ESTIMATION OF THE NEUTRAL FACE SHAPE USING GAUSSIAN MIXTURE MODELS

Sezer Ulukaya[1,2], *Çiğdem Eroğlu Erdem*[2] *

[1] Boğaziçi University - [2] Bahçeşehir University
Dept. Electrical and Electronics Engineering - Dept. Electrical and Electronics Engineering
sezer.ulukaya@bahcesehir.edu.tr - cigdem.eroglu@bahcesehir.edu.tr

ABSTRACT

We present a Gaussian Mixture Model (GMM) fitting method for estimating the unknown neutral face shape for frontal facial expression recognition using geometrical features. Subtracting the estimated neutral face, which is related to the identity-specific component of the shape leaves us with the component related to the variations resulting from facial expressions. Experimental results on the Extended Cohn-Kanade (CK+) database show that subtracting the estimated neutral face shape gives better emotion recognition rates as compared to classifying the geometrical facial features directly, when the person-specific neutral face shape is not available. We also experimentally evaluate two different geometric facial feature extraction methods for emotion recognition. The average emotion recognition rates achieved with the proposed neutral shape estimation method and coordinate based features is 88%, which is higher than the baseline results presented in the literature, although we do not use the person-specific neutral shapes (94% if we use), and any appearance based features.

Index Terms— neutral face estimation, gaussian mixture models, facial expression recognition

1. INTRODUCTION

Automatic recognition of facial expressions is a challenging task that has attracted a lot of interest in recent years. It is envisioned that the capability of recognizing human emotions will be a part of man-machine interactions and ubiquitous computing scenarios in the future [1]. Facial expression recognition has many other application areas including security [2], driver safety [3] and health-care [4].

1.1. Previous Work

A major problem in classifying facial expressions is defining the emotion classes. A solution to this has been suggested by Ekman [5], who specified a universally displayed set of six emotions: anger, happiness, disgust, fear, sadness and surprise. Another solution, which has gained popularity recently is to use dimensional and continuous labeling of the affective cues in the valence, activation and dominance coordinates [6].

Many studies have been published on affect recognition from facial expressions in the last decade, which are summarized in recent survey papers [7, 6, 8, 9]. Most of these methods use two dimensional spatio-temporal facial features, which are fed to a pattern recognition algorithm. Facial features can be categorized as geometrical features and appearance based features. Geometrical features consist of shapes of facial components (eyes, lips etc.) and salient points on the face (nose tip etc.). Appearance based features provide information about the texture of the face as well (natural wrinkles and creases between the eyes etc.). It is expected that methods that use both geometrical and appearance features give more accurate results [9].

*This work was supported by Turkish Scientific and Technical Research Council (TUBITAK) under project EEAG-110E056.

1.2. Contributions and Outline of the Paper

When we describe a facial expression using locations of a set of points on the face, these geometric locations encode two types of information. The first type information is the identity-specific information, which is constant for that person. The second information is a variable part, which depends on pose and facial expressions. The identity-specific component can be eliminated by subtracting the features obtained from a neutral facial expression of that person from the current frame, which may be the first frame of a video clip [10] as in CK+ database. However, neutral face information of that person may not always be available. In that case, researchers generally average the features of a certain number of images in the video clip, assuming that averaging will resemble a neutral facial expression [11]. However, this assumption is not always true, depending on the content of the video clip.

This paper has two contributions: i) We present a Gaussian Mixture Model (GMM) based method for estimating the neutral face shape for frontal facial expression recognition using geometrical features, when the person-specific neutral face shape is not available. We experimentally show that subtracting the estimated neutral face shape gives better affect recognition rates as compared to classifying the geometrical facial features directly, when the person-specific neutral expression is not available. ii) We also experimentally evaluate two different geometric features, which we call the *coordinate based features* (CBF) [10] and *distance and angle based features* (DABF) [12]. CBF features have been observed to give higher emotion recognition rates on the CK+ database.

In Section 2, a more detailed description of the used geometrical features are described. In Section 3, the details of the GMM-based estimation of the neutral face shape are given. Experimental results are presented in Section 4, which are followed by conclusions in Section 5.

2. GEOMETRICAL FACIAL FEATURES

In this paper, we utilize the face tracking data provided in the CK+ database to form two types of geometrical facial features as described below. The CK+ database provides the locations of 68 points on the face at each frame, which are tracked using Active Appearance Models [10, 13].There are 123 subjects and 327 emotion labeled image sequences in the CK+ database belonging to the six

basic emotions. Image sequences start with a neutral (onset) frame and end with a peak frame (apex) of the expression (see Fig 1).

First frame Fifth frame Last (33^{rd}) frame

Fig. 1. AAM based tracking of landmark coordinates in CK+ database (©J. Cohn)

We need to align the face shapes described by the tracked landmark points for all frames in the database to eliminate any rotation, translation and scale effects that may exist between subjects and/or within a video clip.

Alignment of the face shapes for all frames of the CK+ database is carried out using the landmark points that are affected the least from the facial expressions such as the nose tip and the inner corners of the eyes. The inner corners of the eyes (points 40, 43 in Figure 2) are not affected much from facial expressions and they are robust to track [12]. First, we move the nose tip to the origin (point 31 in Figure 2). In order to compensate for in-plane head rotations, all the landmarks are rotated such that the line connecting the inner corners of the eye becomes horizontal (i.e., parallel to the x-axis). Another set of points that are expected to be affected from facial expressions the least are the landmarks located at the outer borders of the cheeks (points 1, 2, and 16, 17 in Figure 2). In order to compensate for any scale differences between frames, we scale the landmarks coordinates such that the sum of distances between three point pairs is constant:

$$d(p^1_{n,i}, p^{17}_{n,i}) + d(p^2_{n,i}, p^{16}_{n,i}) + d(p^{40}_{n,i}, p^{43}_{n,i}) = \alpha, \qquad (1)$$

where $p^j_{n,i} = [x^j_{n,i}, y^j_{n,i}]$, $k = 1, \ldots, M$ denotes the vector representing the j^{th} landmark point in the i^{th} frame of the n^{th} image sequence, and $M = 68$. The operator $d(.,.)$ denotes the Euclidean distance between two landmarks. The constant was chosen as $\alpha = 10$ during the experiments to remove scale differences and zoom factor.

The **coordinate based features (CBF)** consist of the x and y coordinates of the M aligned landmarks points in the last (peak) frame of an image sequence (CBF). When the landmarks points of the person-specific neutral facial expression are available (which is the first frame in CK+ database), they can be subtracted from the peak frame, and will be referred to as **coordinate based features with neutral subtraction (CBF-NS)**.

Another set of geometrical features that we evaluate are derived from the CBF features and they consist of distances and angles between certain landmark points as described below. **We call them distance and angle based features (DABF)**. A total of 20 features (f_1 - f_{20}) are obtained from the last frame of an image sequence as follows [12]. When the person-specific neutral face shape is available, we can subtract the 20 DABF features of the first frame from the peak frame to obtain another set of features that we call as **DBAF-NS** features, where NS stands for "neutral subtraction".

Fig. 2. The 68 landmark points tracked on the face as given the CK+ database (©J. Cohn).

3. ESTIMATION OF THE NEUTRAL FACE SHAPE USING GAUSSIAN MIXTURE MODELS

Neutral face shapes of people in a population are quite different from each other. Some people have long and thin faces while others have round faces. Therefore, we first aim to identify typical face shapes in the population, by fitting a Gaussian Mixture Model to the shape features of neutral faces. We expect the mean vectors of each Gaussian component to represent a typical face shape cluster.

3.1. Fitting a Gaussian Mixture Model to Neutral Face Shapes

The data set of neutral face shapes is constructed from the first frames of all image sequences that are provided in the CK+ database (593 sequences in total) that belong to 123 subjects. Let us represent our neutral shape data set as: $\chi = \{\mathbf{s}_{n,1}\}, n = 1, \ldots, N$, where $\mathbf{s}_{n,1} = [p^1_{n,1}, p^2_{n,1}, \ldots, p^M_{n,1}]$, represent the face shape in the first frame if image sequence n, based on the normalized coordinates of 68 landmark points. Here the parameters are $M = 68, N = 593$.

We want to model the distribution of neutral face shapes using a *mixture of densities* as follows:

$$p(\mathbf{s}) = \sum_{k=1}^{K} p(\mathbf{s}|G_k) P(G_k), \qquad (2)$$

where G_k are the *mixture components*, which are also called *clusters*. $p(\mathbf{s}|G_k)$ are the *component densities* and $P(G_k)$ are the mixture proportions (mixing coefficients). The number of components K is either specified beforehand or can be estimated using Akaike's information criterion as described below. If the component densities are multivariate Gaussian, we have $p(\mathbf{s}|G_k) \sim \aleph(\mathbf{s}|\mu_k, \Sigma_k)$ and $\Phi = \{P(G_k), \mu_k, \Sigma_k\}_{k=1}^K$ are the parameters that should be estimated from the data set $\chi = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$. We look for component density parameters that maximize the likelihood of the data set (sample). The likelihood of the sample assuming that the data points are drawn independently from the distribution is:

$$\begin{aligned} p(\chi|\Phi) &= \prod_{n=1}^{N} p(\mathbf{s}_n|\Phi) \qquad (3) \\ &= \prod_{n=1}^{N} \left(\sum_{k=1}^{K} P(G_k) \aleph(\mathbf{s}_n|\mu_k, \Sigma_k) \right), \end{aligned}$$

and the log likelihood of the data set is given by:

$$\ln p(\chi|\Phi) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} P(G_k) \aleph(\mathbf{s}_n|\mu_k, \Sigma_k) \right). \quad (4)$$

The log likelihood function given in (4) is maximized using the Expectation-Maximization algorithm [14].

The parameter K can be determined experimentally using Akaike's information criterion [15]. It is often used to determine an appropriate number of mixture components when the number of components is unspecified. Akaike information is the negative log-likelihood for the data with a penalty term for the number of estimated parameters:

$$AIC = 2m - 2L_m, \quad (5)$$

where m is the number of parameters in the statistical model and L_m is the maximized value of the log likelihood function. The GMM fitting process is carried out for a range of K values, and the value that maximizes the AIC is selected.

After fitting a Gaussian Mixture Model to the data set of neutral face shapes, the mean vectors $\mu_k, k = 1, \ldots, K$ of the K Gaussian mixture components will represent the typical neutral face shapes in the population.

3.2. Estimation of the Neutral Face Shape for a Facial Expression

Given a shape vector s_i extracted from an image with a facial expression, we assume that it can be decomposed as follows:

$$\mathbf{s}_{n,i} = \hat{\mathbf{s}}_{n,i} + \mathbf{v}_{n,i}, \quad (6)$$

where $\hat{\mathbf{s}}_{n,i}$ represents the person-specific part of the shape and $\mathbf{v}_{n,i}$ represents the variable part of the shape due to pose and facial expression, which are mostly related to the emotional state of the subject. If the neutral face shape of that person is available, it can be subtracted from $\mathbf{s}_{n,i}$, to give the variable part of the shape, which can then be classified.

However, if the person-specific neutral face is not available, it is beneficial in terms of increasing the correct classification rate to estimate the 'best' fitting neutral face shape and subtract it from $\mathbf{s}_{n,i}$. In order to select the best fitting neutral shape among the K face shapes which were estimated using GMM fitting, we use the landmarks that are not affected from facial expressions much. The point set selected for this purpose consist of the left and right sides of the cheeks and the inner corners of the eye: $\{p^1_{n,i}, p^2_{n,i}, p^3_{n,i}, p^{15}_{n,i}, p^{16}_{n,i}, p^{17}_{n,i}, p^{40}_{n,i}, p^{43}_{n,i}\}$.

Let us relabel the above points for the i^{th} frame of sequence n as $\{P^j_{n,i}\}$ and let us denote the corresponding points in the mean vector of the k^{th} Gaussian mixture component as $\{\hat{\mu}^j_k\}$, where $j = 1, \ldots, 8$ and $k = 1, \ldots, K$. In order to select the best fitting neutral shape we minimize the following Mahalanobis distance:

$$D_k(P^j_{n,i}, \hat{\mu}^j_k) = \sqrt{(P^j_{n,i} - \hat{\mu}^j_k)^T \hat{\Sigma}^{-1}_k (P^j_{n,i} - \hat{\mu}^j_k)}, \quad (7)$$

where $\hat{\Sigma}_k$ is the 16×16 covariance matrix for the x and y coordinates of the landmark points $1, 2, 3, 15, 16, 17, 40, 43$, and is formed from the full covariance matrix Σ_k, which is 136×136. The index of the best fitting neutral shape is:

$$k^*_{n,i} = \arg\min_k D_k(P^j_{n,i}, \hat{\mu}^j_k) \quad (8)$$

After the index of the best fitting neutral face is estimated, the mean shape corresponding to that Gaussian mixture is assigned to the person-specific component in (6):

$$\hat{\mathbf{s}}_{n,i} \simeq \hat{\mu}_k. \quad (9)$$

Hence, the variable part of shape due to the facial expression can be approximated as:

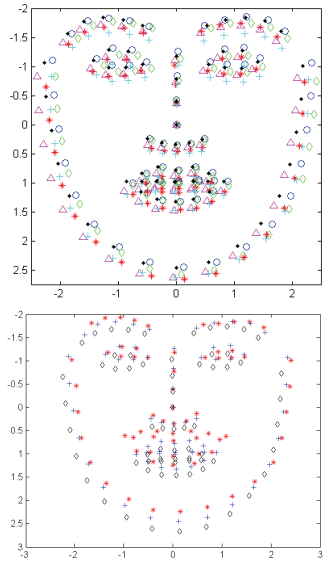
$$\mathbf{s}_{n,i} - \hat{\mu}_k \simeq \mathbf{v}_{n,i}, \quad (10)$$

which is classified using a support vector classifier (SVC) with a polynomial kernel.

4. EXPERIMENTAL RESULTS

Experiments are done on the CK+ database [10]. The Gaussian Mixture fitting to the neutral face shapes is carried out using the first frames of all sequences for various values of K, and $K = 6$, which gives the minimum AIC value is selected. During GMM fitting, we used a small non-negative regularization number added to the diagonal of covariance matrices to make them positive-definite. The mean shapes of the estimated Gaussian mixtures for $K = 6$ is shown in Figure 3 (a). We can observe that the estimated mean shape vectors reflect the person-specific variations of the face shape in the population.

For comparison purposes, we show the CBF features (red *), the best fitting neutral shape (blue +) and the worst fitting neutral face shape (gray diamond) in Figure 3(b) for subject 106. We can see that the best fitting neutral face shape follows the person specific characteristics of the face better than the worst fitting neutral shape, especially if we observe the landmarks around the inner corners of the eyes and the sides of the face. Hence, we can say that the proposed algorithm is successful in estimating a reasonable neutral face shape based on the GMM of the population.



(a)

(b)

Fig. 3. (a) The means of estimated Gaussian mixtures neutral face shapes for $K = 6$. Each mean vector is shown with a different marker. (b) The happy face shape (red *), the best fitting neutral face shape (blue +) and the worst fitting neutral face shape (gray diamond) for subject 106.

We compare the two geometric feature extraction methods (CBF and DABF) under different neutral face shape estimation scenarios. A Support Vector Classifier with a second order polynomial kernel [15] is used to classify the facial features. In order to maximize the training set and to guarantee subject-independence, we use a leave-one-subject-our cross validation scheme. As can be seen in 1, the recognition rate for the proposed CBF-ENS features (88 %) is higher than the CBF features (83%), which shows that estimating the neutral face shape and subtracting it from the shape under test is beneficial. Our results are also better than the baseline method given in [10], although we do not use any appearance based features.

The highest recognition rate is achieved for the CBF-NS (94%) features as expected, since person-specific neutral face information is used. We can observe that the recognition rates achieved with DABF features are lower than CBF features. However, the proposed neutral face shape estimation method is also beneficial for this feature set, since the recognition rate of DABF-ENS (74 %) is higher than the recognition rate of DABF features (69 %).

We also tried the k-means, and using the average of a video clip to estimate the neutral face shape, but they were not better than the proposed GMM based approach in terms of increasing the emotion recognition rate.

Method / Feature Used	Average Recognition Rate
CBF	83 %
CBF-NS	94 %
CBF-ENS	88 %
DABF	69 %
DABF-NS	77 %
DABF-ENS	74 %
Baseline Method [10]	83 %

Table 1. The average emotion recognition rates for the six compared feature sets using a SVC with a second order polynomial kernel. The proposed CBF-ENS features gives higher recognition rates as compared to the CBF features and the baseline method. **CBF**: Coordinate based features, without neutral shape subtraction. **CBF-NS**: Coordinate based features with subtraction of person-specific neutral shape. **CBF-ENS**: Coordinate based features with subtraction of the estimated neutral shape. **DABF**: The distance and angle based features, without neutral shape subtraction. **DABF-NS**: The DABF after subtracting the person-specific neutral shape. **DABF-ENS**: The DABF after subtracting the features calculated from the estimated neutral shape.

5. CONCLUSIONS

We presented a Gaussian Mixture Model (GMM) fitting method for estimating the unknown neutral face shape for frontal facial expression recognition using geometrical features.

Experimental results on the CK+ database [10], show that estimating the neutral face shape and subtracting it from the landmarks of the test frame is beneficial for increasing the average emotion recognition rate. The average emotion recognition rates achieved with the proposed neutral shape estimation method and coordinate based features is 88%, which is higher than the baseline results presented in [10], although we do not use the person-specific neutral shapes, and any appearance based features. If we use person-specific neutral face shapes, the recognition rate increases to 94%.

We also observed that coordinate based features [10] perform better than distance and angle based features [12] for the emotion recognition task.

6. REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] A. Ryan, J. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, and A. Rossi, "Automated facial expression recognition system," in *Proceedings of the International Carnahan Conference on Security Technology*, 2009, pp. 172–177.

[3] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Automated drowsiness detection for improved driving safety," in *Proceedings of the International Conference on Automotive Technologies*, 2008.

[4] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face - pain expression recognition using active appearance models," *Image and Vision Computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[5] P. Ekman and W. V. Friesen, "Constants accross cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[6] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.

[7] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[8] M. Pantic, "Machine analysis of facial behaviour: naturalistic and dynamic behaviour," *Philosophical Transactions of the Royal Society B-Biological Sciences*, vol. 364, no. 1535, pp. 3505–3513, 2009.

[9] Z. H. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[10] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proceedings of IEEE workshop on CVPR for Human Communicative Behavior Analysis*, San Francisco, USA, 2010.

[11] R. Gajsek, V. Struc, and F. Mihelic, "Multi-modal emotion recognition using canonical correlations and acoustic features," in *International Conf. Pattern Recognition (ICPR)*, 2010.

[12] J. Jiao and M. Pantic, "Implicit image tagging via facial information," in *ACM Multimedia, Workshop on Social Signal Processing (SSPW'10)*, Firenze, Italy, 2010, pp. 59–64.

[13] T. f. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[15] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716 – 723, 1974.