Analysis of Feature Space for a 2d/3d Vision based Emotion Recognition Method

Robert Niese, Ayoub Al-Hamadi, 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 (HCI), emotion recognition is becoming an imperative characteristic. The quest for effective and reliable emotion recognition in HCI has resulted in a need for better face detection, feature extraction and classification. In this paper we present results of feature space analysis after briefly explaining our fully automatic vision based emotion recognition method. We demonstrate the compactness of the feature space and show how the 2d/3d based method achieves superior features for the purpose of emotion classification. Also it is exposed that through feature normalization a widely person independent feature space is created. As a consequence, the classifier architecture has only a minor influence on the classification result. This is particularly elucidated with the help of confusion matrices. For this purpose advanced classification algorithms, such as Support Vector Machines and Artificial Neural Networks are employed, as well as the simple k-Nearest Neighbor classifier.

Keywords—Facial expression analysis, Feature extraction, Image processing, Pattern Recognition, Application.

I. INTRODUCTION

In recent years there has been a growing interest in improving the modalities 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, what is focused in this article. Generally, facial expression analysis facilitates information about emotions [1], person perception and it gives insight to interpersonal behavior. 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, accurate, quantitative measurements in diverse applications, it is needed to develop automated methods for face detection, robust feature extraction and classification, which still cannot be done by conventional methods under real world conditions in real-time. A common demand is the reliability across changes in pose, illumination and expressions (PIE). Even though a number of visual emotion analysis methods have been introduced in the literature, for still images as well as for image sequences [2, 3, 4, 5], the criteria of varying PIE still cannot be met reliably. Often this is due to inappropriate processing of image features.

In this paper we show, that by incorporating 3d context information [6] a reliable feature space can be constructed for robust emotion recognition from images and image sequences. After briefly explaining our method in the next section, in chapter 3 we demonstrate the compactness of the feature space and show how the 2d/3d based technique achieves superior features for the purpose of emotion classification. Also it is exposed that through feature normalization a widely person independent feature space is created. As a consequence, the classifier architecture has only a minor influence on the classification result. This is particularly elucidated with the help of confusion matrices. For this purpose advanced classification algorithms, such as Support Vector Machines and Artificial Neural Networks are employed, as well as the simple k-Nearest Neighbor classifier.

II. APPLIED METHODS

In the presented work for automated visual emotion recognition, the underlying methods are based on 2d/3d processing. For the interchange between 2d image and 3d space a calibrated monocular color camera is used. Also subject registration is done in order to generate a personalized 3d surface model analogues to [7], and to determine the feature set of the person’s neutral face. Summarizing, our method consists of three parts, i.e. preprocessing, computation of expression features and classification (Fig. 1). After a brief introduction of the method, evaluation results of the feature space and classification are presented in the next chapter.

A. Preprocessing and pose estimation

Inspired by the Facial Animation Parameter (FAP) system that is contained in the MPEG-4 framework [8]; our method includes the processing of a set of meaningful facial feature points. However, unlike the FAP system for animation we only use a subset of points for the process of recognition. In the first step of the processing chain, an Adaboost cascade classifier is applied that detects the subject’s face [9].

Correspondence to: R. Niese, Institute for Electronics, Signal Processing and Communications, University of Magdeburg, Germany.
E-mail: robert.niese@ovgu.de
This work was supported by DFG-Schmurerkennung (FKZ: BR3705/1-1), DFG-Transregional Collaborative Research Centre SFB/TRR 62, and BMBF Bernstein-Group (FKZ: 01GQ0702).
The goal of the pose estimation method is to reduce error measure $e$ (2), which is the sum of squared distances between the image projections of the 3d anchor points $a_i$ and the fiducial image points $i_j$, determined through image processing. This fitting method is solved iteratively.

$$e = \sum_{j=1}^{N} \| t(M_p \cdot a_j) \|_2^2 \rightarrow \min$$  \hspace{1cm} (2)

where $t(.)$ is the world to image transformation based on the camera model, $M_p$ is the pose matrix with respect to the current pose parameters, $i_j$ and $a_j$ are corresponding image and 3d model anchor points, while $N$ is the number of anchor points.

After pose determination the image feature points are projected to the surface model at its current pose, resulting in set $P_f(3)$ consisting of 3d points (Fig. 2b).

$$P_f = \{ p_{be}, p_{ae}, p_{fe}, p_{le}, p_{me}, p_{de}, p_{de}, p_{de}, p_{de}, p_{de}, p_{de}, p_{de} \}, P_f \in \mathbb{R}^6$$  \hspace{1cm} (3)

Using 3d measures according to (3) one automatically compensates issues such as perspective foreshortening and varying face sizes due to back and forth movement what is commonly referred to as pose problem. Also it enables the normalization of features.

### B. Feature extraction and normalization

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 $f$ (4) is directly inferred from point set $P_f$ (3). The features comprise six Euclidean 3d distances $d_{ai}$ (5) across the face and four angles $a_i$ (6), which expose information about the characteristics of the current mouth shape and the overall facial expression state (Fig. 2d). 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 $d_6$.

$$f = ( d_1, \ldots, d_6, a_1, \ldots, a_6 )^T, \quad f \in \mathbb{R}^{10}, \quad d_{ai}, a_i \in \mathbb{R}$$  \hspace{1cm} (4)

$$d_i = \| p_{preb} - p_{pre} \|, \quad p_k \in \mathbb{R}^3, \quad \text{etc.}$$  \hspace{1cm} (5)
Feature vector $f_{\text{neutral}}$ 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 $f_{\text{neutral}}$ and $f_i$ resulting in $r_{\text{ratio}}$ (8). In particular, the operator $\#$ for component wise division of two feature vectors $a$ and $b$ shall be defined as $a \# b = (a_1/b_1, a_2/b_2, \ldots, a_9/b_9) \in \mathbb{R}^{10}$, $a, b \in \mathbb{R}^{10}$ (7)

$$f_{\text{ratio}} = f_i \# f_{\text{neutral}}, f_i, f_{\text{neutral}} \in \mathbb{R}^{10}$$ (8)

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 $c_{\text{min}}$ and $c_{\text{max}}$ (9) have been computed for each feature distribution across the training data. Applying normalization to the feature ratio vector, the ultimate feature vector $f_{\text{norm}}$ is created (10).

$$c_{\text{min}} = \mu - 2\sigma, c_{\text{min}} \in \mathbb{R}^{10}$$

$$c_{\text{max}} = \mu + 2\sigma, c_{\text{max}} \in \mathbb{R}^{10}$$ (9)

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

$$f_{\text{norm}} = (f_{\text{ratio}} - c_{\text{min}}) / (c_{\text{max}} - c_{\text{min}}), f_{\text{norm}} \in \mathbb{R}^{10}$$ (10)

### C. Classification

In the analysis of normalized feature vectors, three supervised classifiers have been compared. The classifier input is the normalized ten dimensional feature vector $f_{\text{norm}}$ and the output one of the predefined emotion classes. At the moment, we distinguish five classes. Also the feature space has been scrutinized by applying an unsupervised learning algorithm based on a Self Organizing Map (SOM) [12].

**k-Nearest Neighbors:** The k-NN classifier [13] generally achieves good classification results when the training data is well representative and consistent. This technique is one of the simplest machine learning algorithms and requires only an accumulation of labeled template samples for training, which are further used during decision. The distance between a test and the training samples can be computed in several ways. In this work, the Euclidean distance metric is applied and a simple majority vote is used with the parameter selection of $k=5$, which has been determined through the heuristic technique of cross validation.

**Multi Layer Perceptron:** The classification technique of multi-layer artificial neural networks is applied in this work, whereas 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 [14], the input layer has $n_i=10$ neurons and the output layer $n_o=5$ neurons. The Fast Artificial Neural Network Toolbox [15] has been used for the implementation.

**Support Vector Machines:** Generally, the SVM learner is based on an underlying two-class or binary classification in which it is attempted to maximize the hyper plane margin between the classes [16]. The Pairwise Coupling extension is used to adapt SVM for the multi-class problem [17]. In this work, the Radial-Basis-Function (RBF) Gaussian kernel is used which has performed robustly with the given number of features and provided optimum results as compared to other kernels. For the optimization, kernel width $\sigma=3$ and the penalty parameter $C=5$ are used. For more details the reader may refer to [16]. The libSVM implementation has been used for software realization [18].

**Self Organizing Map:** Characteristically, the SOM also known as Kohonen map is a type of artificial neural network that consists of neurons that are arranged in a grid [12, 19]. Thereby, associated with each neuron is a weight vector of the same dimension as the input data and a position in the map space. Typically, the net is trained using unsupervised learning. In theoretical consideration, the SOM represents an approximation of the probability density function of the input data. Biologically motivated, the SOM has the specialty to preserve topological properties of the input space while producing a low, typically two dimensional, discretized representation of that input space. This makes it useful for evaluating and visualizing high-dimensional feature data.

### III. Experimental Results

In the following, results of feature space analysis are presented that have been gained with our 2d/3d based method. In particular, a database has been built for training and testing, comprising two sets of data, every one containing about 3800 normalized feature samples of five facial expressions from 10 subjects each. The persons in the two sets are different. Included are four emotion relevant expressions, i.e. Joy, Surprise, Anger and Disgust and a fifth one of neutrally talking subjects, thus, creating variations in the area of the mouth region. This is of special interest for practical HCI applications. Also the scenarios contain pose variations.

Analysis of the high dimensional feature space was carried out in order to estimate the quality of the feature extraction method. For this purpose, the neural network of a Self-Organizing Map has been learned with the feature data set. Particularly, the so-called U-matrix (unified distance matrix) representation [20] has been determined which visualizes the distances between the adjacent neurons and offers a fast way to get insight to the inherent data distribution (Fig. 3a). The distance is represented by different shadings between the adjacent nodes. Dark shading between the neurons reflects a large distance and thus a gap between the values in the input space. A light coloring between the neurons signifies that the feature vectors are close to each other in the input space. Thus, light areas can be thought of as clusters and dark areas as cluster
separators. Additionally, the class labels are plotted. As can be seen in Fig. 3a, the 2d SOM feature space representation clearly provides excellent separateness between the classes, which gives evidence that the feature space is suitable for classification.

Same observation can be made through the linear dimensionality reduction technique, namely, principal components analysis (PCA). In Fig. 3b the first three principal components $K_1, K_2, K_3$ of the dimension reduced feature samples are plotted. Obviously, all classes are represented by relatively separated clusters. There is only a certain overlap between the classes $C_4$ and $C_5$, i.e. anger and disgust. The first three components contain more than 86 percent of the overall variance.

The compactness of feature data samples belonging to the same class can also be evaluated in the following way. For every of the five classes $C_i$ let $\mu_i$ be the feature mean, in other words the class center in $R^{10}$. Considering Euclidean distances $d_{C_i}$ (10) between these class centers and all training samples $f_{norm}$, the class separateness becomes evident (Fig. 4a-d).

$$d_{C_i} = || \mu_i - f_{norm} ||, \quad d_{C_i} \in R, \mu_{norm} \in R^{10}$$

The feature space has been analyzed for both of the data sets, thus, for the data of groups of different subjects resulting in analogous outcome.

Further, classification results underline the assumption of a well separated feature space. The classification accuracy for our test data can be analyzed by the following confusion matrices (Table 1, 2, 3) [20], which contain information about the actual classes $C_i$ and their prediction $P(C_i)$, based on the particular classifier. For the classes $C_1$ to $C_4$ the recognition rates are high and the results are mostly independent of the classifier. Only with $C_5$ there is noticeable confusion in the classification results which is strongest with k-NN and lowest with SVM. This follows from the mixing in the feature space between samples belonging to $C_4$ and $C_5$. Inverted training and testing between the two feature sets gives comparable results for all of the confusion matrices. Stratified cross validation also confirms the presented results. Consequently, as the two feature sets contain groups of 10 different subjects each, the results support the hypothesis that, with the normalized features, we have created a largely person independent feature space.

<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>$C_1$</td>
<td>86.63</td>
<td>0.17</td>
<td>8.29</td>
<td>4.57</td>
<td>0.34</td>
</tr>
<tr>
<td>$C_2$</td>
<td>1.90</td>
<td>86.20</td>
<td>0.38</td>
<td>3.67</td>
<td>7.85</td>
</tr>
<tr>
<td>$C_3$</td>
<td>1.59</td>
<td>0.00</td>
<td>99.41</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$C_4$</td>
<td>7.46</td>
<td>0.00</td>
<td>0.00</td>
<td>86.78</td>
<td>5.76</td>
</tr>
<tr>
<td>$C_5$</td>
<td>4.74</td>
<td>0.00</td>
<td>0.00</td>
<td>46.02</td>
<td>49.24</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix: k-NN, with $C_1$ Neutral, $C_2$: Happy, $C_3$: Surprise, $C_4$: Anger, $C_5$: Disgust

<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>$C_1$</td>
<td>93.40</td>
<td>0.00</td>
<td>6.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$C_2$</td>
<td>4.94</td>
<td>91.39</td>
<td>0.00</td>
<td>1.27</td>
<td>2.41</td>
</tr>
<tr>
<td>$C_3$</td>
<td>7.63</td>
<td>0.00</td>
<td>92.37</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$C_4$</td>
<td>6.95</td>
<td>0.34</td>
<td>0.00</td>
<td>81.69</td>
<td>11.02</td>
</tr>
<tr>
<td>$C_5$</td>
<td>0.61</td>
<td>1.68</td>
<td>3.82</td>
<td>32.87</td>
<td>61.01</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix: MLP

<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>$C_1$</td>
<td>91.03</td>
<td>0.00</td>
<td>7.61</td>
<td>0.85</td>
<td>0.51</td>
</tr>
<tr>
<td>$C_2$</td>
<td>3.92</td>
<td>90.25</td>
<td>0.00</td>
<td>0.89</td>
<td>4.94</td>
</tr>
<tr>
<td>$C_3$</td>
<td>0.16</td>
<td>0.00</td>
<td>99.84</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$C_4$</td>
<td>0.52</td>
<td>0.00</td>
<td>0.00</td>
<td>95.93</td>
<td>3.56</td>
</tr>
<tr>
<td>$C_5$</td>
<td>0.61</td>
<td>0.00</td>
<td>1.53</td>
<td>20.80</td>
<td>77.06</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix: SVM

Fig. 3 Analysis of training samples; a-e) contain a section with data of each of the five classes, the plots show Euclidean distances $d_{C_i}$ to the center of class $C_i$ in $R^{10}$.
Fig. 4 Analysis of feature space demonstrates the good separability of the classes; a) U-Matrix of a Self Organizing Map with added class labels, b) distribution in $K_1$, $K_2$, $K_3$ space after PCA dimension reduction.

Var $K_1$ = 0.4358
Var $K_2$ = 0.2709
Var $K_3$ = 0.1556

Fig. 5 Example sequence “Happy”, a) image snapshots, b) normalized features, c, d, e) classification based on k-NN, MLP and SVM.

Class $C_1$, c) k-NN Classification

Class $C_2$, d) MLP Classification

Class $C_3$, e) SVM Classification
IV. CONCLUSION AND VIEW

An automatic 2d/3d approach for the recognition of basic emotion expressions has been presented, which through feature normalization, creates of a nearly person independent feature space. Analysis of that space has shown that fine separation between the classes has been achieved. This in turn leads to the observation that in the proposed method, we are relatively independent of the classifier used. At he moment, typical benchmark tests with competitive approaches on public databases cannot readily be performed due to the need of 3d context information, such as camera calibration and person specific surface model data. However, in future work we are going to use generic surface models to address this issue. Also, this will offer us new opportunities to gain training and testing data. Here, the current framework is ready to include additional classes. As shown in the analysis, the achieved feature space still offers room for this.

REFERENCES