Robust facial action recognition from real-time 3D streams

Filareti Tsalakanidou and Sotiris Malassiotis
Informatics and Telematics Institute, Centre for Research and Technology Hellas
6th km Charilaou-Thermi Road, Thessaloniki 57001, Greece
filareti@iti.gr, malasiot@iti.gr

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. Good performance is achieved under relatively non-controlled conditions.

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 [11]. 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 psychologists, neuroscientists and lately computer vision researchers [5][15].

Several approaches have been reported towards automatic facial expression recognition from 2D static images or video sequences [15]. 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 analyzers recognize expressions corresponding to the basic emotions, i.e. happiness, sadness, anger, fear, surprise and disgust, the display of which is thought 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 [5].

Facial features used for expression recognition are geometric, e.g. distances between facial points [9], appearance-based such as Gabor filter responses [1] or holistic such as optical flow fields [2]. Classification methods can be roughly divided to static and dynamic ones. Static classifiers use feature vectors related to a single frame. 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 [9, 16]. Temporal classifiers on the other hand try to capture the temporal pattern in the sequence of feature vectors related to each frame [4].

A problem with existing techniques is that the subtle skin deformations that characterize facial expressions are not captured 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 [16] or 3D images. Although the advantages of using 3D facial images are self evident, very few works have examined 3D facial expression recognition. Wang *et al*. [8] use a surface labelling approach based on the distribution of principal curvature descriptors defined over different face regions. Tang and Huang [19] propose a feature selection technique based on maximizing the entropy of class conditional feature distributions and apply it on a pool of normalized 3D Euclidean distances between preselected feature points. These distances are subsequently classified using AdaBoost and a set of different classifiers. A similar approach is followed by Soyel and Demirel [17], which employ six characteristic facial feature distances (eye opening, mouth width, etc) and a probabilistic neural network classifier. These works do not address the problem of face localization, i.e. facial feature points are manually selected. 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. A hier-

This work was supported by the EU FP6 project "PASION: Psychologically Augmented Social Interaction over Networks"(IST-027654).

archical framework based on deformable shape models is proposed by Huang *et al*. in [7] for tracking facial expressions in sequences of face scans. Fitting between face models and range scans is based in cubic B-spline based Free Form Deformations. No facial expression recognition results are reported in this work, while an initial fitting between the model and the first frame scan is manually performed. Chang *et al*. [3] propose a framework for facial expression analysis and editing based on a generalized expression manifold, which is built by transferring facial deformations computed in range sequences of six subjects to a standard face model. Face tracking is based on 2D feature tracking followed by fitting a coarse 3D mesh model. In [18], a spatio-temporal approach is adopted based on primitive 3D surface descriptors [8] 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.

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 [13]. To achieve real-time performance we do not rely on dense mesh registration algorithms, 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 4 facial expressions as well as 10 facial action units using a rule-based approach. Finally, the efficiency of the 3D face analyzer is evaluated in a database with more than 50 subjects and 1000 sequences.

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:

- Unlike related 2D or multiview tracking techniques, the proposed tracker is drift-free, works robustly on noisy or incomplete 3D data and does not require any first frame initialization, calibration or per user adaptation. In addition, it can withstand moderate pose and illumination variations.

- We explore 3D geometric distances as well as 3D surface features. This alleviates the need for feature normalization and enables detection of subtle facial deformations around the nose and 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 set of rules for facial expression and facial action unit classification is outlined in Section 4. The performance of the proposed system is evaluated in Section 5. Finally, Section 6 concludes the paper.

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 [10], 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. Although ASMs have been demonstrated less accurate than AAMs, they have the advantage of robustness to illumination variations (using local gradient search) and are very efficient computationally.

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 [13] (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.a. The PDM is then expressed as

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

where $\mathbf{s} = \{x_1, y_1, z_1, ..., x_n, y_n, z_n\}$ is the vector of n landmark coordinates, $\mathbf{s}_i$ are the basis shapes computed by applying the 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 here, that image alignment involves 3D rotation and translation of original image pairs so that all faces have

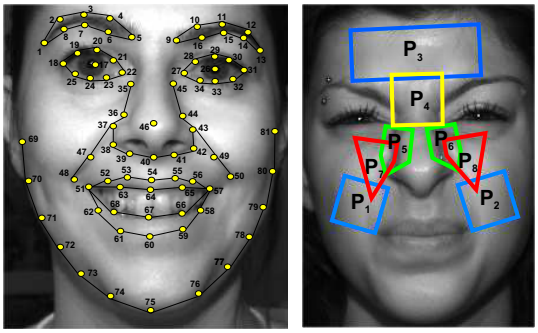

Figure 1. a) The 81 landmarks and corresponding segments of the ASM. b) Polygon areas defined in the face surface for detecting the presence of wrinkles (see Section 3).

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 [12].

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.a. 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 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

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

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 built a local model of gradient changes associated with this feature 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

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

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 ASM 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 3D pose estimation technique proposed in [12]. 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 the Horn's quaternion method [6]. 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 proportional 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 and if the model has not converged we re-initialize the tracker, i.e. repeat face detection, pose estimation and model fitting. For faster convergence we use a multi-resolution scheme with three layers.

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 Fig. 2. 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

Eyebrow localization may be inaccurate if the eyebrow hair is light coloured or very sparse or the eyebrow itself is very thin. In such cases, gradient changes in eyebrow boundaries may be very subtle thus leading to erroneous landmark estimates.

To enhance the eyebrows estimation, we use a local 3D ASM with 16 landmarks and introduce a new model fitting technique. Instead of matching the local gradient profiles of candidate points against the local appearance model, we employ a simpler selection criterion based on area intensity differences. For each candidate landmark position, we define two rectangle areas lying on the positive and negative values of the axis defined by the normal in this position, and we compute their average intensities S_1 and S_2. Since the eyebrow landmarks lie in the boundary between dark (eyebrow hair) and light-coloured (skin) areas, 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 be 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.

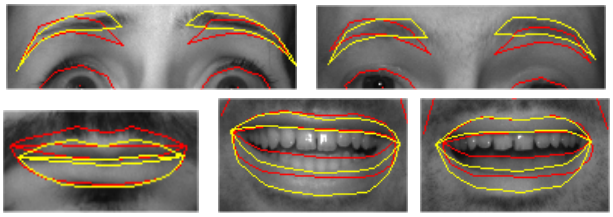

Figure 2. Examples of eyebrow and mouth boundary localization using the global model (red line) and local detectors (yellow line).

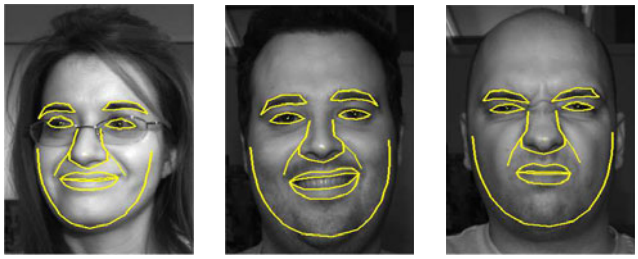

Figure 3. Examples of facial feature tracking using the proposed global tracker and local feature detectors.

The proposed 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 here that in this case, landmark points are weighted proportional to the corresponding intensity difference $S_1 - S_2$. As can be seen in Fig. 2, the proposed local eyebrow detector greatly enhances the estimation provided by the global ASM.

2.2. Local mouth detector

Lip boundaries are also problematic due to the unimodal Gaussian distribution assumption used for the representation of local mouth appearance patterns, which however is not appropriate for landmarks lying in the inner lip boundaries, since their local gradient patterns are significantly affected by whether the mouth is open or closed. 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 openness 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. After the mouth is classified as open or closed, the corresponding model, comprised of 18 points corresponding to lip boundaries, is fitted on the face. Model fitting is based on image gradient profiles, but instead of

	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.	$d_{54,40}$

Table 1. Geometric facial measurements. $d_{i,j}$ is the 3D Euclidean distance between landmarks L_i, L_j. $\varepsilon_{i,j}$ is the 2D line defined by the projections of L_i, 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.

searching for candidate points along the normal, we search in a narrow rectangular area centred on the current landmark position.

Experimental results show that the local mouth detector significantly improves the initial lip boundary estimation, especially when the mouth is open (see Fig. 2).

3. Extraction of facial feature measurements

To encode facial movements, we adopt the Facial Action Coding System (FACS) developed by Ekman and Friesen [5]. In this system, 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. To detect the presence of an AU, several geometric and surface deformation measurements have to be computed, but once a system detects these 44 AUs, then the emotional state of the human subject can be inferred usually with the help of context information. In our case, AU estimation is achieved by a means of 22 geometric, appearance and surface deformation measurements denoted as M_1-M_{22}.

Geometric measurements such as eye opening, eyebrow displacement, mouth shape, etc are computed using the estimated positions of the 81 landmarks L_i depicted in Fig. 1.a. We have developed a set of 19 measurements, which are presented in Table 1.

Surface deformation measurements are associated with wrinkles appearing on the skin due to muscle contractions. These include cheek wrinkling (M_{20}), forehead wrinkling (M_{21}) and nose wrinkling (M_{22}). The approximate position of the wrinkles is easily determined using the estimated landmark positions.

To detect the presence of cheek wrinkles, we define two rectangles P_1 and P_2 enclosing the left and right cheek lines and cheek surface and we compute the average intensity gradient perpendicular to segments defined by the 2D projections of landmarks L_{47}, L_{48} and L_{49}, L_{50} respectively (see Fig. 1.b). We also compute the mean surface gradient and mean Gaussian curvature inside these rectangles using the corresponding 3D image. When cheek wrinkles appear in the face, then the mean Gaussian curvature (M^3_{20}) increases significantly due to cheek raising. The ratio of maximum to mean intensity gradient (M^1_{20}) and the mean depth gradient (M^2_{20}) also increase, 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. However, depth gradient changes are still detectable.

Measurement of forehead wrinkling is based on edge tracing inside a rectangle area P_3 on the forehead, which is defined using the middle points of the upper eyebrow boundary segments. The appearance of wrinkles in the forehead, usually caused by eyebrow raising, results in significant increment of the percentage of pixels representing edge points (M_{21}).

Finally, when someone crinkles up his nose, usually to express disgust or displeasure, wrinkles appear in the nose surface, 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. 1.b) and compute a set of 6 measurements including intensity gradient changes in the 2D image and surface curvature measurements (Gaussian, mean and principal curvatures) in the corresponding 3D image. Nose wrinkling is associated with a) increment of the number of edges in P_4 usually accompanied by increment of surface gradient, b) increment of mean depth gradient, mean Gaussian and mean curvature in $P_5 \cup P_6$ and c) increment of mean Gaussian curvature in $P_7 \cup P_8$. These measurements are defined as $M^1_{22} - M^6_{22}$.

Our experiments showed that wrinkling detection can be significantly improved by the use of 3D information especially in the case of nose and cheek wrinkles, since changes in curvature values are usually stronger compared to changes caused to appearance descriptors such as intensity gradient.

4. Facial action unit and facial expression recognition

The set of facial measurements presented in Section 3 is used for detecting a set of 10 action units (AU1, AU2, AU4, AU5, AU7, AU9, AU12, AU15, AU26 and AU27) [5] and recognizing four basic expressions (happy, sad, sur-

AU1	Raises the inner eyebrow corners IF (inc(M_1)>10 AND inc(M_3)>10) OR inc(M_{21})>30 THEN AU1=true
AU9	Wrinkles the nose IF inc(M^1_{22})>15 AND inc(M^3_{22})>20 AND inc(M^4_{22})>20 AND inc(M^5_{22})>20 AND inc(M^6_{22})>15 AND (dec(M_7)>10 OR inc(M_8)>10) THEN AU9=true
AU12	Pulls lip corners upwards obliquely IF inc(M_{11})>5 AND inc(M_{12}) AND M_{12}>5^o AND dec(M_{17})>5 AND inc(M_{16})>8 THEN AU12=true
E1	Disgust IF AU9=true AND dec(M_4)>15 AND dec(M_5)>10 AND inc(M^3_{20})>40 THEN E1=true
E2	Happy IF inc(M_{11})>10 AND inc(M_{13})>10 AND inc(M_{12}) AND M_{12}>5^o AND (inc(M_9)>8 OR inc(M_{16})>8) AND dec(M_{17})>10 AND (inc(M^1_{20})>10 OR inc(M^2_{20})>10) AND inc(M^3_{20})>40 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_{20})>10 OR inc(M^2_{20})>10) AND inc(M^3_{20})>40 AND AU9=false THEN E2=true
E3	Sad IF dec(M_{12}) AND M_{12}<-4^o AND dec(M_{19})>15 AND dec(M_3)>10 AND dec(M_5)>10 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 inc(M_{18})>10 AND dec(M_{17}) AND dec(M_{16})>10 THEN E4=true IF inc(M_1)>15 AND inc(M_2)>15 AND inc(M_5)>15 THEN E4=true

Table 2. Rules for recognizing facial action units and facial expressions. 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. All threshold values have been determined experimentally.

prise, disgust) in 2D+3D image sequences. A rule-based approach is adopted, which is based on direct comparison of the facial measurements extracted from a new image to the measurements obtained from a reference (neutral face) image of the same subject. More specifically, 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 neutral face measurement vector.

To recognize the facial expression 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 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.

These rules have been defined based on [5] and thorough examination of video sequences of facial behavior. 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. For example, happy (smiling) expressions are associated with lip corners being raised obliquely, lower lip getting a ∪-shape, wrinkles appearing on the cheeks and eyelids narrowing. Based on such observations, a set of rules was defined for detecting action units and recognizing facial expressions. Some of these rules are presented in Table 2.

For example, the first rule for *happy* 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}) and the mouth corners distance (M_{13}) increase, 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 significantly. Finally, cheek wrinkling measurements (M_{20}) also increase.

5. Experimental results

A new 2D+3D image database was recorded using the prototype 3D sensor [13]. The database consists of 1040 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 (13 in total) or mimics a facial expression (6 basic expressions + neutral) 2-3 times. Facial action periods last approximately 10s and are proceeded and followed by short neutral state periods. The duration of each recording is about 40-50s and the framerate is 10 fps. Facial action and neutral face periods were manually identified in each of these sequences and an appropriate tag was assigned to each frame. Fig. 4 illustrates examples of recorded image pairs. The resolution is 582×782 pixels, while the depth accuracy is better than 0.3 mm at a mean distance of 60 cm.

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 action unit or facial expression at its peak. The 81 landmarks positions were manually located in each image.

To test the proposed face tracker, we use another set of 600 images, where we manually mark the positions of facial

Figure 4. Examples of facial expression database images. In depth images warmer colours like red correspond to points closer to the sensor. Black pixels correspond to undetermined depth values.

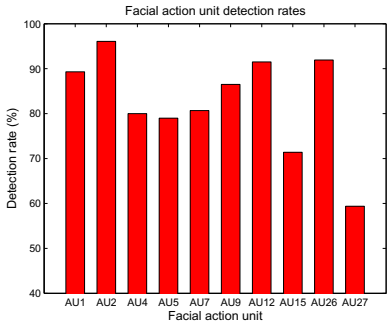

Figure 5. Facial action unit detection rates.

True/Classified	Neutral	Disgust	Happy	Sad	Surprise
Neutral	**94.92**	1.15	2.02	1.35	0.56
Disgust	5.99	**81.47**	6.38	6.03	0.13
Happy	5.86	1.66	**90.43**	0.29	1.76
Sad	23.79	2.47	0.38	**73.36**	0.00
Surprise	2.66	0.78	9.69	0.00	**86.87**

Table 3. Facial expression recognition rates (%).

landmarks. These serve as the ground truth. For each test image, we first fit the global model and then enhance this estimation using the local detectors. The estimated feature positions are finally compared against their ground-truth positions. Using the proposed face tracker, we achieve a mean localization error of 5.35 pixels when the mean face dimensions are 280×370 pixels. On the contrary using the global detector only, the corresponding error is 7.85 pixels.

Next, we evaluate the performance of the facial action unit detector and the facial expression classifier using the recorded image sequences. 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 52×10=520 test sequences, i.e. 52 sequences per action unit (one per subject). The evaluation results are illustrated in Fig. 5. The mean detection rate is 82.5%. The lowest detection rate, 59%, 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. A relative low detection rate is also observed for AU15 (lip corners pressed down). This is mainly 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.

Finally, we evaluate the proposed facial expression recognition technique. 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 52×5=260 test sequences, i.e. 5 sequences per subject (4 expressions + neutral). The evaluation results are presented as a confusion matrix in Table 3. 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 84%. 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 in a very subtle way, only by slightly pressing lip corners down.

To evaluate the benefits obtained from the use of 3D facial data, we compare the proposed facial expression recognition system against a system based exclusively on 2D images. The latter is comprised from 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 sequence frame (2D image), first we localize the position of the 81 facial landmarks using the global 2D ASM and the local feature detectors. In 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 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

	2D+3D	2D
Neutral	94.92	83.60
Disgust	81.47	70.72
Happy	90.43	81.21
Sad	73.36	61.85
Surprise	86.87	79.75

Table 4. Facial expression recognition rates (%) obtained for the proposed 2D+3D method and the 2D appearance-based classifier.

260 sequences used for the evaluation of the proposed face tracker. Table 4 compares its performance against that of the 3D system. It is clear that use of 3D face geometry information significantly aids facial expression recognition.

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

6. Conclusions

A fully automated system for 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 geometric, appearance-based and surface-based measurements, which are effectively classified to emotional related expressions using a rule-based approach. The proposed techniques were evaluated in large database with more than 50 subjects and 1000 sequences demonstrating an average accuracy of 84% and robustness under pose variations.

Future work will exploit the dynamics of facial measurements towards automatic decoding of all action units and their combinations.

References

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and application to human computer interaction. In *Proc. Int Conf. on Computer Vision and Pattern Recognition*, 2003. 1

[2] M. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997. 1

[3] Y. Chang, M. Vieira, M. Turk, and L. Velho. Automatic 3D facial expression analysis in videos. In *2nd Int. Workshop on Analysis and Modelling of Faces and Gestures (AMFG05)*, LNCS 3723, pages 293–307, Oct. 2005. 2

[4] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang. Facial expression recognition from video sequences: Temporal and static modeling. 91:160–187, 2003. 1

[5] P. Ekman and W. V. Friesen. *The Facial Action Coding System: A technique for measurement of facial movement*. Consulting Psychologists Press, Palo Alto, CA, 1978. 1, 5, 6

[6] B. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4:629, April 1987. 3

[7] X. Huang, S. Zhang, Y. Wang, D. Metaxas, and D. Samaras. A hierarchical framework for high resolution facial expression tracking. In *Proc. IEEE Workshop on Articulated and Nonrigid Motion*, page 22, 2004. 2

[8] X. W. Jun Wang, Lijun Yin and Y. Sun. 3D facial expression recognition based on primitive surface feature distribution. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 1399–1406, 2006. 1, 2

[9] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. on Image Processing*, 16(1):172–187, Jan. 2007. 1

[10] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):743–756, July 1997. 2

[11] M. Lewis, J. Jones, and L. Barrett. *Handbook of Emotions*. Guilford Publications, Inc, 2008. 1

[12] S. Malassiotis and M. G. Strintzis. Robust real-time 3D head pose estimation from range data. *Pattern Recognition*, 38(8):1153–1165, Aug. 2005. 3

[13] D. Modrow, C. Laloni, G. Doemens, and G. Rigoll. A novel sensor system for 3D face scanning based on infrared coded light. In *Proc. SPIE Conf. on Three-Dimensional Image Capture and Applications 2008*. 2, 6

[14] I. Mpiperis, S. Malassiotis, and M. G. Strintzis. Bilinear models for 3-D face and facial expression recognition. *IEEE Trans. on Information Forensics and Security*, 3(3):498–511, Sept. 2008. 1

[15] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, Dec. 2000. 1

[16] M. Pantic and L. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Trans. on Systems, Man, and Cybernetics - Part B*, 34(3):1449–1461, May 2004. 1

[17] H. Soyel and H. Demirel. Facial expression recognition using 3D facial feature distances. In *Proc. Int. Conf. on Image Analysis and Recognition*, pages 831–838, Aug. 2007. 1

[18] Y. Sun and L. Yin. Facial expression recognition based on 3D dynamic range model sequences. In *10th European Conference on Computer Vision (ECCV08)*, LNCS 5303, pages 58–71, July 2008. 2

[19] H. Tang and T. Huang. 3D facial expression recognition based on automatically selected features. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshops*, pages 1–8, June 2008. 1