature tracking algorithm, which results in erroneous facial measurements.

Acquisition of 2D+3D image sequences at higher frame rates would result in smaller displacements of facial landmarks between subsequent frames, which should in turn result in faster and more accurate facial feature localization and thus increased facial measurement accuracy and improved facial expression recognition performance. Moreover, a higher frame rate would enable detection of fleeting expressions, which are difficult to detect at low frame rates.

Experiments were performed on an Intel Core Duo 2.0 GHz PC with 4 GB RAM. The total time for processing a single frame is between 0.1 and 0.3 s: 50 ms for face detection, 0.15–0.25 s for facial feature extraction and 10 ms for facial expression recognition.

6. Conclusion

A fully automated system for facial action unit detection and facial expression recognition in sequences of 2D and 3D images was presented in this paper. The proposed system is based on a novel real-time model-based face tracker and a set of special local feature detectors, which effectively combine 3D face geometry and 2D appearance data. The use of 3D information facilitates detection of surface deformations even in case of subtle facial muscle movements. Facial action is represented by a set of geometric, appearance-based and surface-based measurements, which are effectively classified into emotional related expressions using a rule-based approach. A method for detecting temporal events related to action unit or facial expression activation periods was also proposed. The proposed techniques were evaluated in a large database with more than 50 subjects and 800 sequences demonstrating increased accuracy and robustness under pose variations.

Future work will exploit the dynamics of facial measurements towards automatic decoding of all action units and their combinations. 3D information will be further exploited for facial feature tracking and facial action unit recognition. More specifically, the standard ASM fitting technique, which is based on image gradient profiles derived from 2D images, will be extended to include local 3D surface information in the form of curvature or surface gradient descriptors. Such an approach is expected to offer increased localization accuracy as well as increased robustness against pose and illumination variations. In addition, new curvature measurements will be proposed for detecting action units related to the mouth and chin. Finally, the proposed techniques will be extended to cope with large head poses.

Acknowledgments

This work was supported by research project “PASION-Psychologically Augmented Social Interaction over Networks” (FP6-IST-027654) under the Information Society Technologies (IST) priority of the 6th Framework Programme of the European Community.

References

[1] M. Lewis, J. Jones, L. Barrett, Handbook of Emotions, Guilford Publications, Inc, 2008.
[2] P. Ekman, W.V. Friesen, The Facial Action Coding System: A technique for measurement of facial movement, Consulting Psychologists Press, Palo Alto, CA, 1978.
[3] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1424–1445.
[4] I. Kotsia, I. Pitas, Facial expression recognition in image sequences using geometric deformation features and support vector machines, IEEE Transactions on Image Processing 16 (1) (2007) 172–187.
[5] M.S. Bartlett, G. Littlewort, I. Fasel, J.R. Movellan, Real time face detection and facial expression recognition: development and application to human computer interaction, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, vol. 5, 2003, p. 53.
[6] M. Black, Y. Yacoob, Recognizing facial expressions in image sequences using local parameterized models of image motion, International Journal of Computer Vision 25 (1) (1997) 23–48.
[7] M. Pantic, L. Rothkrantz, Facial action recognition for facial expression analysis from static face images, IEEE Transactions on Systems, Man, and Cybernetics—Part B 34 (3) (2004) 1449–1461.
[8] I. Cohen, N. Sebe, A. Garg, L. Chen, T. Huang, Facial expression recognition from video sequences: temporal and static modeling, Computer Vision and Image Understanding 91 (2003) 160–187.
[9] M. Pantic, I. Patras, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, IEEE Transactions on Systems, Man, and Cybernetics—Part B 36 (2) (2006) 433–449.
[10] Z. Wen, T. Huang, Capturing subtle facial motions in 3D face tracking, in: Proceedings of the IEEE International Conference on Computer Vision, vol. 2, 2003, pp. 1343–1350.

[11] J. Wang, L. Yin, X. Wei, Y. Sun, 3D facial expression recognition based on primitive surface feature distribution, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1399–1406.
[12] H. Tang, T. Huang, 3D facial expression recognition based on automatically selected features, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–8.
[13] H. Soyel, H. Demirel, Facial expression recognition using 3D facial feature distances, in: Proceedings of the International Conference on Image Analysis and Recognition, Lecture Notes in Computer Science, vol. 4633, Springer, Berlin, 2007, pp. 831–838.
[14] I. Mpiperis, S. Malassiotis, M.G. Strintzis, Bilinear models for 3-D face and facial expression recognition, IEEE Transactions on Information Forensics and Security 3 (3) (2008) 498–511.
[15] X. Huang, S. Zhang, Y. Wang, D. Metaxas, D. Samaras, A hierarchical framework for high resolution facial expression tracking, in: Proceedings of the IEEE Workshop on Articulated and Nonrigid Motion, 2004, p. 22.
[16] Y. Chang, M. Vieira, M. Turk, L. Velho, Automatic 3D facial expression analysis in videos, in: Proceedings of the 2nd International Workshop on Analysis and Modelling of Faces and Gestures, Lecture Notes in Computer Science, vol. 3723, Springer, Berlin, 2005, pp. 293–307.
[17] Y. Sun, L. Yin, Facial expression recognition based on 3D dynamic range model sequences, in: Proceedings of the 10th European Conference on Computer Vision, Lecture Notes in Computer Science, vol. 5303, Springer, Berlin, 2008, pp. 58–71.
[18] J. Liebelt, J. Xiao, J. Yang, Robust AAM fitting by fusion of images and disparity data, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2483–2490.
[19] D. Modrow, C. Laloni, G. Doemens, G. Rigoll, A novel sensor system for 3D face scanning based on infrared coded light, in: Proceedings of the SPIE Conference on Three-Dimensional Image Capture and Applications 2008, vol. 6805, 2008.
[20] F. Tsalakanidou, S. Malassiotis, Robust facial action recognition from real-time 3D streams, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009, pp. 4–11.
[21] A. Lanitis, C.J. Taylor, T.F. Cootes, Automatic interpretation and coding of face images using flexible models, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 743–756.
[22] S. Malassiotis, M.G. Strintzis, Robust face recognition using 2D and 3D data: pose and illumination compensation, Pattern Recognition 38 (12) (2005) 2537–2548.
[23] S. Malassiotis, M.G. Strintzis, Robust real-time 3D head pose estimation from range data, Pattern Recognition 38 (8) (2005) 1153–1165.
[24] B. Horn, Closed-form solution of absolute orientation using unit quaternions, Journal of the Optical Society of America A 4 (4) (1987) 629–642.

About the Author—FILARETI TSALAKANIDOU received the Diploma and Ph.D. degrees in Electrical and Computer Engineering from the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2000 and 2006, respectively. Currently, she is a post doctoral researcher with the Informatics and Telematics Institute, Thessaloniki, Greece where she has been working as a postgraduate research assistant since 2000. Her research interests include 3D image processing, 3D biometrics and affective computing.

About the Author—SOTIRIS MALASSIOTIS was born in Thessaloniki, Greece, in 1971. He received the B.S. and Ph.D. degrees in Electrical Engineering from the Aristotle University of Thessaloniki, in 1993 and 1998, respectively. From 1994 to 1997 he was conducting research in the Information Processing Laboratory of the Aristotle University of Thessaloniki. He is currently a senior researcher in the Informatics and Telematics Institute, Thessaloniki. He has participated in several European and National research projects. He is the author of more than 20 articles in refereed journals and more than 30 papers in international conferences. His research interests include range image analysis, pattern recognition, and computer graphics.