International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:3, No:11, 2009

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.

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-Schmerzerkennung (FKZ: BR3705/1-1), DFG-Transregional Collaborative Research Centre SFB/TRR 62, and BMBF Bernstein-Group (FKZ: 01GQ0702).

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].

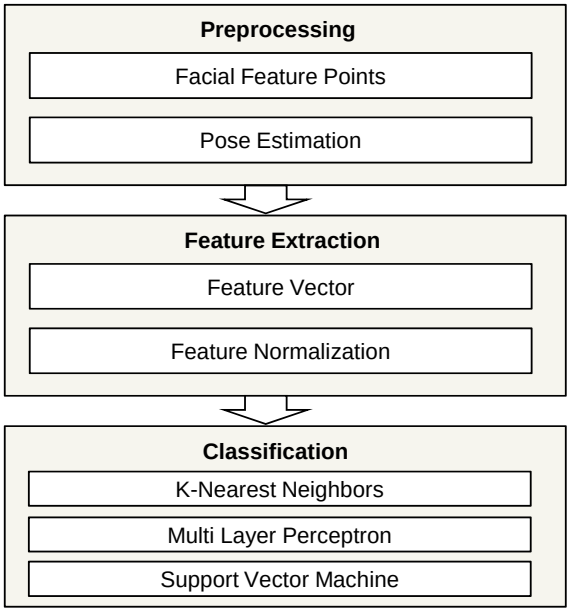

Fig. 1 The applied methods for vision based emotion recognition

Next step is the localization of facial points through image processing techniques. Here, the eye center points i_{le} / i_{re} and mouth contour m_c are extracted (Fig. 2a). From this contour, the mouth corner points i_{lm} / i_{rm} plus upper and lower lip points i_{ul} / i_{ll} are selected. The HSV color model [10] is used to extract the facial feature points, which is appropriate in order to exploit the behavior of facial feature areas under different color channels. The eyebrow points i_{leb} / i_{reb} are determined next using gradient information [6]. The complete set of feature points is summarized in (1) (see Fig. 2a).

$$I_f = \{i_{le}, i_{re}, i_{leb}, i_{reb}, i_{lm}, i_{rm}, i_{ul}, i_{ll}\},\ i_j \in R^2 \quad (1)$$

It is necessary to determine the current face pose in order to infer 3d facial measures. The current face position and orientation is defined through model pose parameter set $t_v = \{t_x, t_y, t_z, t_\omega, t_\varphi, t_\kappa\}$, which contains three translation and three rotation components [11]. So called anchor points $\{a_{re}, a_{le}, a_{rm}, a_{lm}, a_n\}$ are defined across the 3d face model (see Fig. 2c), which are used to estimate the current pose. According to camera model C_m, the image projection of each anchor point is determined. 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.

$$e = \sum_{j=1}^{N} \| i_j - t(M_p \cdot a_j) \|^2 \rightarrow \min \quad (2)$$

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).

$$P_f = \{ p_{le}, p_{re}, p_{leb}, p_{reb}, p_{lm}, p_{rm}, p_{ul}, p_{ll} \},\ p_j \in R^3 \quad (3)$$

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_m (5) across the face and four angles α_n (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.

$$f = (\, d_1 \ldots d_6\ \alpha_1 \ldots \alpha_4)^{\mathrm{T}},\ f \in R^{10},\ d_m,\ \alpha_n \in R \quad (4)$$

$$d_1 = \| p_{reb} - p_{re} \|,\ p_k \in R^3, \text{ etc.} \quad (5)$$

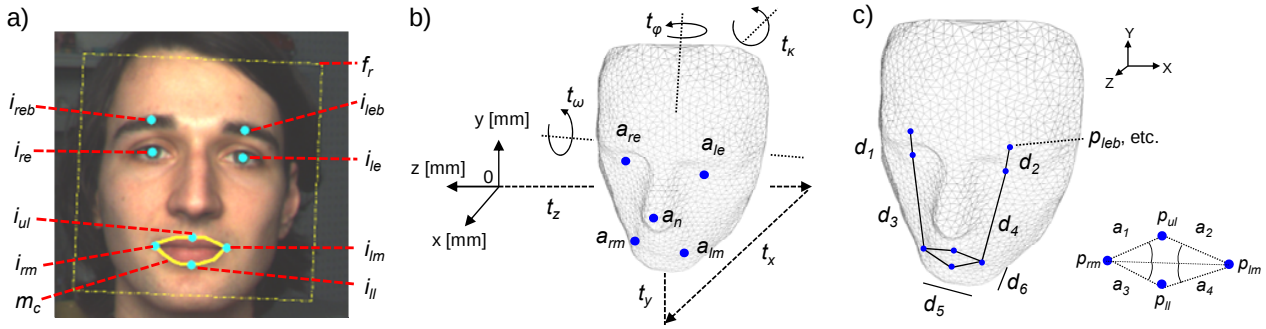

Fig. 2 Feature points, a) image processing, face detection with face rectangle f_r, mouth contour m_c, and feature points i_j, b) face model with anchor points a_i and pose parameters, c) face model with projected feature points p_j and 3d features, i.e. distances d_i and angles α_j

$$a_1 = \arccos\left(\frac{(p_{lm} - p_{ul}) \cdot (p_{ll} - p_{ul})}{\|p_{lm} - p_{ul}\| \cdot \|p_{ll} - p_{ul}\|}\right),\ p_k \in R^3, \text{ etc.} \qquad (6)$$

Feature vector $f_{neutral}$ 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 $f_{i\,ratio}$ (8). In particular, the operator # for component wise division of two feature vectors a and b shall be defined as (7)

$$a \# b = (a_1 / b_1 \;\; a_2 / b_2 \;\; \ldots \;\; a_{10} / b_{10}) \in \mathrm{R}^{10},\ a, b \in \mathrm{R}^{10} \qquad (7)$$

$$f_{i\,ratio} = f_i \# f_{neutral},\ \ f_{i\,ratio},\ f_i,\ f_{neutral} \in \mathrm{R}^{10} \qquad (8)$$

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

$$\begin{aligned} c_{min} &= \mu - 2\sigma,\ c_{min} \in \mathrm{R}^{10} \\ c_{max} &= \mu + 2\sigma,\ c_{max} \in \mathrm{R}^{10} \end{aligned} \qquad (9)$$

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

$$f_{i\,norm} = (f_{i\,ratio} - c_{min}) \# (c_{max} - c_{min}),\ \ f_{i\,norm} \in \mathrm{R}^{10} \qquad (10)$$

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_{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 ni=10 neurons and the output layer no=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 σ=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 μ_i be the feature mean, in other words the class center in R^{10} Considering Euclidean distances dC_i (10) between these class centers and all training samples f_{norm}, the class separateness becomes evident (Fig. 4a-d).

$$dC_i = || \mu_i - f_{norm} ||, \; dC_i \in R, \mu_i, f_{norm} \in R^{10} \qquad (11)$$

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.

Class	$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$
C_1	86.63	0.17	8.29	4.57	0.34
C_2	1.90	86.20	0.38	3.67	7.85
C_3	1.59	0.00	99.41	0.00	0.00
C_4	7.46	0.00	0.00	86.78	5.76
C_5	4.74	0.00	0.00	46.02	49.24

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

Class	$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$
C_1	93.40	0.00	6.60	0.00	0.00
C_2	4.94	91.39	0.00	1.27	2.41
C_3	7.63	0.00	92.37	0.00	0.00
C_4	6.95	0.34	0.00	81.69	11.02
C_5	0.61	1.68	3.82	32.87	61.01

Table 2: Confusion matrix: MLP

Class	$P(C_1)$	$P(C_2)$	$P(C_3)$	$P(C_4)$	$P(C_5)$
C_1	91.03	0.00	7.61	0.85	0.51
C_2	3.92	90.25	0.00	0.89	4.94
C_3	0.16	0.00	99.84	0.00	0.00
C_4	0.52	0.00	0.00	95.93	3.56
C_5	0.61	0.00	1.53	20.80	77.06

Table 3: Confusion matrix: SVM

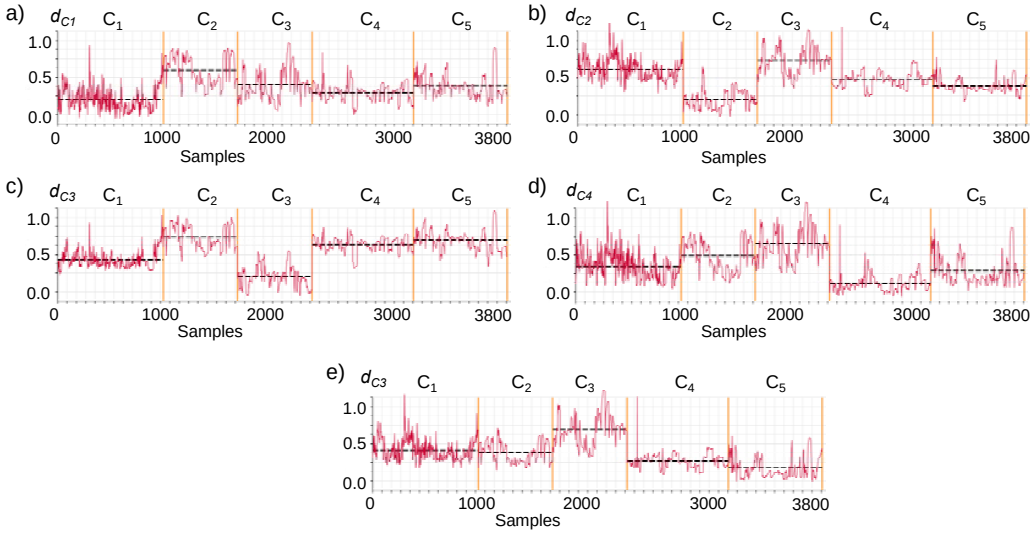

Fig. 3 Analysis of training samples; a-e) contain a section with data of each of the five classes, the plots show Euclidean distances dC_i to the center of class C_i in R^{10}

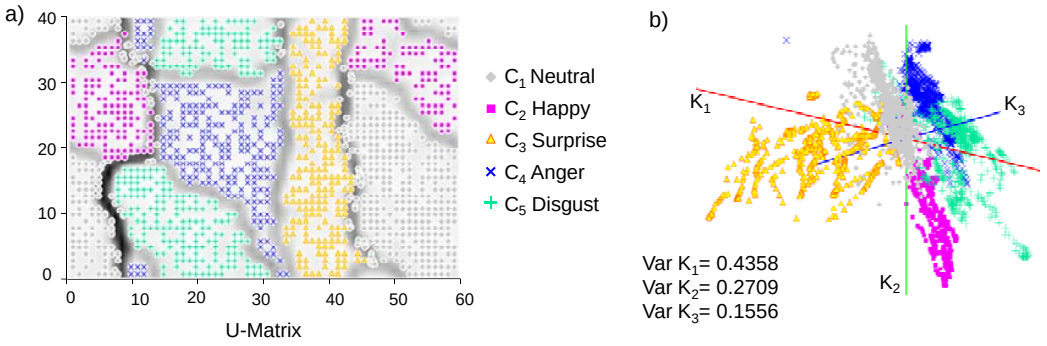

Fig. 4 Analysis of feature space demonstrates the good separability of the classes; a) U-Matrix of a Self Organizing Map with added class labels, b) distribution in K_1, K_2, K_3 space after PCA dimension reduction

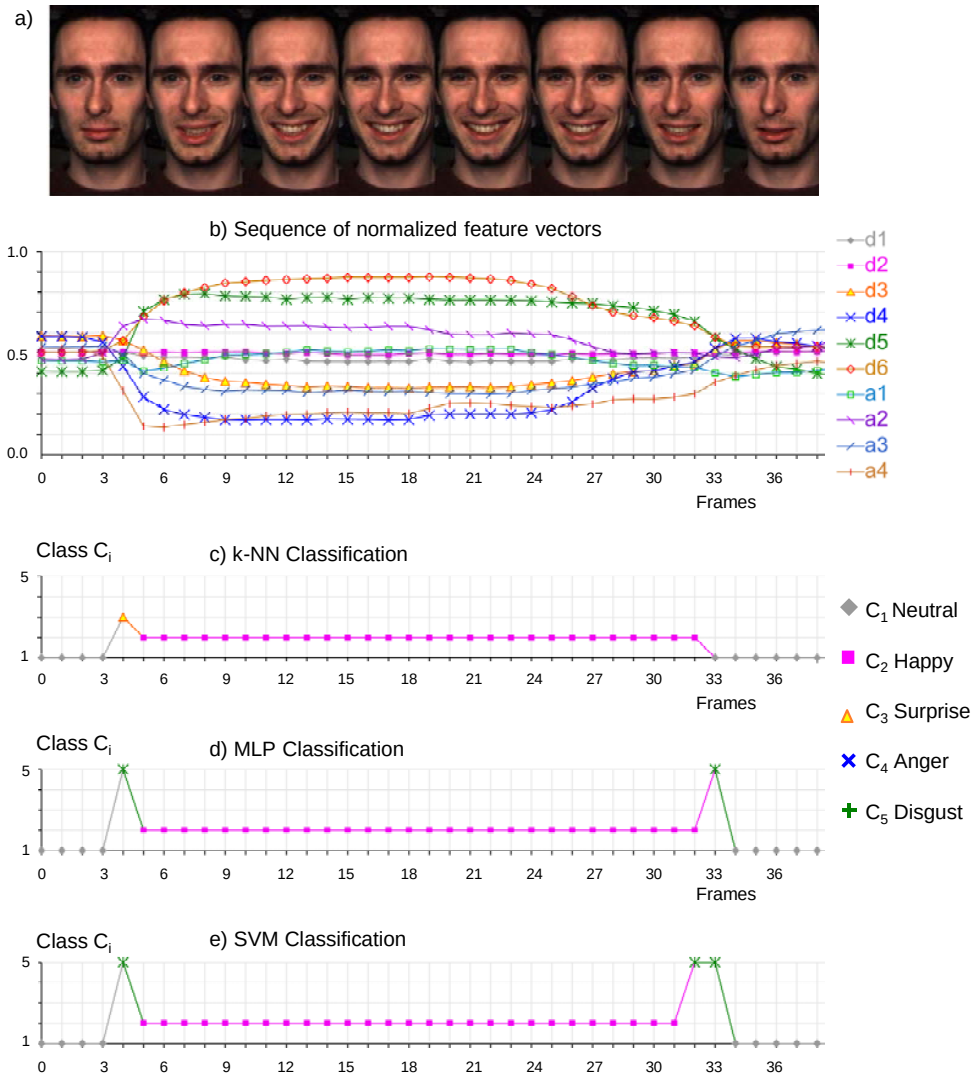

Fig. 5 Example sequence "Happy", a) image snapshots, b) normalized features, c, d, e) classification based on k-NN, MLP and SVM

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

[1] P.E. Ekman, W.V. Friesen. Facial Action Coding System. Consulting Psychologists Press, Palo Alto, CA, 1978.

[2] M. Kunz, V. Mylius, K. Schepelmann, S. Lautenbacher: Impact of age on the facial expression of pain. J of Psychosom Res 2008; 64:311-318.

[3] C.A. Gilbert, C.M. Lilley, K.D. Craig, P.J. McGrath, C.A. Court, S.M. Bennett, C.J. Montgomery: Postoperative Pain Expression in Preschool Children: Validation of the Child Facial Coding System. Clin J Pain 1999; 15:192-200.

[4] G. Littlewort, M.S. Bartlett, I. Fasel, K. Lee: Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain. In ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces, pp. 15–21. 2007.

[5] N. Esau, L. Kleinjohann, B. Kleinjohann: Integration of Emotional Reactions on Human Facial Expressions into the Robot Head MEXI. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IEEE/RSJ IROS 2007). 2007.

[6] C. Hu, R. Feris, M. Turk, Real-time View-based Face Alignment using Active Wavelet Networks, Proc. IEEE, IEEE Int'l Workshop on Analysis and Modeling of Faces and Gestures, pp.215, 2003.

[7] D. Vukadinovic, M. Pantic: Fully automatic facial feature point detection using Gabor feature based boosted classifiers, Systems, Man and Cybernetics, 2005 IEEE International Conference on, vol 2, pp. 1692-1698, 2005.

[8] N. Esau, E. Wetzel, L. Kleinjohann, B. Kleinjohann: Real-Time Facial Expression Recognition Using a Fuzzy Emotion Model. In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2007). 2007.

[9] A.B. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, B.J. Eheobald: The Painful Face - Pain Expression Recognition Using Active Appearance Models: In ICMI. 2007.

[10] M. Monwar, S. Rezaej, K. Prkachin: Eigenimages Based Pain Expression Recognition. In IAENG International Journal of Applied Mathematics, 36:2. 2007.

[11] S. Brahnam, C.F. Chuang, F.Y. Shih, M.R. Slack: SVM Classification of Neonatal Facial Images of Pain. Fuzzy Logic and Applications, WILF 2005, Crema, Italy, 15-17, 2005, LNCS, vol. 3849, 2006b.

[12] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of Facial Expression Extracted Automatically from Video. Image and Vision Computing, vol. 24, pp. 615–625, 2006.

[13] Pantic, M., Pentland, A., Nijholt, A., Huang, T.S.: Human Computing and Machine Understanding of Human Behavior: A Survey, in Artificial Intelligence for Human Computing, 2007.

[14] P. Hancock, C. Frowd, E. Brodie, C. Niven: Recognition of Pain Expressions. Progress in Neural Processing, vol. 16, pp. 339–348, 2005.

[15] S.Z. Li., A.K. Jain: Handbook of Face Recognition, ISBN: 0-387-40595-X, 2005.

[16] P. Viola, M. Jones: Rapid object detection using a boosted cascade of simple features, Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, 2001.

[17] B. Lucas, T. Kanade: An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679, 1981.

[18] J. Albertz, W. Kreiling: Photogrammetric Guide, Herbert Wichmann Verlag GmbH, Karlsruhe, 1989.

[19] A. Al-Hamadi, R. Niese, B. Michaelis: A robust approach for contour extraction and tracking of moving objects in video sequences, Signal processing, pattern recognition, and applications (Greece June 30 - July 2, 2003) proceedings: Acta Press, pp. 336 - 341, 2003.

[20] V. Blanz, T. Vetter: A Morphable Model for the Synthesis of 3D Faces, SIGGRAPH, Conference Proceedings, 187-194, 1999.

[21] P. Albrecht, B. Michaelis: Stereo Photogrammetry with Improved Spatial Resolution, ICPR, pp. 845, 1998.

[22] R. Niese, A. Al-Hamadi, B. Michaelis: A Stereo and Color-based Method for Face Pose Estimation and Facial Feature Extraction. ICPR 2006 (IEEE), The 18th International Conference on Pattern Recognition, 2006, Hong Kong, Volume: 1, pp. 299-302, 2006.

[23] S. Rusinkiewicz, M. Levoy: Efficient variants of the ICP algorithm, Proc. of the 3rd Int. Conf. on 3D Digital Imaging & Modeling, pp. 145–152, 2001.

[24] R. Calow, B. Michaelis: Markerless Analysis of Human Gait with a Multi-Camera-System. Proceedings of IASTED International Conference Biomedical Engineering (BioMED), pp. 270-275, 2005.

[25] S. Wachter: Verfolgung von Personen, Universität Karlsruhe, Dissertation, ISBN-13: 978-3980321266, 1997.

[26] J.L. Bentley: Multidimensional binary search trees used for associative searching, ACM, 18, pp. 509-517, 1975.

[27] N. Cristianini and J.S Taylor, "An Introduction to Support Vector Machines and other kernel based learning methods", ISBN: 0-521-78019-X, 2001.

[28] R. Herbrich: Learning kernel Classifiers: theory and algorithms, ISBN:0-262-08306-X, 2003.

[29] T.F. Wu, C.J. Lin: Probability Estimates for Multi-class Classification by Pair wise Coupling, Journal of Machine Learning Research 5, 975-1005, 2004.

[30] Chang, C.-C., Lin., C.-J., LIBSVM: a library for support vector machines, Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2009.