Proc. IEEE Int. Conf. on Systems, Man and Cybernetics (SMC 2004), Den Haag, pp. 2203-2208, IEEE omnipress 2004

Statistical and Neural Methods for Vision-based Analysis of Facial Expressions and Gender*

T. Wilhelm, H.J. Böhme and H.M. Groß
Department of Neuroinformatics
Ilmenau Technical University
P.O.Box 100565
98684 Ilmenau, Germany
Torsten.Wilhelm@TU-Ilmenau.de

A. Backhaus
School of Psychology
University of Birmingham
Edgbaston
Birmingham B15 2TT, UK
axb388@bham.ac.uk

Abstract – *It is a prerequisite for the successful application of service robots in human dominated domains to have intuitive and natural man-machine-interfaces. In this context, it is desirable to extract information, such as the gender, age, or emotional state (via facial expressions) about the user to aid in communication. We developed a method to estimate the facial expression and the gender of a person based on statistical data anylysis and neural classifiers.*

Keywords: Image processing, human-robot interaction, automatic facial expression analysis.

1 Introduction

The application of service robots has become more realistic even for tasks in natural environments in the last few years. However, in most cases these robots are still not able to communicate with their human operators in a natural manner. Orders are typically given by touch displays, keyboards or in more advanced systems by speech or gestures. The robot fulfills its task and reports failures and successes on a display or by speech output. But there is still one important part of human to human communication missing. Humans use facial expressions to insinuate agreement and disaffirmation or to express emotions occuring during the communication process. It is an intrinsic part of human communication to be able to infer the emotional state of the communication partners from their facial expressions. This is clearly missing in most of todays human-robot interaction applications but, to our opinion, it is a prerequisite for their successful application in human dominated domains. Furthermore, other information like the gender of the communication partner are of relevance to the communication process, because it enables the robot to adapt itself more to the needs of its current user. This adaptivity and flexibility of the man-machine interface could be achieved by learning strategies that associate characterizations of the user to appropriate behavioral patterns during the communication process. In this work, we address the problem of extracting the needed information from images based on statistical and neural methods.

2 State of the art

The analysis of facial expressions and gender is a rather new research topic in computer vision. Nevertheless, there are different methods addressing this problem. A widely used method which already has commercial applications is the elastic graph matching approach [11] which uses jets of gabor filters placed on a grid describing the facial structure. With this method, it is possible to find faces in images and to extract the relevant information. Elastic graph matching has been applied succesfully to gender recognition and facial expression analysis. Another approach, the Active Apearance Models [1], describes faces with a rather small number of shape and gray value parameters. Starting from an initial estimate, a model can be adapted to describe a given face image and the extracted appearance parameters can be classified according to the task at hand. There are applications of this approach to the classification of gender or facial expressions. A combination of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) for feature extraction and a Nearest Neighbor classifier has been used in [2] to extract facial expressions from static images. We based the work presented in this paper on this approach, because its underlying concepts suggest that it would expand well to the extraction of information other than facial expressions. We show that it can be used to extract the gender from the same image and that the robustness of the system can be improved by using more powerful classifiers.

*0-7803-8566-7/04/$20.00 © 2004 IEEE.

3 Emotions and facial expressions

Note that we are interested in emotions, not facial expressions. In MRI, we want to know whether a user is satisfied with operating our robot or not. But since we are not able to detect emotions directly, we have to look for some visible features that people use to express emotions, which obviously could be facial expressions. However, it is clear, that using a system for analyzing facial expressions and interpreting the output as emotions is a very problematic thing to do. Very tiny differences in facial expressions can correspond to completely different emotions, as shown in the results section. Dealing with emotions and corresponding facial expressions, the first questions that arise are: How many basic emotions can be represented by facial expressions and what are these basic emotions? In psychology, there are six established basic emotions: happiness, sadness, surprice, fear, anger, and disgust [4].

The very same problem of the relation between facial expressions and emotions arises when it comes to recording training or test data. We could advice a subject to show different facial expressions during a recording session, but it is doubtful that the recorded images really represent the expected emotions. Fortunately, there is a way to avoid this problem. The facial action coding system or FACS [5] provides a means of describing facial expressions just by movements of muscles and not by emotions. Instead of advising a subject to look in a certain way, he is advised to move facial muscles in a certain way. The correspondence between emotions, facial expressions and muscle movements was established by psycologists and is taken as ground truth [4].

4 System architecture

A general system for the analysis of facial expressions and gender from images consists of a data acquisition, a facial feature extraction, and a classification step, see Figure 1. These submodules are described in detail in the following sections.

4.1 Data acquisition

4.1.1 Database

We trained our system with the Cohn-Kanade database from the CMU in Pittsburg [10]. For each person there are several sequences with facial expressions starting from a neutral expression. The data is labeled with the Facial Action Coding System FACS, but single sequences can also be attributed with one of the six basic emotions: happiness, sadness, surprice, fear, anger, and disgust [4]. For the classification of emotions we used 6 images from each sequence, the neutral one and 5 others showing a facial expression in various strengths. For gender classification, we used the neutral image from each sequence. Thus, we had 49 female and 23 male examples in our training data set.

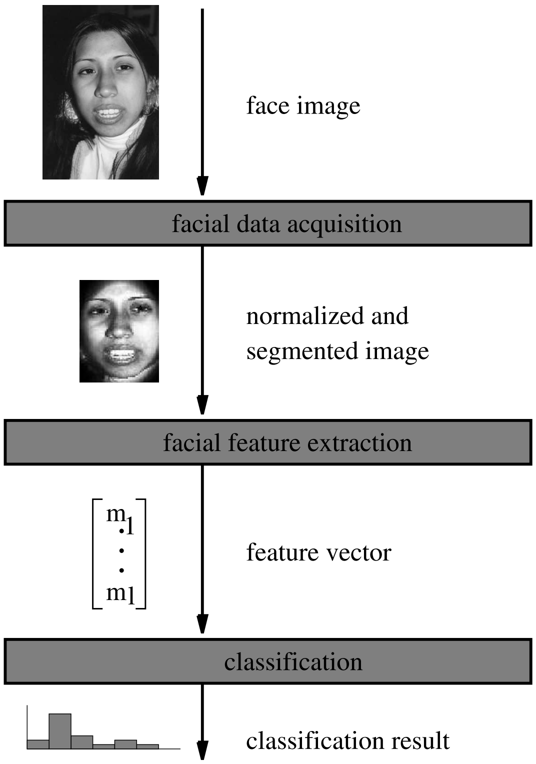


Figure 1: A general system for the analysis of facial expressions and gender consists of data acquisition, facial feature extraction and classification.

4.1.2 Deformation or movement

There are basically two ways of extracting information about facial expressions. The first is purely deformation based, i.e. a snapshot of the face is analyzed for certain local deformations that are typical for some facial expression [13][12]. The other approach uses image sequences to calculate movements of local image regions [3][6]. Often, a neutral image of a person is used as reference to estimate these movements. Since we want to apply our facial expression analysis system on a mobile robot, and we want to estimate the facial expressions of a person during normal operation, it is not clear how such a neutral image could be obtained. Also, for other information like the gender or age a movement based approach would not be suitable. Thus, our system needs to be based on deformations and single images only.

4.1.3 Face normalization

This section describes how the training and test data was extracted from the database. First we manually labeled facial feature points: namely the center of the eyes and the tip of the nose. Then, we applied geometrical transformations such that the facial feature points are on the same position in every image and the eyes are on a horizontal line. The size of these transformed images is 60×70 pixels. For an online version, these facial feature points have to be found automatically. The first version uses an edge orientation model [7] for fine positioning, but the quality of the detection is not yet satisfactory. After this geometrical transformation of the image, we apply a histogram equalization to achieve a further normalization of the appearance.

4.2 Facial feature extraction

Since our feature extraction methods, PCA and ICA, work on one-dimensional data, we have to serialize our images. We use k images of size 60×70 pixels. Serialization can be done by concatenating all rows of one image one after another. In this case, the observation matrix X is a sequence of 4200 k-dimensional observation vectors where observation column i consists of the gray values of pixel i in all images. This representation is called image space. The other possibility is to put each image in one column of the observation matrix X. Here, X is a sequence of k 4200-dimensional observation vectors. This representation is called pixel space.

4.2.1 Gray values

In the most simple case, we did not use any feature extraction, but used the raw image data for classification directly. This version was implemented as a test reference for the other feature extraction methods. Here the training data is simply:

$$D = X \tag{1}$$

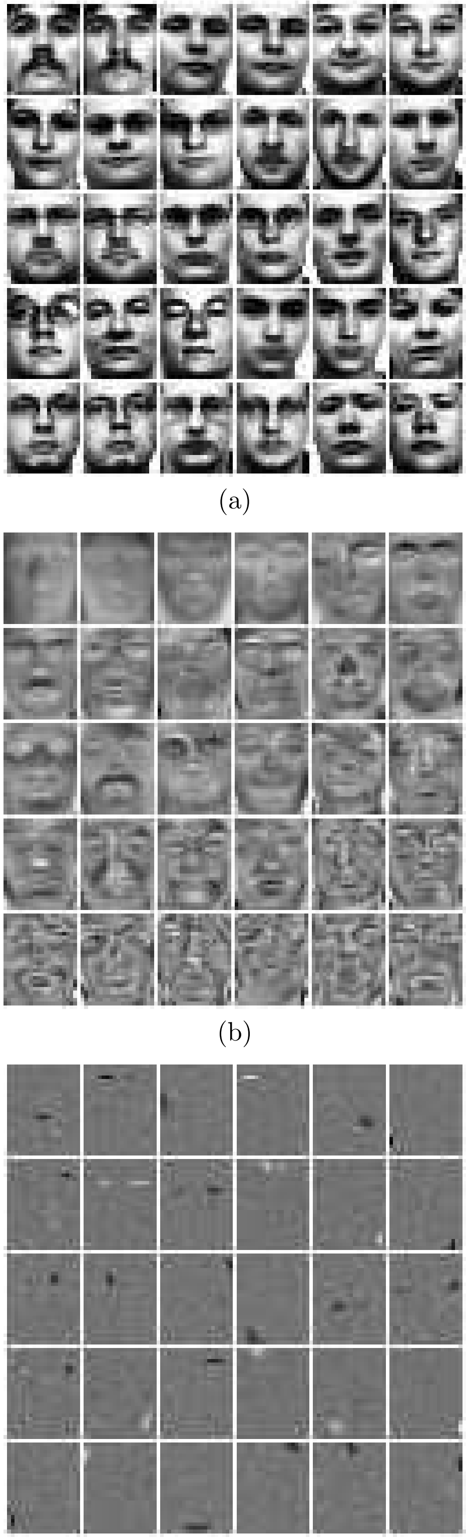
4.2.2 Principal Component Analysis

The Principal Component Analysis (PCA) is carried out in pixel space. The covariance matrix of the observation matrix X is calculated and its eigenvalues and eigenvectors are determined. The eigenvectors can be visualized as images, see Figure 2(b). We reduced the number of eigenvectors spanning the subspace by using only the eigenvectors corresponding to the 200 largest eigenvalues which account for the subspace axes with the highest variance (the sum of the used eigenvalues should be larger than 98% of the sum of all eigenvalues). With T beeing the matrix with eigenvectors in its columns the training data is:

$$D = T^T X \tag{2}$$

4.2.3 Independent Component Analysis

The Independent Component Analysis (ICA) [8][9] is applied in image space. First, the matrix X is whithened such that $E\{X_{sphering} X_{sphering}^T\} = I$. The number of eigenvalues and eigenvectors are calculated as described in the preceeding section. We used the



(a)

(b)

(c)

Figure 2: Illustration of principal and independent components calculated from images from our own database NIFACE. (a) Example faces. (b) Principal components of the faces shown in (a). (c) Independent components calculated from (a) with the FastICA algorithm.

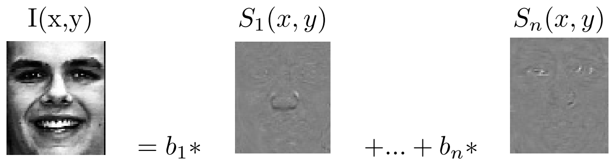


Figure 3: Visualization of the ICA subspace with the independent components shown as basis images. Each image is composed of a weighted sum of basis images which in the case of the independent components represent local facial features.

FastICA-algorithm to calculate the demixing matrix $\hat{W}_{ica}$. The observation matrix X is transformed to its independent components with:

$$\hat{S} = \hat{W}_{ica} W_{sphering} X_{training}. \tag{3}$$

The rows of matrix $\hat{S}$ represent the independent components and can be visualized as images, see Figure 2(c). They are local and represent single facial features. These independent components are the axes of the subspace, the input images are projected on. Figure 3 shows how an image is represented by a set of independent basis images. In this case, the training data is obtained by:

$$D \quad = \quad \hat{S} X^T \tag{4}$$

4.2.4 Discriminant analysis

Using PCA, the size of the eigenvalues provides a measure for choosing the appropriate eigenvectors for classification. The ICA does not provide any order criteria for its components. Thus, we used a criteria introduced in [2], which calculates the suitability of the projection vectors for the classification task at hand. For every row of the matrix D we calculate:

$$\sigma_i = \frac{\sum_{k=1}^{c} (\overline{d}_i^{(k)} - \overline{d}_i)^2}{\sum_{k=1}^{c} \sum_{j \in K_c} (d_{ij}^{(k)} - \overline{d}_i^{(k)})^2} \tag{5}$$

where $\overline{d}_i^{(k)}$ is the mean value of all coefficients of row i that belong to class k. The value $\overline{d}_i$ is the mean value of all coefficients of row i. The $d_{ij}^{(k)}$ are the coefficients of column i in matrix D, that belong to class k. The quotient σ_i is thus the ratio of intra-class variance to the sum of the inner-class variances of the projection values of one independent component. Figure 4 shows an example. Consequently, we use this type of discriminant analysis for the principal components, too. Thus, the training data is obtained by projecting the observation matrix X on the subspace spanned by the 75 best class discriminating eigenvectors or independent components, respectively.

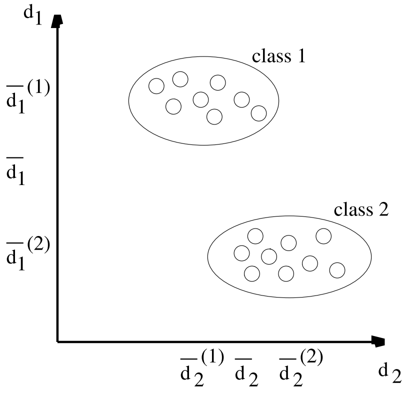


Figure 4: Example for the selection of the features, which discriminate the classes best. For feature d_1 the quotient σ is higher then for d_2, since the projection values on this axes have a higher intra-class variance than for feature d_2 and the distribution of the values from class 1 und 2 are more narrow over d_1, and thus, the sum of inner-class variances is smaller than for d_2.

4.3 Classification

In [2] a very simple Nearest Neighbor (NN) classifier was used. The representative examples from the training set are stored and any new pattern is compared with all the saved patterns by use of the normalized dot product. The pattern is classified with the same classification as the most similar pattern from the training set. In addition to the Nearest Neighbor classifier, we used a Multi Layer Perceptron (MLP) and a Radial Basis Function (RBF) network. All classifiers were trained with gray values directly, reduced principal components, and reduced independent components, separately. In case of the pure gray values, the input consists simply of the gray values of all images, i.e. the unaltered observation matrix X, otherwise the input consists of the projection of this matrix on the principal components or the independent components, respectively.

5 Results

The results for different combinations of the three classifiers with different feature extraction methods are shown in Table 1 and Figure 5. According to the table, the best results are obtained with a RBF network and an independent basis for the classification of facial expressions and with the MLP network and an independent basis for the classification of gender. A significant increase in the detection rates can be shown for neural classifiers in comparison to the Nearest Neighbor classifier used in [2] when using the independent basis.

5.1 Facial expressions

Table 2 shows the confusion matrix of the RBF classifier and Table 3 that of the MLP classifier for the clas-

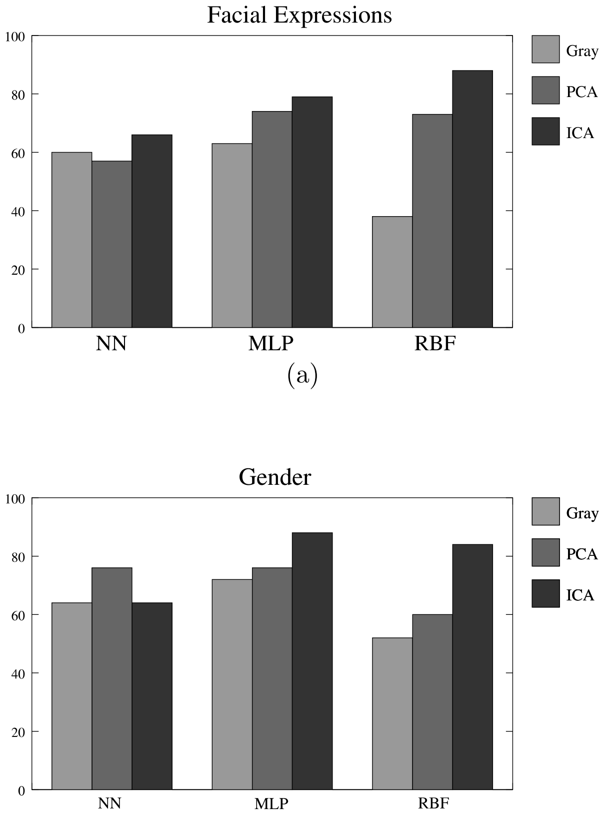
Table 1: Detection rates on the test data for different feature extraction methods and the Nearest Neighbor (NN), Multi Layer Perceptron (MLP), and Radial Basis Function (RBF) classifier, respectively. (a) Classification of facial expressions. (b) Classification of gender.

Facial Expressions	NN	MLP	RBF
gray values	60	63	38
PCA	57	74	73
ICA	66	79	**88**

(a)

Gender	NN	MLP	RBF
gray values	64	72	52
PCA	76	76	60
ICA	64	**88**	84

(b)



(a)

(b)

Figure 5: Detection rates of all classifiers on the test data, see Table 1. (a) Classification of facial expressions. (b) Classification of gender.

Table 2: Confusion matrix for classification of facial expressions using the RBF classifier operating on the 75 best class discriminating independent components.

class **label**	surprise	sadness	anger	fear	disgust	happiness
surprise	67	1	1	1	0	0
sadness	0	81	2	0	0	0
anger	1	9	24	0	4	1
fear	0	0	1	54	0	4
disgust	0	0	3	0	32	0
happiness	0	1	1	9	0	89

Table 3: Confusion matrix for classification of facial expressions using the MLP classifier operating on the 75 best class discriminating independent components.

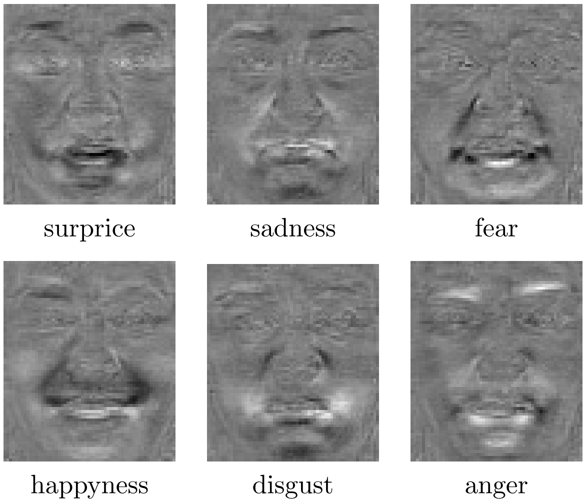
class **label**	surprise	sadness	anger	fear	disgust	happiness
surprise	66	1	2	1	0	0
sadness	0	77	4	0	2	0
anger	1	2	33	1	2	0
fear	2	1	0	53	0	4
disgust	0	5	2	0	28	0
happyness	1	0	6	7	0	86

sification of facial expressions. The numbers on the diagonal are the correct classified patterns and on the off-diagonal the wrong classified patterns. Table 3 shows that the system mixes up ”anger” with ”sadness” or ”happiness” with ”fear”. Looking on the facial expressions of ”anger” and ”sadness” reveals, that the eyebrows are moved down and together for both emotions. Such a similarity exists for ”happiness” and ”fear” as well. In the case of ”happiness”, the corners of the mouth are pulled back and up while in the case of ”fear”, back only.

For a Radial Basis Function network with six hidden neurons, each of the hidden neurons should match to one basic emotion. This can be tested by reconstructing the face images from the independent components and the neuron weights from the hidden layer, see Figure 6.

5.2 Gender

Table 4 shows the confusion matrices of the three classifiers for the classification of gender.



surprice sadness fear

happyness disgust anger

Figure 6: Basis emotions, visualized by use of the ICA basis images and the neuron weights of the hidden layer of the trained RBF network with only six hidden neurons.

Table 4: Confusion matrices for the gender estimation by means of the Nearest Neigbor (NN), the Multi Layer Perceptron (MLP), and the Radial Basis Function (RBF) classifier operating on the 75 best class discriminating independent components.

	NN		MLP		RBF	
class **label**	male	female	male	female	male	female
male	8	4	9	3	8	4
female	5	8	0	13	0	13

6 Conclusions

We developed a system using statistical data analysis methods for feature extraction and neural networks for classification of facial expressions and gender. The shown results are promising, but for deploying these methods on our mobile robot under real world conditions, we have to solve the problem of finding the eyes and the nose tip of the user more robustly.

References

[1] Andreas Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *ICCV*, pages 368–373, 1995.

[2] M.S. Bartlett. *Face image analysis by unsupervised learning.* Kluwer Academic Publishers, 2001.

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

[4] W.V. Ekman, P. und Friesen. *Unmasking the face. A guide to recognizing emotions from facial clues.* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[5] W.V. Ekman, P. und Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.

[6] A. Essa, I. und Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997.

[7] Fröba, B. and Küblbeck C. Face detection and tracking using edge orientation information. *SPIE Visual Communications and Image Processing*, pages 583–594, 2001.

[8] E. Hyvärinen, A. und Oja. Independent component analysis: A tutorial. *Neural Networks*, 1999.

[9] J. Hyvärinen, A.und Karhunen. *Independent Component Analysis.* John Wiley& Son, Inc., 2001.

[10] Kanade, T., Cohn, J.F., and Tian, Y. Comprehensive database for facial expression analysis. *In Proc. of the Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG'00), Grenoble*, pages 46–53, 2000.

[11] L. Wiskott, J. M. Fellous, N. Kruger, and C. v. d. Malsburg. Face recognition and gender determination. In *Proc. Int. Workshop on Automatic Face and Gesture Recognition, Zurich*, pages 26–28, 1995.

[12] Lyons, M.J. and Budynek, J. Automatic classification of single facial image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

[13] G.W. Padgett, C. und Cottrell. Representing face image for emotion classification. *Advances in Neural Information Processing Systems*, 9:894–900, 1997.