British Journal of Applied Science & Technology
16(6): 1-14, 2016, Article no.BJAST.27294
ISSN: 2231-0843, NLM ID: 101664541

SCIENCEDOMAIN *international*
www.sciencedomain.org

Emotional Trace: Mapping of Facial Expression to Valence-arousal Space

Ayoub Al-Hamadi[1], Anwar Saeed[1*], Robert Niese[1], Sebastian Handrich[1] and Heiko Neumann[2]

[1]Institute for Information Technology and Communications, University of Magdeburg, Germany.
[2]Institute for Neural Information Processing, University of Ulm, Germany.

Authors' contributions

This work was carried out in the context of a collaborative research project aiming at the exploration on new companion technology for cognitive technical systems (SFB-TRR 62). Author AAH designed the study, performed the statistical analysis, wrote the protocol, contributed to the experiments and wrote the first draft of the manuscript. Author AS contributed to the analysis of the study, to the literature searches and to the manuscript writing. Author RN participated in the experiments. Author SH contributed to the data gathering. Author HN managed the interpretation of the expressions' intensity. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJAST/2016/27294
Editor(s):
(1) Samir Kumar Bandyopadhyay, Department of Computer Science and Engineering, University of Calcutta, India.
Reviewers:
(1) Gorti Satyanarayana Murty, Jawaharlal Nehru Technological University, Kakinada, A.P, India.
(2) Kamakhaya Kishan Argulewar, RTM Nagpur University, India.
(3) Zahid Akhtar, University of Udine, Italy.
Complete Peer review History: http://sciencedomain.org/review-history/15337

Original Research Article

Received 27th May 2016
Accepted 28th June 2016
Published 9th July 2016

ABSTRACT

The automated analysis of facial expression is a long investigated subject in the computer vision community and has been boosted by applications in the field of human computer interaction (HCI). Besides mapping of facial expressions to basic emotion categories, what is often of limited use for HCI due to sparse occurrence of real emotions, other approaches have been proposed to transform facial expression to the two dimensional so-called valence arousal space. With these affective user state parameters available, the course of the interaction can basically be guided smarter, i.e. the computer can provide help to an apparently confused user. However, it has been shown that the valence arousal space transformation can be impaired due to inaccuracies in image based feature extraction. In this article we present an advanced method using image processing

**Corresponding author: E-mail: asaeed@ovgu.de, anwar.saeed@ovgu.de;*

and 3-D computer vision technology that on the one hand suppresses this problem through hierarchical analysis. Further, our concept enables the assignment of an intensity level of the affective state, which can be a valuable parameter for the interaction. In this paper we give details on the system concept with the different processing steps and respective results. By the application of our method we achieve improvement of facial expression recognition compared to other state-of-the-art methods. In particular we can distinguish roughly 15 percent more classes while maintaining the high recognition rate.

Keywords: Facial expression recognition; human computer interaction; pattern recognition; application.

1. INTRODUCTION

In order to create smart interfaces in state-of-the-art Human Computer Interaction (HCI) applications the analysis and evaluation of facial expression is of increasing importance [1,2] as it may provide information about a person's affective state. Prospectively, this information can be used for the control of the interaction. For instance, through observation of the user's facially and also verbally expressed utterances related to emotion, stress and affect, the automated system can be enabled to support the user offering help. Additionally, a feedback of a satisfaction in wide activities can be measured through the facial analysis (the valence value). For example, we can improve the one-to-one tutoring by adapting it to the student performance through a cognitive process based on nonverbal behavior recognition [3]. A student engagement, which is considered as an important measure for a contemporary education, can be measured from the faces [4]. Nowadays, it is also used to measure the responses to advertisement [5].

Over the last few years, several approaches have been proposed to recognize facial expressions. As each expression is composed of several action units (AUs) simultaneously occurring with different intensities, many researchers build an expression recognizer on top of AUs detection [6], where each AU codes small visible changes in facial muscles [7]. Other approaches employ geometric- and appearance-based features that are directly extracted from the face; clearly they implicitly incorporate the AUs. Littlewort et al. [6] utilize a filter bank of 72 Gabor filters of eight orientations and nine spatial frequencies to extract features used afterwards to detect the AUs using support vector machine classifiers (SVM). On top of the classifiers output, they built a multivariate logistic regression classifier (MLR) to estimate the facial expression. With the help of 68 facial points manually labeled at the first frame and tracked over the image sequence, Lucey et al. [8] infer the facial expressions from the point displacement using SVM. To provide a frame-based decision about the facial expression, Saeed et al. [9] use the relative location of 8 facial points with respect to the enclosing box returned by a face detector. Several approaches assign an expression to the face based on appearance features extracted from the entire face patch, such as local binary patterns (LBP) [10] or local phase quantisers (LPQ) [11]. To recognize the facial AUs and expressions from spatiotemporal features, Valstar et al. [12] utilize the motion history and Zhu et al. [13] the moment invariants. The main shortcoming across the aforementioned approaches is that they extract their features with respect to 2D images, with an assumption of near frontal face pose. This condition cannot be satisfied in real scenarios. To overcome this drawback, our geometric features are extracted with respect to the point location in the real world (3D).

However, in the past and in most state of the art methods discrete emotion categories are being used, which do not always allow an easy interpretation and inference for the interaction. For automated machine analysis some researchers have started implementing a dimensional description of human emotion, especially in combination with audio-visual data where an emotional state is characterized in terms of a set of latent dimensions [14]. Nicolaou et al. propose a multi-layer hybrid framework that derives symmetric spatio-temporal features [15]. These dimensions do mostly refer to the so-called circumplex model of affect, introduced by Russel [16]. In that model two dimensions are considered sufficient for explaining most of the affective variability. In particular these are valence and arousal (V-A), which indicate, how negative or positive and active or inactive an emotional state is. Even though in some works the valence-arousal transformations have already been applied, it has shown that the V-A space mapping can be impaired due to inaccuracies in image based feature extraction [17]. In this article we present a new method that on the one hand suppresses this problem through hierarchical analysis. Further, our new concept enables the assignment of an intensity level of the affective state, which is a valuable parameter. In this paper the major components and principles are explained along with the processing chain and results from the analysis of a 3-D database as well as online examples. The results show the improvement of our method over state-of-the-art techniques by distinction of roughly 15 percent more classes while maintaining the high recognition rate.

2. SUGGESTED METHODS

The major components of the suggested method involve extraction of geometric facial features, assignment of 2-D positions in the Circumplex plane with hierarchical temporal analysis and determination of the intensity level of the affective state. Feature extraction is based on face models, which are particularly used for face pose estimation and feature transformation. Further, camera parameters are being used along with image processing and computer vision techniques. Subsequently the important components of our method are explained briefly and the processing chain is introduced with insight to the applied dynamic temporal analysis. Extensive examples are showing the capabilities of our method.

2.1 Face Model

Throughout this work facial feature processing is based on a geometric 3-D face model, which utilizes the Facegen Photofit routine [18]. This is a morphable model that is adapted to a frontal face image, using facial landmarks. These are robustly found with the IntraFace detector by Xiong et al. [19] in conjunction with gradient data and the active contour model algorithm of Cootes [20] (Fig. 1a/b). In order to attain the correct size of the Facegen based model, we apply scaling in X- and Y-dimension using point cloud data derived from the depth image and the ICP-algorithm of section 2.2 with scaling as free model parameters (Fig. 1c/d).

For further processing we only need a rigid 3-D mesh description of the adapted Facegen model, which is denoted by **M** (2.1).

$$\mathbf{M} = \left(\left\{\mathbf{v}_1,\ldots,\mathbf{v}_n\right\},\left\{w_1,\ldots,w_m\right\}\right), \mathbf{v}_i \in \mathbf{R}^3, w \in \mathbf{N} \quad (2.1)$$

with $\mathbf{v}_i$ as mesh vertices and $\mathbf{w}_j$ as triangle indices.

2.2 Facial Expression Features

In the presented work we use geometric facial features to characterize the current expression at every image frame *t*. In particular we evaluate spatial distances and angles that are computed from a set of characteristic 3-D feature points.

2.2.1 Facial expression related feature points

In automated facial expression recognition, the evaluation of feature points is a common approach. Accordingly, in our work the choice of feature points has been motivated by the so-called Facial Animation Parameter (FAP) system [21], which has been created for animation purposes in the context of the MPEG-4 standard. In the FAP 88 feature points are used for the controlled definition of facial expression. In our investigations we have found that a subset of 8 key points (Fig. 2b) already performs excellent and robust for the recognition task. Specifically, we apply point set $\mathbf{P}_f$ (2.2). The computation of these feature points is done model based and requires the detection of corresponding image points beforehand.

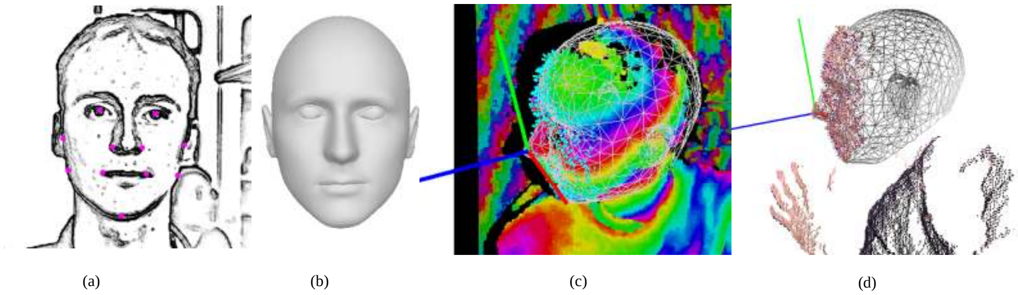

Fig. 1. Face model adaptation, (a) Frontal face gradient image with landmark points detected using IntraFace and active contours, (b) Reconstructed 3-D model, (c) Depth image encoded in cyclic rainbow colors with scaled model at pose, (d) Point cloud with scaled model

$$\mathbf{P}_f = \{\mathrm{p}_{le}, \mathrm{p}_{re}, \mathrm{p}_{leb}, \mathrm{p}_{reb}, \mathrm{p}_{lm}, \mathrm{p}_{rm}, \mathrm{p}_{ul}, \mathrm{p}_{ll}\},\ \mathrm{p}_i \in \mathbb{R}^3 \quad (2.2)$$

2.2.2 Extraction of image feature points

For the determination of feature points, i.e. eyes, eyebrows and mouth, in the first step, a Haar-like feature (HLF) based Adaboost classifier [22] is utilized to find the subject's face and to confine the search region in the image. Using the IntraFace detector [19] the complete set of image feature points $\mathbf{I}_f$ (2.3) is reliably found.

$$\mathbf{I}_f = \{\mathrm{i}_{le}, \mathrm{i}_{re}, \mathrm{i}_{leb}, \mathrm{i}_{reb}, \mathrm{i}_{lm}, \mathrm{i}_{rm}, \mathrm{i}_{ul}, \mathrm{i}_{ll},\},\ \mathrm{i}_i \in \mathbb{R}^2 \quad (2.3)$$

2.2.3 Definition of geometric features

Basically, the evaluation of 3-D features has the advantage of invariance from the current pose, opposed to plain 2-D image features, which suffer from perspective distortions. The determination of geometric features derived from the 3-D feature point set $\mathbf{P}_f$ (2.2) makes use of this attribute. These raw features are combined to vector **f** (2.4), which is the starting point for the normalization and successive classification.

Generally, the appearance of faces reveals specific changes during facial expression, especially when compared to the neutral expression. In our approach the neutral facial expression $\mathbf{f}_{neutral}$ is captured once per subject during initialization. The geometric features include seven Euclidean 3-D distances d_i within the face and four angular parameters α_k in the mouth region (Fig. 2b, c).

$$\mathbf{f} = (d_1\ \dots\ d_7\ \alpha_1\ \dots\ \alpha_4)^{\mathrm{T}},\ d_i, \alpha_k \in \mathbb{R},\ \mathbf{f} \in \mathbb{R}^{11} \quad (2.4)$$

Here the distances d_i are defined as

$$\begin{aligned} &d_1 = \|\mathbf{p}_{reb} - \mathbf{p}_{re}\|,\ d_2 = \|\mathbf{p}_{leb} - \mathbf{p}_{le}\|,\\ &d_3 = \|\mathbf{p}_{re} - \mathbf{p}_{rm}\|,\ d_4 = \|\mathbf{p}_{le} - \mathbf{p}_{lm}\|,\\ &d_5 = \|\mathbf{p}_{rm} - \mathbf{p}_{lm}\|,\ d_6 = \|\mathbf{p}_{ul} - \mathbf{p}_{ll}\|, \end{aligned} \quad (2.5)$$

and angles α_k as follows

$$\begin{aligned} &\alpha_1 = \arccos\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \cdot \|\mathbf{v}_2\|}\right),\ \alpha_2 = \arccos\left(\frac{\mathbf{v}_2 \cdot \mathbf{v}_3}{\|\mathbf{v}_2\| \cdot \|\mathbf{v}_3\|}\right),\\ &\alpha_3 = \arccos\left(\frac{-\mathbf{v}_2 \cdot \mathbf{v}_4}{\|\mathbf{v}_2\| \cdot \|\mathbf{v}_4\|}\right),\ \alpha_4 = \arccos\left(\frac{-\mathbf{v}_2 \cdot \mathbf{v}_5}{\|\mathbf{v}_2\| \cdot \|\mathbf{v}_5\|}\right), \end{aligned} \quad (2.6)$$

With

$$\mathbf{v}_1 = \mathbf{p}_{rm} - \mathbf{p}_{ul},\ \mathbf{v}_2 = \mathbf{p}_{ll} - \mathbf{p}_{ul},\ \mathbf{v}_3 = \mathbf{p}_{lm} - \mathbf{p}_{ul},$$
$$\mathbf{v}_4 = \mathbf{p}_{rm} - \mathbf{p}_{ll},\ \mathbf{v}_5 = \mathbf{p}_{lm} - \mathbf{p}_{ll},\ \mathbf{v}_i, \mathbf{p}_j \in \mathbb{R}^3.$$

The determination of the distances plus angles requires the computation of the 3-D facial points $\mathbf{P}_f$ and the present head orientation beforehand, what is explained in the following system concept section, where the complete processing chain is introduced.

2.3 System Concept Processing Chain

The processing chain of our system concept for the automated image based facial expression analysis contains five parts that are explained in the subsequent points.

- Capturing of color image and depth data,
- Point cloud based pose estimation,
- Feature normalization,
- Valence-arousal estimation,
- Dynamic integration and determination of intensity.

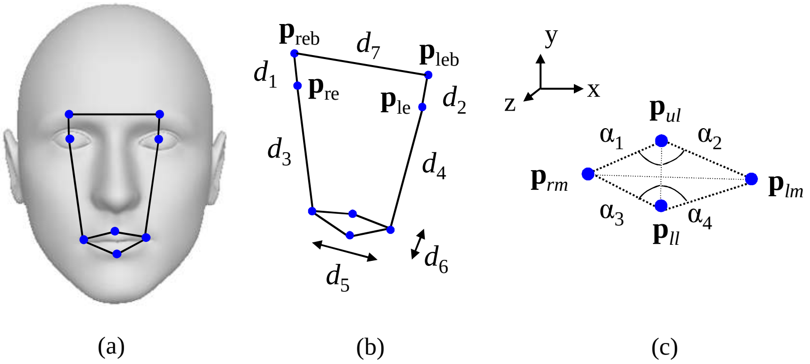

Fig. 2. (a) Face model M with features, (b) Points $\mathbf{p}_j$ and, distance features d_i, and (c) Angles α_k along the 3-D model's mouth region

2.3.1 Capturing of color image and depth data

In the presented approach we use the Kinect camera system for online experiments, which provides color images along with depth data, respectively 3-D world points for each pixel at 30Hz in VGA resolution. The depth image information enables fast and accurate face pose estimation, which itself is a prerequisite for correct facial feature extraction. Facilitating the available intrinsic camera parameters **K**, the transformation of 3-D world points in camera system coordinates to image points can be described using so-called projective geometry [23]. For this purpose, in the following the projection of any world point **w** in camera coordinates to the image coordinate **i** is referred to as function *k*.

$$\mathbf{i} = k(\mathbf{w},\mathbf{K}) \qquad (2.7)$$

with $\mathbf{i} \in \mathbf{R}^2, \mathbf{w} \in \mathbf{R}^3$ and camera model **K**.

The inverse function k^{-1} (2.8) creates a 3-D world point **w** from an image point **i**. Since this transformation requires additional information for the third dimension. Here we use the scene depth *d*, which is the distance on the viewing ray from the camera's image plane at coordinate **i** down to the facial surface. This is realized with face model **M** (2.1), which aligned to the current pose. The intersection is determined using ray casting [24].

$$\mathbf{w} = k^{-1}(\mathbf{i},d,\mathbf{K}) \qquad (2.8)$$

with $\mathbf{i} \in \mathbf{R}^2, \mathbf{w} \in \mathbf{R}^3$, depth *d* and camera model **K**.

2.3.2 ICP-based pose estimation using point cloud data

Commonly, the works on pose estimation are based on the determination of rigid body motion with six degrees of freedom [25], i.e. translation plus rotation, what is subsequently referred to as pose vector *t* (2.9).

$$\mathbf{t} = \left(t_x\, t_y\, t_z\, t_\omega t_\varphi\, t_\kappa\right)^T, \mathbf{t} \in \mathbb{R}^6 \qquad (2.9)$$

We realize the pose estimation with help of an Iterative Closest Point (ICP) algorithm variant, in which we approximate a person adapted ICP model to point cloud **W** gained from the processed depth map, while introducing a normal vector constraint (Fig. 3). The approach is explained in detail in [26]. The goal of the ICP algorithm is to iteratively reduce error measure *e*(**t**) while improving pose vector **t** (2.9). Besides translation and rotation angles, also model scaling can be estimated.

$$e(\mathbf{t}) = \sum_j \left(d_j(\mathbf{t})\right)^2 \rightarrow min,\ d_j(\mathbf{t}) = \left(\mathbf{a}_j(\mathbf{t}) - \mathbf{p}_i\right) \cdot \mathbf{b}_j \qquad (2.1)$$

with $\mathbf{t} \in \mathbb{R}^6, \mathbf{a}_j, \mathbf{b}_j, \mathbf{p}_i \in \mathbb{R}^3, d_j \in \mathbb{R}$

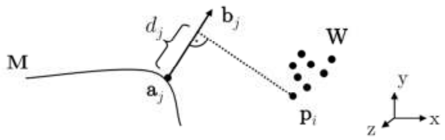

Fig. 3. Pose estimation minimization principle. Orthogonal model-point cloud distance d_j constraint. Point cloud W, face mesh model vertex $\mathbf{a}_j$, normal $\mathbf{b}_j$

2.3.3 Feature normalization

Based on the geometric model at the current pose, every image feature point in $\mathbf{I}_f$ (2.3) is transformed to a 3-D coordinate through application of function k^{-1} (2.8), yielding in point set $\mathbf{P}_f$ (2.2). On that basis feature vectors can be computed according to (2.4). When inspecting the facial features, large deviations become obvious between different persons and different facial expressions. Thus, in order to enable classification, additional steps are required, i.e. building ratios and normalization. Here, the facial feature vector $\mathbf{f}_{neutral}$ is computed for the face with neutral expression in the initial registration step. Examination of the current image frame *t* leads to feature vector **f**(*t*). Next, we compute the ratios between all components of $\mathbf{f}_{neutral}$ and **f**(*t*) what leads to $\mathbf{f}_{ratio}(t)$ (2.12). For this task we introduce the operator # (2.11) to retrieve the component wise ratio between two feature vectors.

$$\mathbf{a}\#\mathbf{b} = (\mathbf{a}_1 / \mathbf{b}_1\ \mathbf{a}_2 / \mathbf{b}_2\ ...\ \mathbf{a}_{11} / \mathbf{b}_{11}) \in \mathbb{R}^{11} \qquad (2.11)$$

$$\mathbf{f}_{ratio}(t) = \mathbf{f}(t)\#\mathbf{f}_{\text{neutral}},\ \mathbf{f}_{ratio}(t), \mathbf{f}(t), \mathbf{f}_{neutral} \in \mathbb{R}^{11} \qquad (2.12)$$

For every element of vector $\mathbf{f}_{ratio}$, statistical parameters were evaluated in an examination of many persons and facial expressions. In addition to vectors for mean *μ* and standard deviation *σ* these are the minimum and maximum values c_{min} and c_{max} (2.13), which have been computed for all feature distributions in the training data set.

$$\begin{aligned}\mathbf{c}_{min} &= \mu - 2\sigma,\ \mathbf{c}_{min} \in \mathbb{R}^{11} \\ \mathbf{c}_{max} &= \mu + 2\sigma,\ \mathbf{c}_{max} \in \mathbb{R}^{11}\end{aligned} \tag{2.13}$$

Then, the resulting feature vector $\mathbf{f}_{geo}(t)$ (2.14) is achieved through normalization of the ratio vector $\mathbf{f}_{ratio}$. Subsequently, the facial feature vector is used for valence and arousal estimation.

$$\begin{aligned}\mathbf{f}_{geo}(t) &= (\mathbf{f}_{ratio}(t) - \mathbf{c}_{min}) \# (\mathbf{c}_{max} - \mathbf{c}_{min}) \\ &= (\mathbf{f}_{ratio}(t) - \mathbf{c}_{min}) \# 4\sigma,\ \mathbf{f}_{geo}(t) \in \mathbb{R}^{11}\end{aligned} \tag{2.14}$$

2.3.4 Valence-arousal estimation

In literature most facial expression analysis approaches apply discrete emotion categories for classification which is not always optimal. The model we apply is influenced by the observation that the model's labels valence and arousal lead to robust emotion state representation that is continuous in principle. Thus, unlike conventional methods we use the mapping $f_{map}(\mathbf{f}_{geo}(t))$ (2.16) of the 11-dimensional feature vector $\mathbf{f}geo$ to the 2-D plane of the Circumplex model of affect (Fig. 4).

In particular, we use a technical realization of the famous model from psychology. In our implementation the Circumplex model plane is defined by six different positions in polar coordinates P_C (2.15) of discrete emotion categories plus neutral (Fig. 5). This definition represents the results of Russel [16].

$$P_{C_i}\left(r_{c_i};\varphi_{c_i}\right) \in \begin{Bmatrix} 0 & l & l & l & l & l & l \\ 0 & 10 & 85 & 170 & 200 & 125 & 240 \end{Bmatrix},\ l = 0.7,\ \varphi_{c_i} \tag{2.15}$$

with classes $C_i \in$ {Neutral, Happy, Surprise, Anger, Disgust, Fear, Sad}.The radius was empirically set to l=0.7.

For the transformation $f_{map}(\mathbf{f}_{geo}(t))$ (2.16) of feature vectors at frame t an artificial neural network is used, i.e. a so-called Multi-Layer Perceptron (MLP) [28], parameterized with a sigmoid transfer function and a training algorithm utilizing backpropagation.

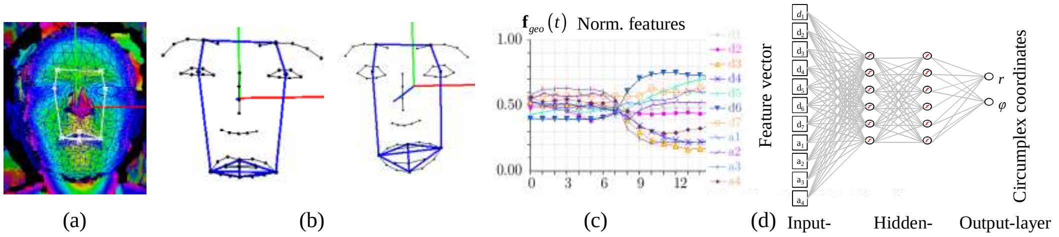

Fig. 4. Valence-Arousal space transformation; (a) Depth image with inferred pose and overlaid point detection, (b) 3-D features in blue, (c) Feature plot, (d) Artificial Neural Network used for transformation $f_{map}(\mathbf{f}_{geo}(t))$ (2.16) of the 11-dimensional feature vector to the valence arousal space position

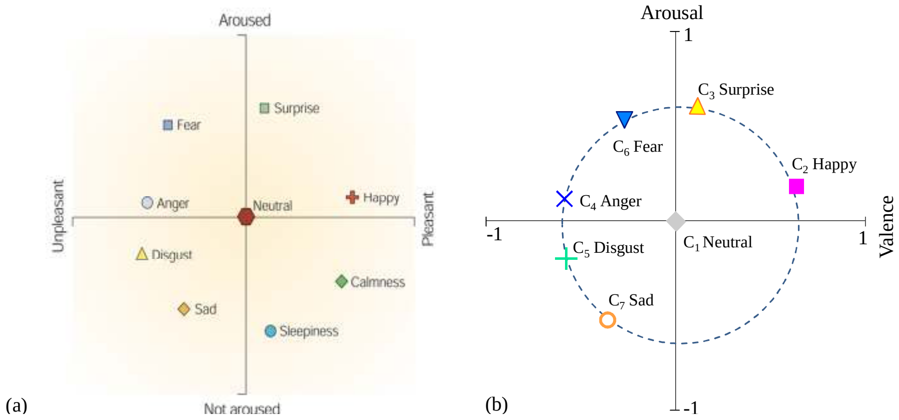

Fig. 5. (a) Circumplex model of affect wrt. Russel (image source [27]), (b) Technical implementation of the Circumplex model with applied polar coordinates (2.15) for each class that define the 2-D model plane

$$f_{map}:\mathbf{f}_{geo}(t)\in\mathbb{R}^{11}\xrightarrow{MLP}\begin{bmatrix}V\\A\end{bmatrix}\in\mathbb{R}^{2} \tag{2.16}$$

with valence *V* and arousal *A*.

The applied network has eleven input and two output neurons with two hidden layers that contain six neurons each. The working assumption is that, based on the training data, we can deduce the 11-D to 2-D transformation in terms of the neural network's adapted weights. Throughout the supervised training all samples have been assigned to the polar coordinate of their respective class in the circumplex plane. Accordingly, in the classification step, each input vector leads to a position in the output plane, which is supposed to be at a place that reflects the presented emotion.

2.3.5 Dynamic integration, determination of intensity

The determination of the current V-A state *z*(*t*) can be regarded as an inverse problem. That means, in the first step, the state results from feature processing and mapping to the 2-D valence-arousal plane, expressed as $f_{map}(\mathbf{f}_{geo}(t))$. Next, the task is finding the unknown state, which is expressed by variable *z*(*t*). From a mathematical point of view, in Hadamard's sense this problem is ill-posed because the reconstruction is likely to be sensitive to noise and ambiguous. The solution is expected to be in range of the observations, as measured by the square norm of (2.17).

$$E_{data}=\left\|z(t)-f_{map}\left(\mathbf{f}_{geo}(t)\right)\right\|^{2} \tag{2.17}$$

In addition, we constrain the potential solution by utilizing a constraint operator *P*(*z*) to simultaneously impose a smoothness property upon the solution. The smoothness property is realized through the 1st order derivative of the desired solution, i.e. $P(z)=\dot{z}$,[1] which is scaled by a weighting constant λ, also called regularization parameter. In combination, the resulting energy measure is defined as the weighted sum of the data and the smoothness term defined above as

$$E_{data}(z)=\int\left\|z(t)-f_{map}\left(\mathbf{f}_{geo}(t)\right)\right\|^{2}+\lambda\cdot\dot{z}^{2}(t)dt\rightarrow\min \tag{2.18}$$

For the minimization of (2.18) we apply the Euler-Lagrange equation to solve the partial differential system of equations, what leads to state variable *z*(*t*). The intensity level *r* (2.19) of the current emotion quantity is inferred from the state variable, while the user state is traced over a temporal period and integrated over time.

$$z(t)=\begin{pmatrix}r\\\beta\end{pmatrix}(t) \tag{2.19}$$

with $r(t)=\sqrt{a^{2}+v^{2}}$, $\beta(t)=\tan^{-1}(v/a)$, where *a* and *v* are the scalar activations along the cardinal dimensions of arousal and valence, respectively.

3. RESULTS AND DISCUSSION

The different modules of the proposed method have been processed with various training and test samples, taken from the BU-4DFE database [29] and exemplary online recordings using the Kinect camera. In particular, from the BU database about 9.000 samples from seven classes (neutral, happy, surprise, anger, disgust, fear, sad) of the database have been used to train the neural network. For testing another 9.000 samples were used. Each sample consists of a high quality texture image and a 3-D depth map.

Our point cloud based pose estimation method has been adapted to process the temporal 3-D samples of the BU database. In order to enable the required 2-D/3-D transformations, video and 3-D depth data have been generated using a virtual camera with defined parameters (Fig. 6). For this purpose the database raw data has been rendered in OpenGL as textured mesh, whereas the color image and depth buffers serve as input for our method. Analysis has been carried out by applying feature extraction and valence-arousal transformation to the preprocessed BU-4DFE data (Fig. 7).

Even though the motivation of this work is to over-come thinking in fixed categories for recognition purposes; for the evaluation of the method, the basic emotion categories are suitable for testing the V-A space transformation of the samples and comparison with the state-of-the-art recognition. In order to gain a qualitative statement about the recognition results, we analyzed the angle *μ*(*t*) (3.1), which reflects the displacement between the computed angle *β*(*t*) (2.19) and the a-priori given sample's class orientation φ_{Ci} (2.15) in the V-A space of the

[1] *We use the dot notation to refer to temporal derivatives of the function with time as the independent variable.*

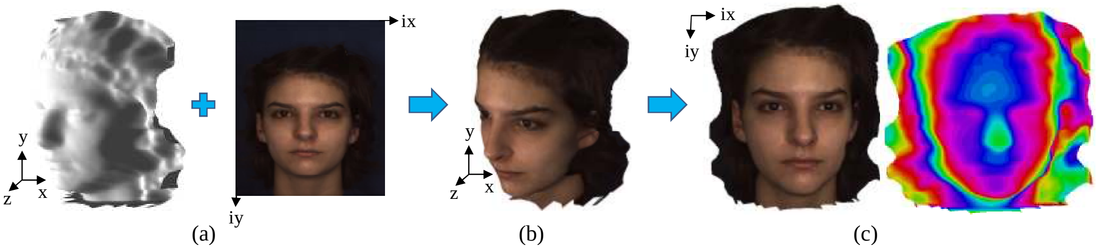

Fig. 6. Preprocessing of BU-4DFE data. (a) 3-D mesh and high resolution texture image, (b) Textured mesh in 3-D view, (c) Rendered color and cyclic rainbow depth image with defined camera parameters

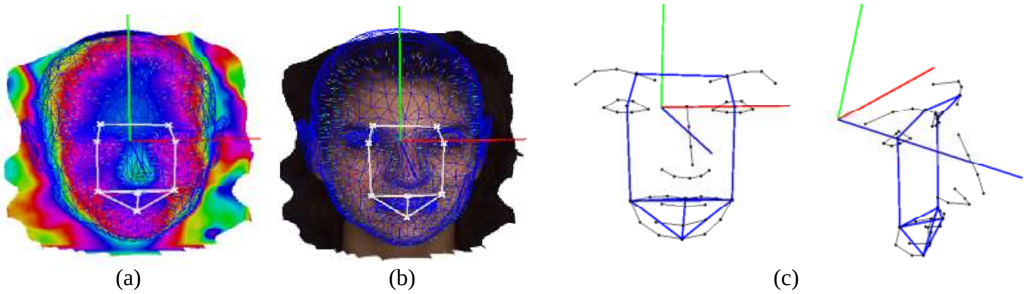

Fig. 7. Feature processing of a BU-4DFE sample. (a) Depth and (b) Color image with pose as RGB coordinate system, further face model M (1) is shown as blue triangle mesh and the extracted facial features as white lines. (c) Shows the projected 3-D facial features as blue lines in two different views along with the face pose, the black lines represent the image processing result after 3-D transformation

Circumplex model (Fig. 8a). Also, in this way we can easily compare the achieved recognition results over other state-of-the-art approaches that evaluate that particular database.

$$\mu(t)=\left|\beta(t)-\varphi_{Ci}\right|\in\mathbb{R}\text{, see (2.15), (2.19)} \qquad (3.1)$$

The neutral class is handled in a different way, i.e. in the recognition step a sample is rated as neutral, if $r(t)<min_r$, according to (2.19). The threshold has been set empirically to min_r=0.25. The ground truth data for the neutral facial expression class C_1 were extracted from the first frames of all database sample sequences, where there is neutral facial expression.

In the evaluation we have randomly separated the database into training and test samples, such that all classes are equally represented (about 1.400 samples per class) and no training sample is contained in the test data. For testing we consider a sample to be correctly classified, if the angle μ is below threshold t_μ. If not then it is considered to be belonging to the closest adjacent class in the Circumplex model plane in terms of its angular value. The following Tables 1, 2 show the resulting confusion matrices for two empirical values of threshold t_μ, i.e. t_μ=30 and 60 degrees.

The tables further show that the recognition results are highest for the classes with the greatest feature distinction, thus, surprise and happy, while confusion rather occurs for the other classes. The average recognition rates are 70.2 and 79.7 percent for t_μ =30 and 60 respectively. This is better, or at least in accordance with solely category based state-of-the-art recognition techniques, especially with that particular database [30,31]. However, apart from the other authors, we also consider the neutral class, which is commonly neglected in literature. This shows an obvious advantage of our method which also deals with this challenging class. Hence, in numbers, our method deals with roughly 15 percent more classes, while maintaining the recognition rate.

Further, thorough inspection of the image material shows that sometimes, the presented facial expressions are not easy to recognize, even for humans. Here the continuous description of the user's emotional state in the V-A space offers more opportunities for further evaluation and action, compared to previous category based classification.

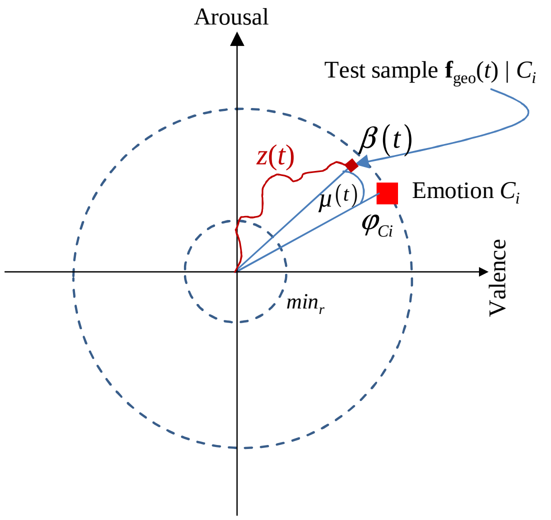

Fig. 8. The method's accuracy for a sample f$_{geo}$(*t*) is determined using angle *μ*(*t*), which reflects the displacement of function *z*(*t*) and the a-priori given sample's class *C$_i$* angle *φ$_{Ci}$* in V-A space

Fig. 9 shows the mapping of the BU-4DFE testing samples, which were not used for model training. Also here one can see overlapping of samples from the classes disgust and anger, as well as sad and surprise and fear. Further empirical tests have been carried out with samples that were taken with the Kinect camera (Fig. 10) with a focus on the evaluation of emotion intensity and the temporal constraint. The processing speed is about 20 frames per second at VGA resolution. Example sequences with presented basic emotions are given in Fig. 11, i.e. the 3-D facial expression model is shown with extracted features, and the projections to the 2-D valence arousal space over time. The graphs show the evolution of function *z*(*t*) (2.19), starting from the center, which represents the neutral state. The samples show that the various facial expressions are clearly separated, what makes it possible to perform an assessment with respect to the emotion model's parameters valence and arousal.

An example of the smoothing effect of the dynamic temporal constraint can be seen in Fig. 12 (corresponds to first plot of Fig. 11a), as well as the respective temporal feature sequence, together with a conventional basic emotion classification and a subset of the high dimensional feature space. In that example the categorized basic emotion is just another perspective of the V-A transformation result, due to the facial expression of a smile.

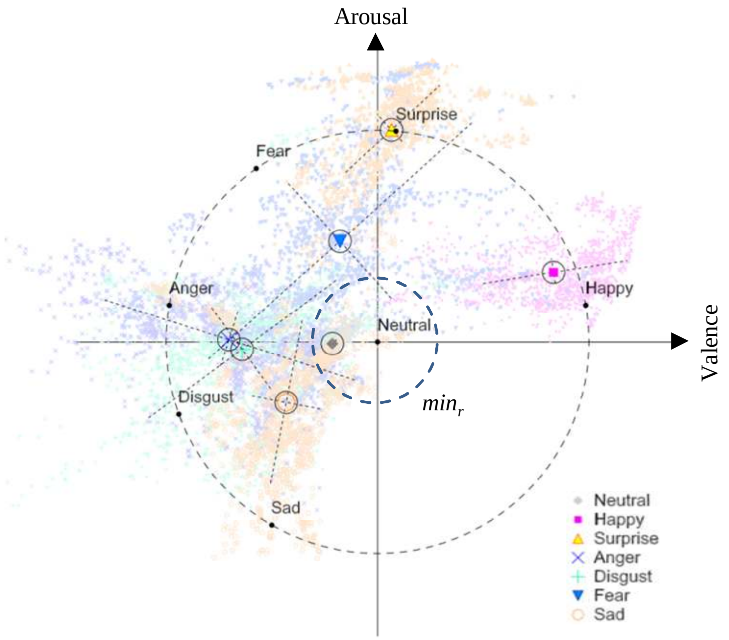

Fig. 9. Projection of test set from BU-4DFE database in light color with centroid and principal axes for each class. Samples with *r*(*t*)<*min$_r$* are considered neutral according to (2.19) due to low expression intensity

Table 1. Confusion matrix in percent, t_μ=30

Class	P(C_1)	P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_1 Neutral	**89.1**	0	0.1	0.4	10	0.1	0.3
C_2 Happy	3.1	**82.2**	13.3	0	0	1.4	0
C_3 Surprise	1.4	0	**93.1**	0.1	0.1	5.3	0
C_4 Anger	0.8	0.5	0	**64.5**	31.6	1.2	1.4
C_5 Disgust	6.2	1.6	0.2	37.8	**50.4**	3.5	0.3
C_6 Fear	4.7	5	14.4	7.3	5.9	**62.6**	0.1
C_7 Sad	8.9	0	1.8	6.5	29.5	0.1	**53.2**

Table 2. Confusion matrix in percent, t_μ=60

Class	P(C_1)	P(C_2)	P(C_3)	P(C_4)	P(C_5)	P(C_6)	P(C_7)
C_1 Neutral	**89.1**	0	0.1	0.4	10	0.1	0.3
C_2 Happy	3.0	**89**	6.6	0	0	1.4	0
C_3 Surprise	1.4	0	**98.4**	0.1	0.1	0	0
C_4 Anger	0.8	0.5	0	**65.7**	31.6	0	1.4
C_5 Disgust	6.2	1.6	0.2	37.8	**50.7**	3.5	0
C_6 Fear	4.7	5.0	1.4	0	5.9	**83**	0
C_7 Sad	8.9	0	1.8	6.5	0	0.1	**82.8**

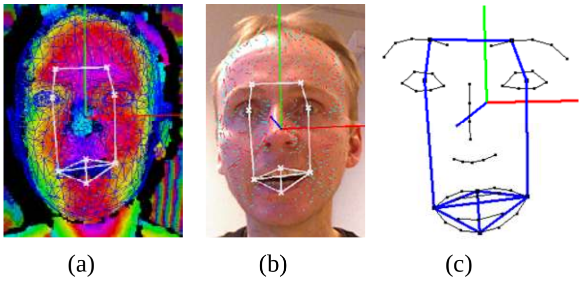

Fig. 10. Evaluation example. Online feature processing with (a) Depth and (b) Color image, (c) Extracted features

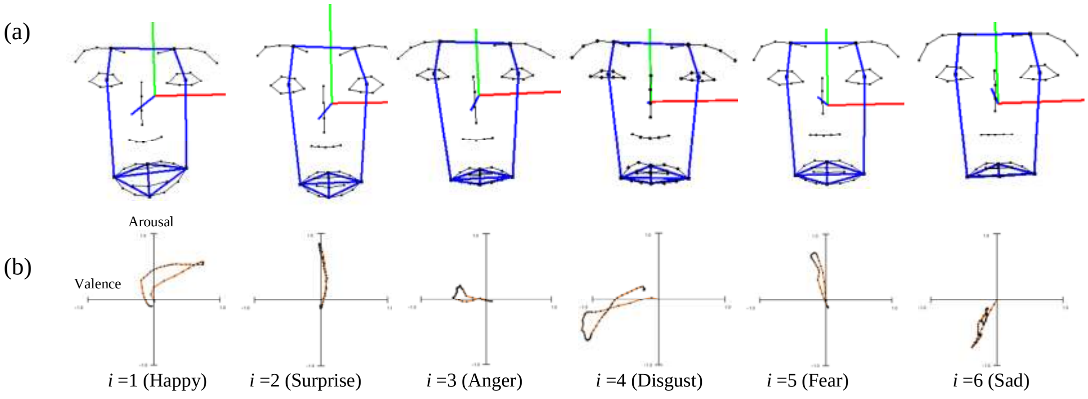

Fig. 11. (a) 3-D Facial expression model with features in blue. (b) Shows the projections of the presented classic basic emotion expressions to the V-A space over time using the temporal constraint

A further exemplary sequence with different expressiveness of a presented emotion is shown in the plots of Fig. 14, along with the corresponding category based classification. This example shows clearly the benefit of the emotion mapping, which provides the intensity information, that cannot be inferred from the category based recognition. Also here the smoothing effect of the dynamic temporal constraint becomes obvious.

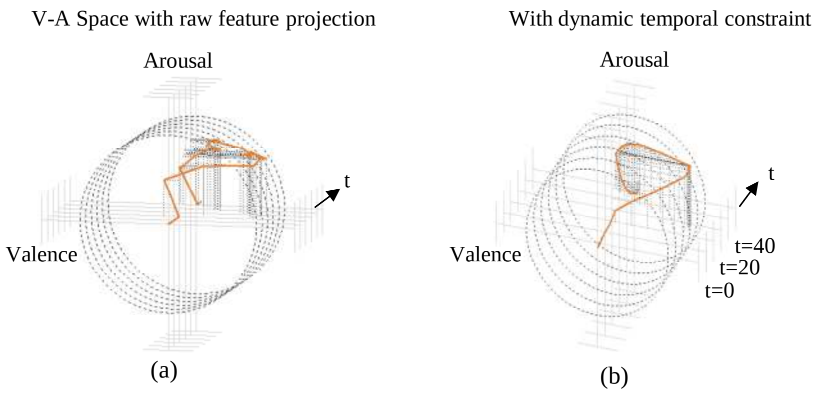

Fig. 12. Valence-Arousal space mapping example. (a) Without and (b) With dynamic temporal constraint

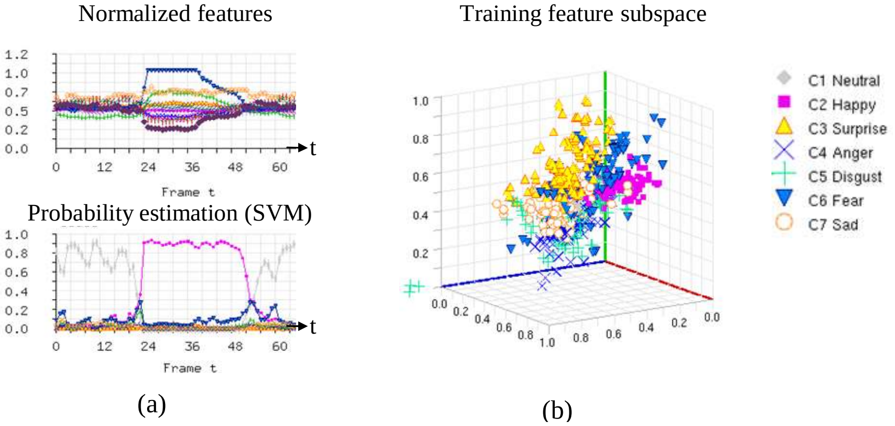

Fig. 13. Facial feature evaluation example. (a) Shows the feature vector over time with category based SVM classification with probability estimation through pairwise coupling [32], and (b) Visualizes the training data feature subspace (features d_1, d_2, d_3)

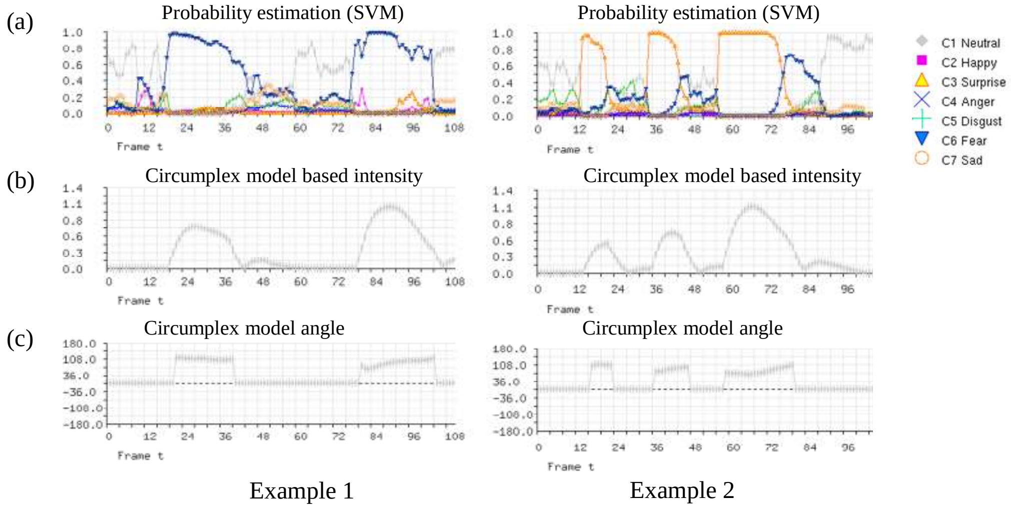

Fig. 14. Intensity example: Emotions have been presented with different expressiveness. (a) Shows the category based classification that does not provide any hint about intensity information, (b) Reveals the intensity through function *r*(*t*) (2.19), while (c) Shows the corresponding angle information *β*(*t*)

4. CONCLUSION AND OUTLOOK

In this work we propose a complete system concept for facial expression analysis with inference of valence-arousal (V-A) information from 3-D and image data. In particular, we apply point cloud based pose estimation, feature normalization, V-A estimation and dynamic integration plus determination of the intensity level. The achieved results provide more information about the affective user state compared to the usual category based emotion classification, which is constricted due to ambiguities on the one hand and rare incidence of basic emotions in authentic, real interaction on the other hand. However, the comparison with sole category based state-of-the-art recognition techniques shows an improvement of our method by roughly 15 percent, in the way that also the neutral class can be regarded, while maintaining the same competitive recognition rate like other methods, e.g. in [30].

Even though overlap of categories does also exist after V-A transformation, this is not necessarily a bad property; it is simply due to the fact that in the emotion model, certain states are nearby. Thus, the V-A space provides a tendency of the user state, like “negative / positive / aroused or not”, rather than one special discrete emotion, which is unlikely to happen. For a HCI system this information can be much more valuable. Thus, the presented methods have fundamentally a reasonable impact on the usability of new human computer interfaces.

In future work we want to explore several modifications in the presented concept. First, in order to ease applicability, we want to generalize and automate the adaptation of the different person specific models that are applied throughout analysis. Also, we will increase the machine's perceptive abilities through the application of advanced sensor technology, like near infrared (NIR) and high speed cameras, what is supposed to provide new features and to increase the application range. Also we want to apply new detection algorithms, like the one of Ren et al. [33], which can robustly and quickly deliver a greater number of image features.

Moreover, we also intend to adapt our method to new application domains. In particular, at the moment, we do not use the lower right part of the valence-arousal plane. However, this part includes states such as sleepiness and calmness. In future work we also want to use this quadrant, because it may enable the analysis of vigilance in medical projects and sleep detection in automotive applications.

ACKNOWLEDGEMENT

This work was supported by DFG-Transregional Collaborative Research Centre SFB/TRR 62.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Pantic M, Pentland A, Nijholt A, Huang TS. Human computing and machine understanding of human behavior: A survey. In Artificial Intelligence for Human Computing; 2007.
2. Cohn JF, Ekman P. Measuring facial action. In: Handbook of nonverbal behavior research methods in the affective sciences. Oxford Univ. Press. 2008;9-64.
3. Littlewort G, Bartlett MS, Salamanca LP, Reilly J. Automated measurement of children’s facial expressions during problem solving tasks. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG ’11). 2011;30–35.
4. Whitehill J, Serpell Z, Lin YC, Foster A, Movellan JR. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. In IEEE Transactions on Affective Computing. 2014;5(1):86–98.
   DOI: 10.1109/TAFFC.2014.2316163
5. Yang S, Kafai M, An L, Bhanu B. Zapping Index: Using smile to measure advertisement zapping likelihood. In IEEE Transactions on Affective Computing. 2014;5(4):432–444.
   DOI: 10.1109/TAFFC.2014.2364581
6. Littlewort G, Whitehill J, Wu T, et al. The computer expression recognition toolbox (CERT). In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition andWorkshops (FG ’11). 2011;298–305.
7. Ekman P, Friesen W. Facial action coding system: A technique for the measurments of facial movements. Consulting Psychologists Press; 1978.
8. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended

Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '10). 2010;94–101.

9. Saeed Anwar, Al-Hamadi Ayoub, Niese Robert. The effectiveness of using geometrical features for facial expression recognition. In Cybernetics (CYBCONF), IEEE International Conference. 2013;122–127.

10. Shan C, Gong S, McOwan PW. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing. 2009;27(6): 803–816.

11. Zhang W, Shan S, Gao W, Chen X, Zhang H. Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A novel non-statistical model for face representation and recognition. In Proceedings of the10th IEEE International Conference on Computer Vision (ICCV '05). 2005;1:786–791.

12. Valstar M, Pantic M, Patras I. Motion history for facial action detection in video. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '04). 2004;1:635–640.

13. Zhu Y, De Silva C, Ko C. Using moment invariants and hmm in facial expression recognition. In Proceedings of the 4th IEEE Southwest Symposium on Image Analysis and Interpretation. 2000;305–309.

14. Wollmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas E, Cowie R. Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies. In Interspeech. 2008;597–600.

15. Nicolaou MA, Gunes H, Pantic M. A multi-layer hybrid framework for dimensional emotion classification. In ACM Multimedia. 2011;933–936.

16. Russel JA. A circumplex model of affect. Journal of Personality and Social Psychology. 1980;39(6):1161-1178.

17. Niese R, Al-Hamadi A, Heuer M, Michaelis B, Matuszewski B. Machine vision based recognition of emotions using the circumplex model of affect. Int'l Conference on Multimedia Technology (ICMT). 2011;6424-6427.

18. "Facegen modeller; 2015.
Available:http://facegen.com/modeller.htm, March

19. Xiong X, De la Torre F. Supervised descent method and its applications to face alignment. IEEE Computer Vision and Pattern Recognition; 2013.

20. Szeliski R. Computer vision - algorithms and applications. Springer; 2011.
ISBN: 978-1-84882-935-0.

21. Pandzic I, Forchheimer R. MPEG-4 facial animation: The standard, implementation and applications; 2002.
ISBN-13: 978-0470844656.

22. Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection. IEEE ICIP. 2002;1:900-903.

23. McGlone C, Mikhail E, Bethe J. Manual of photogrammetry. ASPRS; 2004.
ISBN: 1-57083-071-1, 5th edition.

24. Foley JD, van Dam A, Feiner S, Hughes J. Computer graphics: Principles and practice. Addison-Wesley, 3rd Edition; 2013.

25. Rusinkiewicz S, Levoy M. Efficient variants of the ICP algorithm. Proc. of the 3rd Int. Conf. on 3D Digital Imaging & Modeling. 2001;145–152.

26. Niese R, Werner P, Al-Hamadi A. Accurate, fast and robust realtime face pose estimation using kinect camera. IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2013;487-490.

27. Calder AJ, Lawrence AD, Young AW. Neuropsychology of fear and loathing. Nature Reviews Neuroscience. 2001;2: 352-363.

28. Haykin S. Neural networks: A comprehensive foundation. (3rd Edition) Prentice-Hall, Inc., Upper Saddle River, NJ, USA; 2008.
ISBN-13: 978-0131471399.

29. Yin L, Chen X, Sun Y, Worm T, Reale M. A high-resolution 3D dynamic facial expression database. IEEE International Conference on Automatic Face and Gesture Recognition (FG'08). 2008;1-6.

30. Wang J, Yin L, Wei X, Sun Y. 3D facial expression recognition based on primitive surface feature distribution. IEEE Int'l. Conference on Computer Vision and Pattern Recognition. 2006;1399-1406.

31. Niese R, Al-Hamadi A, Michaelis B. Emotion recognition based on 2D-3D facial feature extraction from color image sequences. JMM, Academy Publisher. 2010;5:488-500.

32. Wu TF, Lin CJ. Probability estimates for multi-class classification by pair wise

coupling. Journal of Machine Learning Research. 2004;5:975-1005.

33. Ren S, Cao X, Wei Y, Sun J. Face alignment at 3000 FPS via regressing local binary features. IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA. 2014;1685–1692.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/15337