Emotion Recognition based on 2D-3D Facial Feature Extraction from Color Image Sequences

Robert Niese, Ayoub Al-Hamadi, Axel Panning and Bernd Michaelis
Institute for Electronics, Signal Processing and Communications (IESK)
Otto-von-Guericke University Magdeburg
D-39016 Magdeburg, P.O. Box 4210 Germany
Email: {Robert.Niese, Ayoub.Al-Hamadi}@ovgu.de

Abstract—In modern human computer interaction systems, emotion recognition from video is becoming an imperative feature. In this work we propose a new method for automatic recognition of facial expressions related to categories of basic emotions from image data. Our method incorporates a series of image processing, low level 3D computer vision and pattern recognition techniques. For image feature extraction, color and gradient information is used. Further, in terms of 3D processing, camera models are applied along with an initial registration step, in which person specific face models are automatically built from stereo. Based on these face models, geometric feature measures are computed and normalized using photogrammetric techniques. For recognition this normalization leads to minimal mixing between different emotion classes, which are determined with an artificial neural network classifier. Our framework achieves robust and superior classification results, also across a variety of head poses with resulting perspective foreshortening and changing face size. Results are presented for domestic and publicly available databases.

Index Terms—Human Computer Interaction, Computer Vision, Feature Extraction, Emotion Recognition

I. INTRODUCTION

In recent years there has been a growing interest in improving all aspects of human computer interaction (HCI). A challenging aspect of future HCI is to give the computer more human like capabilities, such as emotion recognition. For this purpose, much research has been done in the domain of prosodic speech analysis as well as visual emotion recognition from facial expressions, which is addressed in this article. Generally, facial expression analysis facilitates information about emotions, person perception and it gives insight to interpersonal behavior. It indexes physiological as well as psychological functioning and is essential to use in the scenarios like patient state analysis. Previously human-observer methods of facial expression analysis needed more labor and were difficult to work out across laboratories and over time. These factors force investigators to use generalized systems which are easy to adopt in any environment. To make valid, more accurate, quantitative measurements in diverse applications, it is needed to develop automated methods for recognition. Generally, these systems include automatic feature extractors and change trackers in the facial features in static images and video sequences.

II. RELATED WORK

The analysis of faces has received substantial effort in recent years. Applications include face recognition, facial expression analysis, face simulation, animation, etc. [1, 2]. Observing facial expressions is a natural way of humans to recognize emotions. Extensive studies on human expressions have laid a strong basis for the existence of universal facial expressions. Paul Ekman and his fellows introduced the Facial Action Coding System (FACS) to cover all possible expressions in static images [3].

Many approaches have been described for facial expression analysis using still images and video sequences, which can provide more information regarding expressive behavior. Black and Yacoob presented one of the earliest works using local parametric motion models to regions of the face and feeding the resultant parameters to a classifier [4]. Chang and colleagues used the low dimensional manifold technique for modeling, tracking, and recognizing facial expressions [5]. They used the Lipschitz embedding to embed aligned facial features to build the manifold. Gabor wavelets were exploited for the detection of facial changes occurred during expression. These changes were analyzed in the temporal domain for the expression recognition by Valstar and colleagues [6]. Pose invariant facial expression was analyzed by Kumano and colleagues where they used variable intensity templates. They described that these templates are different for different facial expressions and can be used for the classification [7]. Bartlett et al. [8] presented user independent recognition of facial actions in frontal faces in video streams using an Adaboost classifier and Support Vector Machines. Torre and colleagues used spectral graph techniques for facial shape and appearance features and analyzed them temporally for expressions [9]. Zeng and colleagues treated the emotional facial expression as one class classification problem and separated it from non-emotional facial expressions [10]. Furthermore, they made a comparison between Gaussian based classifiers and a support vector data description technique.

Practically, the common hierarchy in all techniques is that firstly feature extraction is done from facial images or sequences, which is followed by the classification module. The resultant class is one of the predefined expressions.
Usually, in these systems variety comes in the feature extraction technique, which ranges from holistic to individual feature analysis. In the second step, the classifier plays an essential role to discriminate the classes under observation.

State of the art methods for automatic face processing, in particular facial expression analysis are often limited to more or less frontal poses and small skin color variations. Further, normally they are sensitive to global motion with resulting perspective distortions. These restrictions have impact on the applicability for human machine interfaces. In our method, we want to overcome these restrictions by incorporating 2D-3D photogrammetric measuring techniques that provide better pose invariance. Further, feature normalization is carried out, which increases classification robustness.

III. SUGGESTED METHOD

In the following paragraphs, we present a method for automatic emotion recognition from facial expressions in video. The current implementation includes up to seven classes, i.e. six basic emotions and the neutral state. Including new classes is straightforward and basically only requires appropriate training data. Our method is based on the analysis of color image sequences while using 3D context information. In particular, photogrammetric techniques are applied for the determination of features, namely distances and angles, which are corresponding to real world measures. Therefore we apply camera calibration and use subject specific surface model data, which is gained in an initial registration step. In this way, in terms of feature definition, we achieve independence from the current head pose and varying face sizes due to head movements and perspective foreshortening like in the case of out of plane rotations.

In the following paragraphs (see Fig. 1), we introduce the components of our facial expression recognition method, starting with the detection of facial points in the image (A.). Further, a brief explanation of the camera and geometric model is given (B., C.), which are required to determine the face pose and establish transformations between 3D world and image data (D., E.). On that basis a normalized feature vector is built and fed to the classifier (F., G.). Finally, classification results are presented, which were gained in various test scenarios.

Our analysis proves that the proposed representation of facial feature data leads to superior classification of different emotion expressions, while reaching high performance.

A. Facial Feature Points in 2D

In the 2D processing part of the suggested method a number of points are detected across the facial area of the current input image. For initialization and refreshing of the system, a nearly frontal view is required. In between we apply a tracking approach to the feature points with special treatment to the eye region in order to cope with eye blinking.

Analogously to surveyors we define fiducial points, which are exploited to establish correspondence between model and real world data. However, unlike for surveying it is fairly not practical to use markers in the face. Thus, in order to determine a six degrees of freedom pose vector, at least three points need to be found in the face, which firstly, have to be visible also in a range of perspectives and secondly, do not change during expression. Even more, these points need to be well distributed in space and must be robustly and accurately detectable in the image.

Practically, the only points that fulfill all of these requirements are the two points at the center of the eyes. Also in the case of closed eyes, these points can well be inferred from the inner eye corner points. Further, we involve the nose tip point which is determined after pose estimation through model support (see section D.). For reinitialization we also include the mouth corners. This means we have two sets $I_1$ and $I_2$ of fiducial image points in pixel coordinates (1) (Fig. 2e). The corresponding model points we refer as model anchor points (see section C.).

$$I = \{i_{cc}, i_{ee}, i_{n} \}, \quad I = \{i_{cc}, i_{ee}, i_{m}, i_{mn} \}, \quad i \in \mathbb{R}^2$$

The detection of the face and the fiducial points ($i_{cc}, i_{ee}, i_{m}, i_{mn}$) is based on a detector using Haar-like features trained by an Adaboost algorithm as introduced by Viola and Jones [11].

![Figure 1](image.png)  
Figure 1. Work flow of the suggested 2D-3D based method with initialization in B. and C., continuous processing is carried out in the horizontal chain from A. to G.
In particular, after applying the face detector, which provides the face size in the image, the search space for the eyes and mouth is considerably constrained. Then the feature point detector is used alike to find the inner eye corner points and the mouth corner points. Falsely detected point candidates are efficiently removed through a validation step in which the point candidates are compared with estimates provided by a generic face model that is placed inside the bounding rectangle of the face detector. Hereafter, the eye center points \( (i_{\text{le}}, i_{\text{re}}) \) are calculated from the inner eye corners. The centers are determined through extension of the eye corner point axis, enlarged proportionally to the face size in the image. The brow points \( (i_{\text{lb}}, i_{\text{rb}}) \) are detected by following a ray perpendicularly to the eye axis (Fig. 2a). The brow point is defined at the minimum peak of gray level intensity and its derivative.

For mouth extraction we utilize the colored highlighting of the lip tone in contrast to skin color, in particular, in our experiments we found the normalized green color channel \( g \) (2) to work robustly throughout the analyzed data, which covers a range of skin colors. However, in general, the choice of the color space, channel and used threshold may depend on the illumination and individual skin- or mouth-color tone.

\[
g = G / (R + G + B), \quad R, G, B \in \text{RGB color space} \quad (2)
\]

The ROI for the mouth blob is basically defined by the distance of the mouth corners, which have been detected by the point detector (Fig. 2b). Figure 2c displays the distribution in the histogram \( H \) of lip and skin color pixels in the ROI of the normalized green channel. The separation of the histogram’s elements into two sets is achieved by automatically applying threshold \( t_{\text{ng}} \) that is derived from the minimum of a least squares polynomial fitting model \( \xi \) for the histogram. After threshold \( t_{\text{ng}} \) is found the ROI is binarized and one or more blobs of mouth pixels are determined. Morphologic operations, such as closing and contour based erosion [12] are applied to the binary shape. This is especially necessary in the open mouth case as the teeth will not belong to the mouth part of the histogram due to different color characteristics.

Finally, the mouth contour \( m_i \) is determined from the contour of the post-processed mouth blob \( m_b \) (Fig. 2b/d). From this contour, the mouth corner points \( I_{\text{lm}} / I_{\text{rb}} \) plus upper and lower lip points \( I_{\text{ul}} / I_{\text{bl}} \) are selected. The complete set of feature points is summarized in (3) (Fig. 2f).

\[
I_f = \{ i_{\text{ls}}, i_{\text{rs}}, i_{\text{lsb}}, i_{\text{rbs}}, I_{\text{lm}}, I_{\text{rb}}, I_{\text{ul}}, I_{\text{bl}} \}, \ i_f \in \mathbb{R}^2 \quad (3)
\]

### B. Camera Model

In the proposed method a pinhole camera model is applied to simulate a number of fundamental properties of the image capturing device. Camera model \( K \) (4) is the basis for all transformations between the world and image space as well as for the applied 3D model to image fitting technique. The camera parameters are gained in a calibration procedure in which we determine external and internal parameters. Calibration is a well known approach in photogrammetry and surveying [13]. In particular we use a calibration target with coded circles and the standard bundle block adjustment technique.

\[
K = \{ K_x, K_y \} \quad (4)
\]

The six external parameters \( K_x \) represent the camera orientation, i.e. translation and rotation relative to the world coordinate system, which is defined through the calibration target (Fig. 3a).

\[
K_x = \{ t_x, t_y, t_z, \omega_x, \omega_y, \omega_z \}, \quad t_i \in \mathbb{R}^3 \quad (5)
\]

Further, the set of six non-linear internal parameters \( K_i \) (6) describes the geometrical camera properties (Fig. 3b).

\[
K_i = \{ K_\text{proj}, f, \alpha, \beta, \phi, \chi \}, \quad t_i \in \mathbb{R}^3 \quad (6)
\]
\[ K_i = \{ c, s_r, h, a_1, a_2 \}, \]  
\begin{align*}
K_{\text{c}} = \{ c, s_r, h, a_1, a_2 \},
\end{align*}
where \( c \in \mathbb{R} \) is the focal length, \( s_r \in \mathbb{R} \) is the pixel ratio, \( h \in \mathbb{R}^2 \) is the principal point in pixels, and \( a_1, a_2 \in \mathbb{R} \) are coefficients of the radial symmetrical lens distortion.

Facilitating the camera parameters \( K_i \) and \( K_{\text{c}} \), the transformation of 3D world points to image points is well described through projective geometry. An intermediate conversion to the camera coordinate system is required here [14]. In the following, the total transformation process of a world point \( w \) to a sub-pixel image coordinate \( i \) is denoted as \( k (w, K) \).

\[ i = k(w, K), \]  
where \( i \in \mathbb{R}^2 \), \( w \in \mathbb{R}^3 \), Camera model \( K \).

The inverse function \( k^{-1} \) (8) transforms an image point \( i \) to the 3D world point \( w \). Since \( i \) has only two components \( x \) and \( y \), an additional parameter \( d \) is required, which is the depth of the scene. In particular depth \( d \) is the distance along the viewing ray that goes through the image plane at the coordinate \( i \) to the surface of the face. This surface is represented through the geometric model of the observed face.

\[ w = k^{-1}(i, d, K), \]  
where \( w \in \mathbb{R}^3 \), \( i \in \mathbb{R}^2 \), depth \( d \in \mathbb{R} \) and camera model \( K \).

C. Geometric 3D Model

A central part of the proposed method is the evaluation of 3D information gained from the 2D image domain. In particular, the transformation of image points to real world coordinates is the key to pose invariant feature extraction. As can be seen in (8), a distance value that corresponds to the depth of the scene is required. This depth value is retrieved by measuring the distance from the camera plane to the geometric model at the present world pose. For this purpose we introduce the person specific mesh model \( M \), which consists of a number of vertices \( v_i \) (triangulated points) and triangle indices \( w_j \).

\[ M = \{ v_1, \ldots, v_n, w_1, \ldots, w_m \}, \quad v_i \in \mathbb{R}^3, \quad w_j \in \mathbb{N} \]  

In literature, there are various techniques for creating geometric face models, i.e. morphable models [15] or accurate striped lighting methods [16]. However, these approaches have the disadvantages of requiring additional equipment for light projection or high amount of manual interaction. We therefore use an automatic method [17] that exploits point clouds (Fig. 4a/b) and additional color image data, which is taken with a passive stereo camera system and ultimately creates triangle mesh model \( M \) of the face (Fig. 4c). For this initial registration step the observed subject is captured one time in frontal pose and with neutral expression. In the proposed method the face is localized in the stereo point cloud by using the observation that generally, surfaces are represented by more or less connected point clusters, erroneous data on the other hand, frequently leads to geometrically separated points. Further, color information is used, which is retrieved for each 3D point via projection to the corresponding color image. As the face shows a sufficiently homogeneous color that mostly varies from the background, we introduce a similarity criterion \( h \) for clustering that combines color and the Euclidean distance of points. On this basis, a dissection of points into a set of clusters \( C_i \) can easily be achieved.

\[ C_i = \{ p_1, \ldots, p_n \}, \quad p_j \in 3D \]  
All clusters are disjoint, such that

\[ \forall i, j \quad C_i \cap C_j = \emptyset \]  
Any two points \( p_i \) and \( p_j \) are similar and in the same cluster if they fulfill the homogeneity criterion (\( h=1 \)).

\[ h(p_i, p_j, d_{\text{dist}}, d_{\text{col}}) = \begin{cases} 1, \text{if } \| p_i - p_j \| < d_{\text{dist}} \wedge \| p_i^{\text{col}} - p_j^{\text{col}} \| < d_{\text{col}} \\ 0, \text{else} \end{cases} \]  
Threshold \( d_{\text{dist}} \) represents the maximum valid distance between two points within one cluster and is depending on the scaling factor of the point coordinates. Color threshold \( d_{\text{col}} \) is depending on the spectral properties of the camera. For defining a color similarity measure skin color models or simply the RGB color space can be used. The dissection of a point cloud is always unique. Extension of the similarity measure is possible but has not been required yet. Efficient implementation of the clustering can be realized using binary search trees.

Evaluating the clusters, depth information can efficiently be used in order to carry out a segmentation of the head region and eliminate the background. For this purpose, we determine a set of features for each cluster \( C_i \) and apply a quality measure. The selection of features is based on a series of independent observations. First, due to scene setup, the probability is higher that cluster \( C_i \) with center of gravity \( \bar{c}_i \) belongs to the face, if the image projection \( k(\bar{c}_i) \) is close to the image center. For this purpose the normalized distance \( \bar{z}_i \) is applied as a measure of centeredness. It is computed through division by the maximum possible distance \( d_{\text{max}} \), which is inferred from the image dimensions.

\[ \tau_i = \frac{1}{d_{\text{max}}} \| k(\bar{c}_i) \|, \]  
projection \( k \) according (7), \( d_{\text{max}} \) as maximum possible distance and

\[ \tau_i = \begin{bmatrix} w/2 \\ h/2 \end{bmatrix}, \quad w, h - \text{Image width and height}, \]  

and

\[ \bar{z}_i = \frac{1}{n_j} \sum_{j=1}^{n_j} p_j, \quad p_j \in C_i \]
Additionally, we use the fact that with monochromatic illumination the face has only a small color value in the blue component. Thus, under normal imaging conditions, the face has different color characteristics from the background. Here, for every cluster \( C \), the average color \( F_{RGB} \) (16) is determined from all points that are associated.

\[
F_{RGB} = \left( R, G, B \right) = \frac{1}{n} \sum_{j=1}^{n} p_{j} \tag{16}
\]

To gain invariance with respect to illumination changes, the normalized blue color component is applied.

\[
\overline{b} = \frac{B}{R + G + B}, \quad \overline{b}, R, G, B \in [0, 1] \tag{17}
\]

The set of features \( F_i \) (18) is evaluated in a weighted similarity function \( Q_i \) (19).

\[
F_i = \left\{ \overline{z}_i, \overline{b}_i, n_i \right\} \tag{18}
\]

\[
Q_i = \frac{w_1 \overline{z}_i + w_2 \overline{b}_i}{n_i}, \quad i \in [1, q], \text{ weights } w \tag{19}
\]

Equal weights \((w_1 = w_2 = 1)\) have shown to work reliably which is due to stability of features. Empirical tests have shown that a minimum of three independent features is necessary to robustly select face clusters. The minimum value of \( Q \) represents the face cluster \( C_{3D} \), which is chosen for further processing.

\[
C_{3D} = \min(Q), \quad i \in [1, q] \tag{20}
\]

The current shape of the face is roughly approximated by face cluster \( C_{3D} \). Due to the nature of the applied passive stereo image capturing method there are inevitable inaccuracies, such as outliers, dents and holes. Subsequent processing is optimized for the reduction of these disturbances. Our method creates a coarse surface reconstruction of the face cluster, whereas adjacency information of the surface representation is used to significantly suppress errors. Common fitting techniques, such as NURBS approximation are critical at high curvature regions like at the nose tip. Here we use a fast 2D triangulation of point cluster \( C_{3D} \) in camera lens direction. Erroneous vertices are substituted with help of adjacency information, i.e. for each vertex \( v_i \) of the triangulation, mean \( v_i \) is derived by averaging all adjacent vertices \( v_j \). Consequently, all outliers are exchanged that have a deviation bigger than a given threshold. The result is regarded as mesh model \( M \) according to (9) (Fig. 4c). Along with the creation of model \( M \), nose tip \( a_n \) is determined through evaluation of the 3D model shape. In particular, the tip of the nose is the most outward lying point across the center of mesh \( M \).

Accuracy of the reconstruction can be inferred from comparison with a precise active stereo scan of the face (Fig. 4d/e). Here the error within the face is usually clearly below \( \pm 2 \) mm. This is where the feature point detection is carried out. At the border the error is much higher, what is due to foreshortening problems. However, for feature and fiducial point detection this is not relevant. Fiducial image points of set \( I_2 \) (1) are detected in the corresponding image and projected to model \( M \) using (8). From the resulting 3D points we define two sets of model anchor points \( A_1 \) and \( A_2 \), which are applied for the estimation of the model pose (10), (Fig. 5).

\[
A_1 = \{ a_{x1}, a_{x2}, a_{x3} \}, \quad A_2 = \{ a_{x4}, a_{x5}, a_{x6} \}, \quad a_{x} \in R^3 \tag{21}
\]

D. Estimation of Face Pose

Pose estimation is a fundamental task in computer vision. In general, the pose of a model represents a set of parameters to describe the current model orientation. Thus, in 3D often the task is to determine three translation and rotation parameters of a rigid model, what is referred to as pose parameter set \( t_i \) (11).

\[
t_i = \{ t_x, t_y, t_z, t_{rot}, t_{x}, t_{y}, t_{z}, t_{rot} \}, \quad t_{i} \in R \tag{22}
\]

The estimation of the model pose is an optimization problem in which the pose parameters become, usually iteratively, optimized with respect to an error measure. Differences arise in the definition of the error measure, the domain of the corresponding model and image features and the definition of correspondences themselves. Commonly, with 3D scene data from stereo and unknown correspondences between model and world data, ICP algorithms are performed [18].

Figure 4. Geometric surface model, a) point cloud from passive stereo with clustering, b) clustered face, c) reconstructed model \( M \) with detected nose tip \( a_n \), d) accurate reference mesh from active stereo, e) displacement between active and passive stereo shows accuracy in the inner face part.
In the case of 3D model to image fitting, usually edges [19] or prominent points guide the matching process. In a setting where the correspondences between 3D model data and image data, for example on the basis of points is a-priori known, like addressed in this article, pose estimation is simplified to a straightforward fitting process. This however, must take the projective transformation of the image capturing device into account, properly.

Once parameter set \( i \) is known, the corresponding pose matrix \( T \) can be derived, simply by multiplying the basis matrices \( E \) in homogeneous coordinates for the current translations and rotations.

\[
T = E_{io}(x)E_{io}(y)E_{io}(z)E_{io}(u)E_{io}(v)E_{io}(w)E_{io}(x) = (m_{rc}), \quad h \in \{1,4\} \tag{23}
\]

According to camera model \( K (4) \) and current pose transformation matrix \( T \), the image projection of the transformed anchor points \( a_j \) is determined using (7). The goal is now to reduce the error measure \( e \) while in an iterative least squares approach, the projected 3D anchor points are approximated to their corresponding fiducial image points \( i_j \) (1). Correspondence between model and fiducial points is a priori known from the preceding image processing part.

\[
e = \sum_{j=1}^{n} || i_j - k(T \cdot a_j) ||^2 \rightarrow \min, \tag{24}
\]

with \( e \in \mathbb{R}, i_j \in \mathbb{R}^2, a_j \in \mathbb{R}^3, \) model pose matrix \( T \) (23), \( k \) is the world to image projection function (7) and \( n \) is the number of corresponding model and image points, which is four for reinitialization and three otherwise.

Using an iterative differential approach, all six pose parameters become stepwise improved with respect to error measure \( e \), until convergence is achieved. Then matrix \( T \) is used to transform surface model \( M \) to the current world pose. For initialization and re-initialization in longer image sequences, the fitting process is performed between the point sets \( I_1 \) and \( A_2 \), thus eyes and mouth points. After that we determine the world position of the surface model’s 3D nose point \( d_n \), which is then projected to the image point \( i_n \). Hereafter, a tracking of the nose point is applied. For that purpose we track a grid of points on the nose tip \( i_n \) using the LK tracker [20]. This tracking has proved to be stable also during a considerable range of rotation. After a specified number of frames the system is reinitialized. During the tracking we compute the face pose on the basis of point set \( I_1 \) and \( A_1 \) (1), (21), thus eyes and nose point.

E. Generation of 3D Geometric Features

The total set of image feature points \( I_1 \) (3) is transformed to world coordinates (Fig. 6a/b) resulting in the facial feature point set \( P_f \) (25).

\[
P_f \in \{ p_{le}, p_{ra}, p_{lbs}, p_{lbs}, p_{leb}, p_{leb}, p_{le}, p_{le}, p_{lm} \}, p_f \in \mathbb{R}^4, \tag{25}
\]

with \( P_f = k^{-1} (I_1, d_n, K) \) according to (8).

The depth values required in (8) are gained from an intersection test. Here a viewing ray is cast through the virtual camera image plane at pixel coordinate \( i_j \) towards the scene. The depth value is then returned from the distance of the virtual image plane to the surface model, which is oriented with respect to the current pose vector \( t \). This intersection test is realized using a binary space partition tree (BSP) for the surface model. BSP trees are known from computer graphics to handle intersection queries efficiently [21].

Using 3D measures according to (25), issues such as perspective foreshortening and varying face sizes due to back and forth movement are automatically compensated. These are commonly referred to as pose problem. Also it enables the normalization of features as presented in the following.
F. Feature Vector

Fundamentally, the feature vector consists of angles and distances between a series of facial feature points in 3D. As compared to the neutral face, facial geometry shows some specific changes during expression. Thus, the combination of these changes can be used for recognition. The ten dimensional vector \( f(26) \) is directly inferred from point set \( P_f \) (25). The features comprise six Euclidean 3D distances \( d_n \) (27) across the face and four angles \( a_n \) (28), which expose information about the characteristics of the current mouth shape and the overall facial expression state (Fig. 6c). The raising and lowering of both of the eyebrows are gained from the distances \( d_1 \) and \( d_2 \). The distances between the mouth corners and eye centers \( (d_3 \) and \( d_4 \)) capture the mouth movement. The widening and opening of the mouth are represented by \( d_5 \) and \( d_6 \).

\[
f(\ldots d_n a_n \ldots )^T, \quad f \in \mathbb{R}^{10}, \quad a_n \in \mathbb{R}
\]

where

\[
d_1 = \| p_{eyeb} - p_{ee} \|, \quad d_2 = \| p_{eyeb} - p_{ue} \|,
\]

\[
d_3 = \| p_{ro} - p_{re} \|, \quad d_4 = \| p_{ro} - p_{lm} \|,
\]

\[
d_5 = \| p_{ro} - p_{lm} \|, \quad d_6 = \| p_{al} - p_{ll} \|
\]

\[
a_1 = \cos^{-1}\left(\frac{v_1 \cdot v_3}{\| v_1 \| \| v_3 \|}\right), \quad a_2 = \cos^{-1}\left(\frac{v_1 \cdot v_4}{\| v_1 \| \| v_4 \|}\right)
\]

\[
a_3 = \cos^{-1}\left(-\frac{v_1 \cdot v_5}{\| v_1 \| \| v_5 \|}\right), \quad a_4 = \cos^{-1}\left(-\frac{v_2 \cdot v_5}{\| v_2 \| \| v_5 \|}\right)
\]

where

\[
v_1 = p_{ro} - p_{re}, \quad v_2 = p_{ro} - p_{lm}, \quad v_3 = p_{eyeb} - p_{ue}, \quad v_4 = p_{al} - p_{ll}, \quad v_5 \in \mathbb{R}^3
\]

The advantage of \( f_{neutral} \) is determined for the neutral face in an initial registration step. Analysis of the currently observed image frame \( i \) results in feature vector \( f_i \). Further, ratios are computed between the components of \( f_{neutral} \) and \( f_i \) resulting in \( f_{ratio} \) (30). In particular, the operator \( \# \) for component wise division of two feature vectors \( a \) and \( b \) shall be defined as follows.

\[
a \# b = (a_1 / b_1, a_2 / b_2, \ldots, a_{10} / b_{10}) \in \mathbb{R}^{10}, \quad a, b \in \mathbb{R}^{10}
\]

\[
f_{ratio} = f_{ratio} \# f_{neutral}, \quad f_{ratio} = f_{ratio} \# f_{neutral} \in \mathbb{R}^{10}
\]

Analysis has been carried out for numerous subjects and facial expressions. Separately, for all ten components of the feature ratio vector, statistical parameters with respect to mean and standard deviation have been determined. Consequently, the minimum and maximum values \( c_{min} \) and \( c_{max} \) (31) have been computed for each feature distribution across the training data. Applying normalization to the feature ratio vector, the ultimate feature vector \( f_{norm} \) is created (32).

\[
c_{min} = \mu - 2\sigma, \quad c_{max} = \mu + 2\sigma
\]

where \( \mu \in \mathbb{R}^{10} \) and \( \sigma \in \mathbb{R}^{10} \) are vectors for mean and standard deviation across the training data.

\[
f_{norm} = (f_{ratio} - c_{min}) \# (c_{max} - c_{min}), \quad f_{norm} \in \mathbb{R}^{10}
\]

G. Classification

Classification is carried out on the basis of normalized feature vectors. In particular, we are dealing with four facial expression classes related to basic emotions as described by Ekman [3], i.e. Happy, Surprise, Anger, and disgust along with an additional class as the neutral state. For training and classification a so-called Multi-Layer-Perceptron implementation of an artificial neural network is applied [22]. In general, Artificial Neural Networks attempt to model the inner workings of nerve cells by structurally resembling neurons. ANNs apply several layers of information beginning with an input layer (Fig. 7). This is where feature based input data is fed to the model. The output layer consists of nodes that answer the classification query at hand. Also, between the input and output, there can be a number of hidden layers, which represent the associations between the features for each object to be classified. Neural networks are trained with example data by inputting features along with the true classification. The model adjusts the weighting of the nodes in the different layers by using a suitable algorithm, i.e. unsupervised, supervised or reinforcement learning.

Figure 7. Applied ANN structure with 10 inputs and 5 classes.

The advantages of ANNs are their ability to learn concepts from training data and running speed. On the other hand, ANNs are often considered as “black boxes” as they contain undecipherable information, what can cause difficulties for further data analysis. Also, under and overtraining of the classifier can occur and lead to misclassifications. For the classification feature data a net topology is favored that can be learned under supervision, as the matching of learning and target data is known. Thus, a feed forward net topology of a fully connected...
back propagation network with a sigmoid transfer function is used and has proved to produce superior results. In particular we use two hidden layers with a number of six hidden neurons each, a configuration that has been determined through cross validation. The Fast Artificial Neural Network Toolbox [23] has been used for the implementation.

IV. EXPERIMENTAL RESULTS AND CONCLUSION

Training and testing has been carried out in the first step on our domestic database, which contains about 7000 samples (images) for the emotion and non-emotion expressions from twenty persons. Training is done for five classes including the neutral face of speaking subjects who show variations in the mouth region. The emotions of fear and sadness were not yet included due to the lack of proper image data. The classification results of the domestic database are based on about 3500 test samples, i.e. image sequences of 50 frames length in average, each starting with the neutral expression. Also, for testing we are using data from different subjects than in the training phase. Our test scenarios contain pose variations including out-of-plane rotation up to ±25 degrees.

The classification accuracy that we have achieved for our test data can be analyzed by the following confusion matrix (Table 1), which contains information about the actual emotion class \( C_i \) and the prediction \( P(C_j) \) [24]. Here we are dealing with five classes, i.e. Happy (C2), Surprise (C3), Anger (C4), Disgust (C5) as well as the neutral face (C1) of speaking subjects.

<table>
<thead>
<tr>
<th>Class</th>
<th>( P(C_1) )</th>
<th>( P(C_2) )</th>
<th>( P(C_3) )</th>
<th>( P(C_4) )</th>
<th>( P(C_5) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>93.40</td>
<td>0.00</td>
<td>6.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C2</td>
<td>4.94</td>
<td>91.39</td>
<td>0.00</td>
<td>1.27</td>
<td>2.41</td>
</tr>
<tr>
<td>C3</td>
<td>7.63</td>
<td>0.00</td>
<td>92.37</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C4</td>
<td>6.95</td>
<td>0.34</td>
<td>0.00</td>
<td>81.69</td>
<td>11.02</td>
</tr>
<tr>
<td>C5</td>
<td>0.61</td>
<td>1.68</td>
<td>3.82</td>
<td>32.87</td>
<td>61.01</td>
</tr>
</tbody>
</table>

The confusion matrix shows a high percentage recognition accuracy for class C1 to C4, which is represented by diagonal elements, whereas confusion is shown by the non-diagonal elements. Here, only with C5 there is some confusion with class C4. That means, in terms of feature definition the disgust class C5 is the closest one to the anger class while happy has also minor misclassification results with disgust. Neutral, happy and surprise class show the least mixing, which is also according to the physiognomy of these expressions.

From these inspections, it can be seen how clearly the ANN is defining boundaries between different classes. Thus, there is only a small mixing between the classifications of different classes. The small number of misclassification results shows the tendency of each emotion class towards its nearby class.

A typical feature vector and classification graph for a short sequence is illustrated in Fig. 8. First half of the graph shows there is neutral expression in the beginning, which is followed by the emotion expression Happy. The X-Axis reflects the frame number; the Y-Axis shows the raw values of the features in a), and the normalized ratios in b) along with the classification result in c).

Robust working, also with a reduced feature set is desired to guarantee proper working of the system. This is especially of interest in the case of occlusions or self occlusions, e.g. due to strong rotations what may lead to missing features. As the facial expressions are usually symmetric, strong correlations in the extracted features are to be expected. In Table 2 correlations between geometric feature vectors have been computed using the Pearson method [25] as given in (33).

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}
\]

with \( \bar{x}, \bar{y} \) as mean value while \( S_x, S_y \) are the standard deviations for features \( x \) and \( y \), respectively and \( n \) is the number of features.

The correlation matrix shows the relation among each feature pair, whereas all emotion expression classes have been included. It shows that some feature pairs have a very high correlation, e.g. \((d_1/d_2), (d_1/d_3), (a_1/a_2), (a_3/a_4)\). From the definition of these pairs’ features according to (26), it becomes obvious that this effect is due to facial symmetry between the left and right half.

Exploiting this observation, analysis has been carried out for feature reduction in the training data set by carrying out principal component analysis. Setting the variance threshold to 95 percent, we determined five resultant features. Figure 9 visualizes the information contained in the features in terms of mean and variance, before and after PCA reduction. This figure clearly reveals the separateness of the classes used.

The resultant PCA features have also been trained by the artificial neural network and tested. Analysis of the confusion matrix (Table 3) shows that the PCA reduced feature set does only mildly degrade the high classification results. From this observation it can be concluded that missing features due to occlusion do not necessarily have adverse impact on performance of the classification process.

Feature robustness has been analyzed in an additional experiment. Here we compared the feature vector \( f \) against the ground truth feature vector \( f_{\text{neutral}} \) during head motion while having neutral expression. In particular in-plane and out-of-plane face rotations up to ±25 degrees and backward-forward head motion leading to an image variation in face size by factor 1.5 have been regarded. For this setting we found that there is only a small deviation \( d<2\text{mm} \) for the measured distances in the face, which shows the reliability of the 3D processing and feature detection process. Further, the testing data contains moderate variations in white skin color, but no black subjects.
Figure 8. Example sequence “Happy”, a) raw features of 3D geometric measures, b) normalized feature vector over time, c) artificial neural network based classification and d) image processing, detected feature points and mouth contour.

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>1.0</td>
<td>0.981</td>
<td>0.473</td>
<td>0.428</td>
<td>0.440</td>
<td>0.046</td>
<td>-0.347</td>
<td>-0.224</td>
<td>-0.096</td>
<td>-0.213</td>
</tr>
<tr>
<td>d2</td>
<td>1.0</td>
<td>0.469</td>
<td>0.419</td>
<td>0.408</td>
<td>0.031</td>
<td>-0.320</td>
<td>-0.238</td>
<td>-0.114</td>
<td>-0.186</td>
<td></td>
</tr>
<tr>
<td>d3</td>
<td>1.0</td>
<td>0.931</td>
<td>0.279</td>
<td>-0.578</td>
<td>-0.786</td>
<td>-0.726</td>
<td>0.118</td>
<td>0.109</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d4</td>
<td>1.0</td>
<td>0.277</td>
<td>-0.548</td>
<td>-0.762</td>
<td>-0.756</td>
<td>0.137</td>
<td>0.134</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d5</td>
<td>1.0</td>
<td>0.041</td>
<td>-0.321</td>
<td>-0.344</td>
<td>-0.696</td>
<td>-0.739</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d6</td>
<td>1.0</td>
<td>0.499</td>
<td>0.536</td>
<td>-0.011</td>
<td>-0.089</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1</td>
<td>1.0</td>
<td>0.692</td>
<td>0.273</td>
<td>0.020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a2</td>
<td>1.0</td>
<td>0.094</td>
<td>-0.216</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a3</td>
<td>1.0</td>
<td>0.685</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a4</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II.
CORRELATION COEFFICIENT \( r \) AMONG ALL FEATURE PAIRS
JOURNAL OF MULTIMEDIA, VOL. 5, NO. 5, OCTOBER 2010

for evaluation the neural network has been trained with better results when including the additional classes. Thus, which is distance also been regarded. Further, one more feature is included, facial variation, like eye blinking and lip movements has speaking subjects. Nevertheless, a neutral state with small includes data for the classes sadness and fear, but no the four previous emotion categories, this database also from 101 subjects have been analyzed. Additionally to Figure 9.

sequences of 3D models with associated color texture expression Database [26]. Overall 606 facial expression proposed processing. However, we have tested the basic databases cannot readily be used for the whole of the approach of 3D feature processing, we do omit the automatic image feature extraction step, which is highly challenging and error prone due to the very wide range of ethnical groups included in the BU-4DFE database. Thus, we assume to have the image feature points (Fig. 10a) according to (1) and (3) extracted correctly from the input images. These images we have rendered from the 3D model sequences of the database, using a defined virtual camera. The virtual camera enables transformations according to (7) and (8). Figures 10b/c/d show the 3D model of the neutral face and smile expression with projected feature points and indicated distance features. Note that the actual surface has been used in Fig. 10e/d. As this information is generally not available, the proposed method is based on the static face model, which is aligned in the world position. This case is shown in Figure 10e together with the projection of feature points against the static model. As can be seen between Fig. 10d and 10e, the projected feature points have a certain displacement on the surface, caused by the deformation from the facial expression. As shown by the example sequence in Fig. 11a, the displacement causes small differences for the extracted distance and angular features, which are generally below 5mm and 10 degrees, respectively. After feature normalization (Fig. 11b) these differences mostly render into insignificance for classification (Fig. 11c), which has been proven empirically (Table IV, V). As can be seen in Fig. 11c short misclassification can occur at the transition between facial expressions, which is due to indistinct features.

Due to the requirement of calibrated cameras, stereo processing and subject specific surface models different databases cannot readily be used for the whole of the proposed processing. However, we have tested the basic idea of our method with the Binghamton 4D-Facial Expression Database [26]. Overall 606 facial expression sequences of 3D models with associated color texture from 101 subjects have been analyzed. Additionally to the four previous emotion categories, this database also includes data for the classes sadness and fear, but no speaking subjects. Nevertheless, a neutral state with small facial variation, like eye blinking and lip movements has also been regarded. Further, one more feature is included, which is distance \( d_e \) (34). This has shown to provide better results when including the additional classes. Thus, for evaluation the neural network has been trained with 11 inputs and 7 classes for output.

\[
d_e = \| p_{reh} - p_{rob} \|, p \in \mathbb{R}^3
\]  

(34)

As we want to demonstrate the validity of the overall approach of 3D feature processing, we do omit the automatic image feature extraction step, which is highly challenging and error prone due to the very wide range of ethnical groups included in the BU-4DFE database. Thus, we assume to have the image feature points (Fig. 10a) according to (1) and (3) extracted correctly from the input

\[
\begin{array}{c|cccccc}
\text{Class} & P(C_1) & P(C_2) & P(C_3) & P(C_4) & P(C_5) & P(C_6) \\
C_1 & 90.36 & 0.00 & 6.09 & 3.55 & 0.00 & 0.00 \\
C_2 & 3.67 & 89.87 & 0.13 & 0.63 & 5.70 & 0.00 \\
C_3 & 0.79 & 0.00 & 99.21 & 0.00 & 0.00 & 0.00 \\
C_4 & 7.63 & 0.00 & 0.00 & 80.85 & 11.53 & 0.00 \\
C_5 & 0.15 & 5.96 & 0.00 & 41.59 & 52.29 & 0.00 \\
\end{array}
\]

TABLE III. CONFUSION MATRIX IN PERCENT FOR THE REDUCED FEATURE VECTOR

\[
\begin{array}{c|cccccc}
\text{Class} & P(C_1) & P(C_2) & P(C_3) & P(C_4) & P(C_5) & P(C_6) \\
C_1 & 95.66 & 0.00 & 0.07 & 1.07 & 0.36 & 1.35 \\
C_2 & 1.83 & 94.86 & 0.20 & 0.00 & 0.41 & 2.70 \\
C_3 & 0.00 & 0.00 & 87.84 & 0.17 & 0.25 & 8.85 \\
C_4 & 7.54 & 0.00 & 0.00 & 60.89 & 8.08 & 23.09 \\
C_5 & 0.61 & 2.21 & 6.39 & 14.30 & 65.40 & 7.98 \\
C_6 & 0.96 & 12.64 & 11.44 & 0.08 & 14.76 & 21.09 \\
C_7 & 8.37 & 0.00 & 0.08 & 10.86 & 8.62 & 9.87 \\
\end{array}
\]

TABLE IV. CONFUSION MATRIX FOR BU-4DFE DATABASE (STATIC 3D MODEL)

\[
\begin{array}{c|cccccc}
\text{Class} & P(C_1) & P(C_2) & P(C_3) & P(C_4) & P(C_5) & P(C_6) \\
C_1 & 95.59 & 0.07 & 0.00 & 1.14 & 0.71 & 1.21 \\
C_2 & 1.83 & 94.79 & 0.00 & 0.00 & 0.41 & 2.97 \\
C_3 & 0.00 & 0.08 & 86.35 & 0.00 & 0.5 & 9.51 \\
C_4 & 5.08 & 0.15 & 0.00 & 62.05 & 8.55 & 0.92 \\
C_5 & 0.08 & 2.28 & 6.24 & 14.90 & 67.07 & 4.11 \\
C_6 & 1.12 & 13.84 & 11.36 & 3.12 & 15.68 & 45.84 \\
C_7 & 8.13 & 1.74 & 0.41 & 14.01 & 11.44 & 6.22 \\
\end{array}
\]

TABLE V. CONFUSION MATRIX FOR BU-4DFE DATABASE (ACTUAL 3D MODEL)
Figure 10. Analysis of BU-4D Facial Expression database, a) color image rendered with virtual camera, semi-automatic image processing in order to prove concept of 3D processing, b) neutral face model, eight 3D feature points and seven distance features, c) with facial expression, d) surface shading, e) projection of features against static mesh at current pose.

Figure 11. Classification example for BU-4DFE database, a) displacement of features according to (26) in millimeters and degrees, w.r.t. static and actual 3D model, b) normalized feature ratios according to (32), c) neural network classification.
As can be seen from the confusion matrix, the method performs best with the neutral, happy and surprise expression, what can be attributed to the strong feature characteristic of these expressions. The anger class \( C_4 \) shows less accuracy than in the five class setting, as it has considerable confusion with the disgust and sadness classes \( C_3 \) and \( C_7 \). Disgust itself has some mixing with anger and performs comparably with the class sadness \( C_7 \). The lowest recognition rate has been achieved for the fear class \( C_2 \), which has mixings with all classes except neutral and anger. When inspecting the BU-4DFE image material it becomes obvious that the subjects actually show a variety of different facial expressions in the fear samples.

We have further examined our test data with a publicly available real-time system for expression classification provided by Fraunhofer Institute for Integrated Circuits [27]. The high quality of the results that we have achieved with our method could not be reached with their method; in particular, in the course of the sequences, the Fraunhofer classifier frequently gave ambiguous results. This is certainly due to the fact that we have a registration step included. In general, methods, which include subject specific knowledge, achieve higher accuracy than those that do not. Referring to other automatic classification approaches that apply subject specific knowledge, our method has a comparable or smaller classification error [28]. However, adequate comparison can only be carried out on the same image material.

The performance of our system is satisfactory; even though not yet optimized, the processing time is 60 ms per frame on a 3 GHz PC and at a camera resolution of 800x600.

V. SUMMARY AND OUTLOOK

In this article we have presented an efficient framework for emotion recognition from facial images by using a 2D-3D approach. In particular the proposed method is based on an exact camera model and a registration step in which a person specific surface model is built automatically. Incorporating photogrammetric techniques, real world measures are calculated as facial features, such as Euclidean distances and angles. Thus, in terms of feature processing, the method can also cope with various head orientations causing perspective foreshortening and changing face size. Further, through feature normalization and artificial neural network classification, minimal mixing between different classes is reached.

The proposed system offers much room for extensions. In order to achieve broader applicability, instead of the stereo based face model, generic face models are intended to be used in future investigations. Also, additional features can be easily extracted using the 3D framework. Our focus will be to analyze different facial regions with respect to texture. Statistical texture parameters as well as motion analysis will provide additional information for the classification. Also the framework can readily be extended by including other classes in addition to the current ones.

ACKNOWLEDGMENT

This work was supported by BMBF (Bernstein-Group, FKZ: 01GQ0702) and Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

REFERENCES


**BIography**

Robert Niese was born in Halberstadt, Germany in 1977. He received his Masters Degree (Dipl.-Ing.) with distinction in computer science at the Otto-von-Guericke-University Magdeburg, Germany in 2004. He gathered broad experience in several international internship investigations on medical image and data analysis, including MRI, CT and EEG. He is currently working at the Magdeburg University on his PhD thesis that focuses on 3D, image processing, tracking and pattern recognition. Robert Niese is the author of more than 30 publications.

Ayoub K. Al-Hamadi was born in Yemen in 1970. He received his Masters Degree in Electrical Engineering & Information Technology in 1997 and his Ph.D. in Technical Computer Science at the Otto-von-Guericke-University of Magdeburg, Germany in 2001. Since 2002 he has been Junior-Research-Group-Leader at the Institute for Electronics, Signal Processing and Communications at Otto-von-Guericke-University Magdeburg. In 2008 he became Professor of Neuro-Information Technology at Otto-von-Guericke University. His research work concentrates on the field of image processing, tracking analysis, pattern recognition and artificial neural networks. Professor Al-Hamadi is the author of more than 120 articles.

Axel Panning was born in in Magdeburg, Germany, in 1980. He received his Masters Degree (Dipl.-Ing.) in Computer Science at the University of Magdeburg, Germany, in 2006. He is currently working on a PhD thesis focusing on image processing, tracking, and pattern recognition.

Bernd Michaelis was born in Magdeburg, Germany in 1947. He received a Masters Degree in Electronic Engineering from the TH Magdeburg in 1971 and his first Ph.D. in 1974. Between 1974 and 1980 he worked at the TH Magdeburg and was granted a second doctoral degree in 1980. In 1993 he became Professor of Technical Computer Science at the Otto-von-Guericke-University Magdeburg. His research work concentrates on the field of image processing, artificial neural networks, pattern recognition, processor architectures, and microcomputers. Professor Michaelis is the author of more than 250 articles.