C. Frank and E. Nöth, *CAIP 2003*, Springer-Verlag, (pp. 646–654).

Optimizing Eigenfaces by Face Masks for Facial Expression Recognition

Carmen Frank and Elmar Nöth

Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Martensstraße 3, 91058 Erlangen, Germany,
frank@informatik.uni-erlangen.de,noeth@informatik.uni-erlangen.de,
http://www5.informatik.uni-erlangen.de

Abstract. A new direction in improving modern dialogue systems is to make a human-machine dialogue more similar to a human-human dialogue. This can be done by adding more input modalities. One additional modality for automatic dialogue systems is the facial expression of the human user. A common problem in a human-machine dialogue where the angry face may give a clue is the recurrent misunderstanding of the user by the system. Or an helpless face may indicate a naive user who does not know how to utilize the system and should be led through the dialogue step by step.
This paper describes recognizing facial expressions in frontal images using eigenspaces. For the classification of facial expressions, rather than using the face whole image we classify regions which do not differ between subjects and at the same time are meaningful for facial expressions.
Important regions change when projecting the same face to eigenspaces trained with examples of different facial expressions. The average of different faces showing different facial expressions forms a face mask. This face mask fades out unnecessary or mistakable regions and emphasizes regions changing between facial expressions.
Using this face mask for training and classification of *neutral* and *angry* expressions of the face, we achieved an improvement of up to 5% points. The proposed method may improve other classification problems that use eigenspace methods as well.

1 Introduction

Dialogue systems nowadays are constructed to be used by a normal human being, i.e. a naive user. Neither are these users familiar with "drag and drop" nor do they want to read thick manuals about a lot of unnecessary functionality. Rather modern dialogue systems try to behave similar to a human-human dialogue in order to be used by such naive users. But what does a human-human dialogue looks like?

A human being uses much more input information than the spoken words during a conversation with another human being: the ears to hear the words and the tone of the voice, the eyes to recognize movements of the body and facial muscles, the nose to smell where somebody has been, and the skin to recognize physical contact. In the following we will concentrate on facial expressions. Facial expressions are not only emotional states of a user but also internal states affecting his interaction with a dialogue system, e.g. helplessness or irritation.

At the moment, there are several approaches to enhance modern dialogue systems. The dialogue system *SmartKom* introduced in [Wah01] which is funded by the BMBF[1] is also one of the new powerful dialogue systems. It is a multimodal multimedial system which uses speech, gesture and facial expression as input channels for a human-machine dialogue. The output is a combination of images, animation and speech synthesis.
One idea of facial expression recognition is to get as soon as possible a hint for an angry user in order to modify the dialogue strategies of the system and to give more support. This prevents the users from getting disappointed up to such an extent that they would never ever use the system again.
If a system wants to know about the users internal state by observing the face, it first has to localize the face and then recognize the facial expression.

Face localization aims to determine the image position of a single face. The literature shows various methods. In [Cha98] a combination of skin color and luminance is used to find the face in an head-shoulder image. A combination of facial components (like eyes and nostrils) found by SVMs and their geometric relation is used in [Hei00]. They used this method to detect faces in frontal and near-frontal views of still grey level images. A probabilistic face detection method for faces of different pose, with different expression and under different lighting conditions is the mixture of factor analyzers used by [Yan99]. Only color information is used by [Jon99] to form a statistical model for person detection in web images.
The task of facial expression recognition is to determine the emotional state of a person. A common method is to identify facial action units (AU). These AU were defined by Paul Ekman in [Ekm78]. In [Tia01] a neural-network is used to recognize AU from the coordinates of facial features like lip corners or the curve of eye brows. To determine the muscle movement from the optical flow when showing facial expressions is the task in [Ess95]. It is supplemented by temporal information to form a spatial-temporal motion energy model which can be compared to different models for the facial expressions.

In this paper we only deal with the second part, the analysis of an already found face.

2 Algorithm

In the method proposed by us, only pixels that are significant for facial expressions are used to create an eigenspace for facial expression recognition. These

[1] This research is being supported by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents of this study lies with the authors.

significant pixels are selected automatically by a training set of face images showing facial expressions. There is no assumption about the spatial relation of these pixels in contrast to [Kir90] where only an oval region of the face is used to omit background and hair. First we give a short introduction to standard eigenspaces. Then we show their disadvantages and introduce our face mask as improvement.

2.1 Introduction to Eigenspaces

Eigenspace methods are well known in the topic of face recognition ([Tur91], [Yam00], [Mog94]). In a standard face recognition system, one eigenspace for each person is created using different images of this person. Later, when classifying a photo of an unknown person, this image is projected using each of the eigenspaces. The reconstruction error of the principal component representation is an effective indicator of a match.

To create an eigenspace with training images a partial Karhunen-Loéve transformation, also called principal component analysis (PCA) is used. It is a dimensionality reduction scheme that maximizes the scatter of all projected samples, using N sample images of a person $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ taking values in an n-dimensional feature space. Let $\boldsymbol{\mu}$ be the mean image of all feature vectors. The total scatter matrix is then defined as

$$S_T = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{\mu})(\boldsymbol{x}_k - \boldsymbol{\mu})^T \tag{1}$$

In PCA, the optimal projection W_{opt} to a lower dimensional subspace is chosen to maximize the determinant of the total scatter matrix of the projected samples,

$$W_{opt} = arg \max_{W} |W^T S_T W| = [w_1, w_2, \ldots, w_m] \tag{2}$$

where $\{w_i | i = 1, 2, \ldots m\}$ is the set of n-dimensional eigenvectors of S_T corresponding to the set of decreasing eigenvalues. These eigenvectors have the same dimension as the input vectors and are referred to as Eigenfaces.

In the following sections we assume that high order eigenvectors correspond to high eigenvalues. Therefore high order eigenvectors hold more relevant information.

2.2 Disadvantages of standard Eigenspaces

An advantage and as well a disadvantage of eigenspace methods is their capability of finding the significant differences between the input samples. This feature enables eigenspace methods to model a given sample of a n-dimensional feature space in an optimal way using only a m-dimensional space.

But if one has significant differences between training samples not relevant for separating the classes, nevertheless they appear in the high order eigenvalues and maybe fudge the classifying result. An example for such differences of training samples is lighting. Training samples created under different lighting

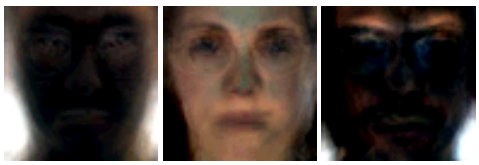

Fig. 1. The first three eigenvectors (eigenfaces) of an *anger*-eigenspace, which model face shape, lighting and eyebrows

conditions constitute an eigenspace which model the light in high order eigenvectors. In Figure 1 the first three eigenvectors (often called eigenfaces) from an *anger* eigenspace can be seen modeling light and face contour but not facial expressions. Therefore in face recognition often the first p eigenvectors are deleted as described in [Bel96].

2.3 Eigenfaces for Facial Expression Recognition

When using eigenfaces for facial expression recognition of unknown faces, one possibility is to calculate one eigenspace for each facial expression from a labeled database of different persons.
The classification procedure corresponds to that of face recognition: project a new image to each eigenspace and select the eigenspace which best describes the input image. This is accomplished by calculating the residual description error.

In addition to the disadvantage mentioned above, a problem for facial expression classification is that the person itself, whose facial expression should be classified, is unknown.
Each person uses a different smile. Each person has a different appearance of the neutral face. But each smile of each person should be classified as *smile*. And even facial expressions result from very subtle changes in the face and therefore do not show up in the high order eigenvectors.

2.4 Adapting Eigenfaces for Facial Expression Recognition

In order to deal with this fact we tried to eliminate parts of the face with a high level of changes between different persons which do not contribute to facial expressions. To find out which parts of the face are unnecessary for classifying facial expression, we also use an eigenspace approach.
Imagine we have a training set F_κ of l samples with similar characteristics for each class Ω_κ, $\kappa \in 1, \dots k$. Thus there is different illumination, different face shape etc. in each set F_κ. Reconstructing one image with each of our eigenspaces results in k different samples. The reconstructed images do not differ in characteristics like illumination, because this is modeled by each eigenspace. But they differ in facial expression specific regions, such as the mouth area.
So we can obtain a mask vector $\boldsymbol{m}$ as the average of difference images using a

Fig. 2. In the first row the original neutral face, the face reconstructed by a neutral and an anger eigenspace and the difference image of both is shown. In the second row an anger face was used.

training set T. For a two class problem this is done in the following way,

$$\boldsymbol{m} = \frac{1}{|T|} \sum_{\boldsymbol{y}_i \in T} V_1^T(\boldsymbol{y}_i - \boldsymbol{\mu}_1) - V_2^T(\boldsymbol{y}_i - \boldsymbol{\mu}_2) \tag{3}$$

where $|T|$ stands for the cardinality of set T and V_κ^T is the eigenspace for class κ . In Figure 2 the neutral and anger face of a man are projected in both eigenspace and the resulting difference images are shown. Before training an eigenspace, we now delete vector components (in this case pixels) from all training samples whose corresponding component of the mask vector $\boldsymbol{m}$ is smaller than a threshold θ. The threshold is selected heuristically at the moment.
The same components must be deleted from an image before classification. A positive side effect is the reduction of feature dimension. The face mask used for our experiments (see. Figure 3) eliminates about 50% of all pixels.

3 Data

All experiments described in this article are performed using the AR-Face Database [Mar98]. From this database we selected one image per person showing a neutral or angry facial expression. The included faces do not show full blown emotion, but natural, weak facial expression. This results in 264 images altogether. The whole set was split into 4 parts (equivalent to the cdroms of the database), 3 parts were used for training and one for testing in a leave one out method. No normalization was done. The tip of the nose, marked by a naive person, served as a reference point to cut the face from the whole image.

4 Experiments

4.1 Facial Expression Mask

The first task is to generate a mask which emphasizes regions of the face important for facial expressions and deletes other regions. We use a set of training images, to create one *anger*- and one *neutral*- eigenspace. The same set of images

Fig. 3. The left image is an average of faces projected to different eigenspaces. This image binarised with a threshold $\theta = 135$ can be seen on the right side. All *white* pixels are deleted before classification.

is used to create a mask using Equation 3.
This means in detail: project and reproject one image with each eigenspace, subtract the resulting images, calculate an average image over all difference images created from the training set.

Using a threshold θ the mask image is converted to a binary image. Preliminary experiments showed 135 to be a suitable value when using 1 byte of color information for each channel. Such a binarised mask with the corresponding average image can be seen in Figure 3.

4.2 Facial Expression Classification

The images used for the experiments have a size of 64×64 pixels. The classes used were *anger* and *neutral*.

To get an idea of the obtained improvement by the face mask, we show both ROC-curves in one figure. When using *rgb* information of an image the improvement is about 2% points for a medium false alarm rate; this can be seen in Figure 4. False alarm rate means the percentage of *neutral* faces which are classified as *anger*.

The improvement in Figure 5 is much clearer to see. Here we used only the intensity information of a given image. This is more reasonable, because the color of a face does not give a clue about the facial expression, except if the person is ashamed.

The reason why the improvement for the *rgb* case does not have these good values as for the grey value, may be the way the face mask is defined and used. There is only one face mask and not three, for each color channel a separate one. So a constant red value at one position means the green and blue channels are not used either.

5 Application

The knowledge about a users internal state is important to a modern personalized dialogue system. The possible internal user state not only include anger and happiness but also helplessness or confusing. An example application for giving useful information to an automatic dialogue system by analyzing facial expression is a dialogue about current television program. A happy face of a user when getting information about a thriller indicates an affectation for thriller. From

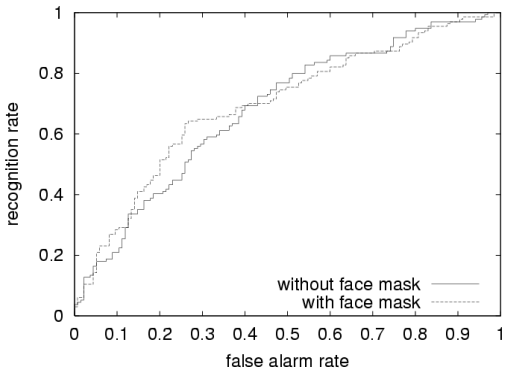

Fig. 4. The percentage of *neutral* faces classified as *anger* is shown on the x-axis (false alarm rate). On the y-axis the percentage of *anger* face classified as *angry* (recognition rate) is shown when using rgb information.

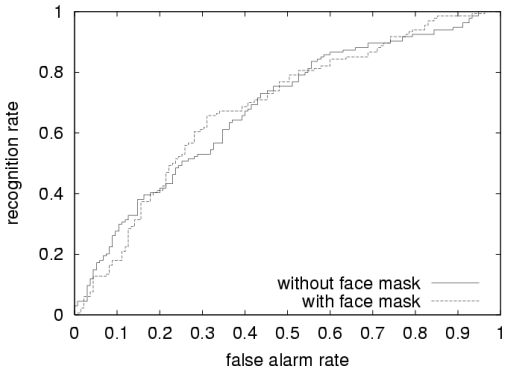

Fig. 5. Recognition rate compared to false alarm rate when using grey level images.

now on this user can be lead to a happy mood when thriller are presented to him first while information about other genres is presented afterwards.

Up to now there are no results from naive persons using a dialogue system which uses information about their emotional state. The reason for this is that on the one hand the users must not know about this functionality of the dialogue system in order to show natural behavior. On the other hand they should be familiar with the system because the dialogue must be as similar as possible to a human-human interaction.

In Wizard-of-Oz experiments the users seemed not to be confused by the facial camera, they forgot being filmed during the dialogue. The questionnaires which are filled out after each dialogue showed the users are not aware of facial

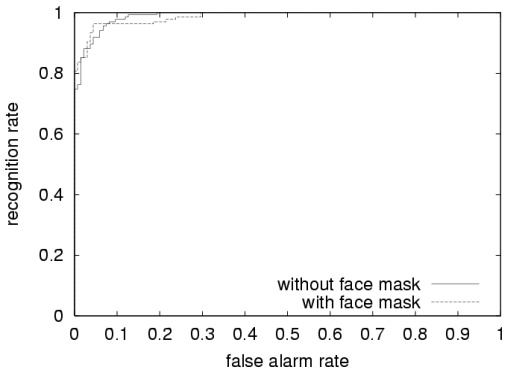

Fig. 6. This is the recognition rate compared to false alarm rate when using rgb-images for classifying *joy* vs. *angry* using a mask with a threshold of $\theta = 135$.

expressions influencing the dialogue. They are content with the system and would like to use it an other time.

6 Conclusion

Our experiments show that significant information for discriminating facial expressions can not be found in the high order eigenvectors of standard eigenspaces. Moreover fudging information is represented by the high order eigenvectors. This is avoided by using masked faces for the eigenspace training. The used facial mask is automatically trained from a set of faces showing different expressions. It emphasizes discriminating facial regions and fades out unnecessary or mistakable parts.
Using this face mask for training and classification of *neutral* and *angry* expressions of the face, we achieve an improvement of 5% when using grey level images. The described method for data selection to train eigenspaces for facial expression recognition may be used for other classification tasks by changing the training data for the mask.
The next steps for us will be to increase the recognition rates for the rgb case by a more detailed mask and the application of the mask method to the detection of faces using eigenspace methods.

7 Remarks

There are lots of differences between facial expressions. E.g.neither does each smile result from the same positive emotional state of a person nor does each person express a positive emotional state with the same smile. A smile may express love to someone else but a slightly different smile says 'I am sorry'.

The same is true for anger. And especially anger is an emotional state which is expressed in very different manners by different individuals. Some form wrinkles at the forehead, others nearly close their eyes, knit their eyebrows or press the

lips together.
But *angry* is besides *helplessness* the most important state for an automatic human-machine dialogue system a user can be in. The anger of a user gives a hint for dialogue problems which should be solved by the dialogue system. Of course it would be nice to know that a user is happy and satisfied with the system but in this case no system reaction is necessary.

The *angry* and *neutral* state of a person are those states which are most difficult to discriminate. A human person produced 50% false alarms and 95% recognition rate when classifying our samples. The reason for this high false alarm rate is that, as mentioned above anger is expressed in very different ways and people often hide anger. The classification of facial expressions of a familiar person is much easier for the human as well as for an automatic system.

The recognition rates for the classification of *angry* and *joyful* user states are much higher than for *angry* and *neutral*. They are shown in Figure 6. A human reaches 97% recognition rate with no false alarms.

References

[Bel96] Belhumeur, P.; Hespanha, J.; Kriegman, D.: *Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection*, in *European Conference on Computer Vision '96*, 1996, S. 45–58.

[Cha98] Chai, D.; Ngan, K. N.: *Locating Facial Regions of a Head–and–Shoulders Color Image*, in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 1998, S. 124–129.

[Ekm78] Ekman, P.; Friesen, W.: *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*, in *Consulting Psychologists Press, Palo Alto, CA*, 1978.

[Ess95] Essa, I.; Pentland, A.: *Facial Expression Recognition Using a Dynamic Model and Motion Energy*, in *Proceedings of the Fifth International Conference on Computer Vision*, 1995, S. 360–367.

[Hei00] Heisele, B.; Poggio, T.; Pontil, M.: *Face Detection in Still Gray Images*, in *MIT AI Memo, AIM–1687*, 2000.

[Jon99] Jones, M.; Rehg, J.: *Statistical Color Models with Application to Skin Detection*, in *Proceedings of Computer Vision and Pattern Recognition*, 1999, S. I:274–280.

[Kir90] Kirby, M.; Sirovich, L.: *Application of the Karhunen-Loève Procedure for the Characterization of Human Faces*, *TPAMI*, Bd. 12, Nr. 1, 1990, S. 103–108.

[Mar98] Martinez, A.; Benavente., R.: *The AR Face Database*, Purdue University, West Lafayette, IN 47907-1285, 1998.

[Mog94] Moghaddam, B.; Pentland, A.: *Face Recognition Using View–Based and Modular Eigenspaces*, in *Vismod, TR-301*, 1994.

[Tia01] Tian, Y.; Kanade, T.; Cohn, J.: *Recognizing Action Units for Facial Expression Analysis*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Bd. 23, Nr. 2, 2001, S. 97–115.

[Tur91] Turk, M.; Pentland, A.: *Face Recognition Using Eigenfaces*, in *Proceedings of Computer Vision and Pattern Recognition*, 1991, S. 586–591.

[Wah01] Wahlster, W.; Reithinger, N.; Blocher, A.: *SmartKom: Multimodal Communication with a Life–Like Character*, in *Eurospeech 2001*, 2001, S. 1547–1550.

[Yam00] Yambor, W. S.; Draper, B. A.; Beveridge, J. R.: *Analyzing PCA–based Face Recognition Algorithms: Eigenvector Selection and Distance Measures*, in *Second Workshop on Empirical Evaluation Methods in Computer Vision*, 2000.

[Yan99] Yang, M.; Ahuja, M.; Kriegman, D.: *Face Detection using a Mixture of Factor Analyzers*, in *Proceedings of the International Conference on Image Processing*, Bd. 3, 1999, S. 612–616.