An Efficient and Robust Facial Expression Recognition System

Juliano J. Bazzo, Marcus V. Lamar, and Rui Seara

Abstract—**This paper describes a new pre-processing strategy to classify facial expression. Previous research works suggest that Gabor wavelets applied to recognize facial expression images subtracted from the neutral face of the same subject could provide a satisfactory recognition rate under controlled conditions, such as eye and mouth alignments. Here, we propose a recognition system in which the Gabor kernels are applied to a facial expression subtracted from an averaged neutral face. A pre-processing stage which generates a reduced output data set is also proposed. By using an artificial neural network-based classifier, recognition rates of 86.6% and 81.6% are obtained for 7 upper and 7 lower face actions, respectively. Considering a heterogeneous subject database with head motion and lighting variations, the new recognition approach is assessed for performance. The obtained results attest the effectiveness and applicability of the proposed technique for recognizing facial actions.**

***Index Terms*—Face recognition, facial action units, Gabor wavelets, artificial neural network.**

I. Introduction

In expression recognition applications very good results have been attained to classify emotions, such as anger and fear [1], [2]. Furthermore, the facial action coding system (FACS) developed by Ekman and Friesen [3] has become an objective model to measure facial activity for behavioral science investigation of the face. FACS defines 46 action units (AUs) corresponding to each independent face motion. FACS has been used to demonstrate difference between genuine and simulated pain [4], differences between when people are telling the truth or lying [5], and differences between the facial signals of suicidal and nonsuicidally depressed patients [6]. The development of a computational system able to recognize single AUs and their combinations is still a very challenging problem [7], [8]. In the telecommunication area, FACS can be used in very low bit rate coders, such as MPEG-4. In this application, the current face image in a video signal can be synthesized at the receiver based only on the parameters of the transmitting face, modifying the facial model according to emotion and lip motions, thus reducing significantly the needed bit rate.

Tian *et. al.* [8] have classified several upper face AUs using Gabor wavelets and geometric features. This experiment was applied to the Cohn-Kanade database [9] in which the images were recorded considering heterogeneous subjects. Forty Gabor filters on images of increasing complexity have been used, and the coefficients from 20 manually adjusted fiducial points have been extracted. Zang *et. al.* [1] have applied 18 Gabor filters, and the geometric feature from 34 manually selected fiducial points have been obtained. A two-layer perceptron neural network has been used to distinguish 7 emotion expressions. Donato *et al.* [7] have considered independently several techniques (optical flow, principal component analysis (PCA), independent component analysis (ICA), local feature analysis (LFA), and Gabor wavelets) for recognizing 6 single upper face AUs and 6 lower face AUs from 24 subjects. Careful face alignments, rotation of eyes, scaling, cropping, contrast enhancement, and histogram equalization have been performed in a pre-processing stage.

The previous presented approaches have suggested systems that require a large amount of pre-processing to align the face, choose the interest points or even to crop, rotate and scale the face to fit it in a specific template. Generally, the recognition rate can be increased if some restrictions and pre-processing are performed: (*i*) a homogeneous subject database is chosen, such as Japanese or Euro-American subjects; (*ii*) head motion is excluded; (*iii*) frames are cropped and aligned; (*iv*) some AU combinations which may change the structure of other AUs are excluded; (*v*) very different AUs are considered for testing.

In our work a heterogeneous subject database is considered. It includes European, African and Asian ancestry subjects, lightning variations, no accurate frame alignment, sequences containing head motion, combinations of AUs, and similar expressions. We propose a recognition system with an innovative pre-processing. The feature extraction technique is based on Gabor wavelets followed by a principal component analysis (PCA), the latter leading to a reduced output data set. This feature extraction procedure is applied to the averaged neutral face image difference.

II. Expression Recognition System

Figure 1 shows the structure of the proposed system.

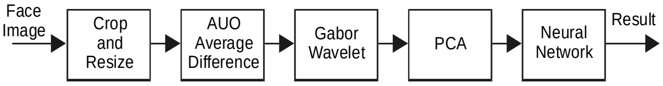


Fig. 1. Block diagram of the proposed system.

Juliano João Bazzo and Marcus Vinícius Lamar, Department of Electrical Engineering, Federal University of Paraná, Curitiba, Brazil, E-mails: julianojoaobazzo@yahoo.com.br, lamar@eletrica.ufpr.br.

Rui Seara, LINSE – Circuits and Signal Processing Laboratory, Department of Electrical Engineering, Federal University of Santa Catarina, Florianópolis, Brazil, E-mail: seara@linse.ufsc.br.

This work was supported in part by the Araucária Foundation (615/007/03), LACTEC, and Brazilian National Research Council (CNPq).

Following the previous diagram, a frontal face input image is roughly cropped generating the upper and lower faces. The average sizes of cropped images are 105×190 pixels for the upper face and 133×192 pixels for the lower face. Each image extracted is normalized to 30×45 pixels. No rotation or accurate eye and mouth alignments are performed. The upper face images are then subtracted from the average of upper neutral faces AU 0, considering the whole database. Similarly, the lower face is also subtracted from the average of lower neutral faces. The image difference is then filtered by a set of Gabor wavelets and the result is decomposed by PCA. The principal component vector feeds an artificial neural network, which performs the classification of the facial actions. The upper and lower faces are treated separately with the same unit processing.

Simulation results are obtained by using the Cohn-Kanade facial expression video database [9]. Ninety-four heterogeneous subjects with ages of 18 to 50 years have been selected. The image sequences used in the experiments contain small amounts of in-plane and limited out-of-plane motions. The images are digitized into 640×480 pixels with 8-bit gray scale.

From all subjects, we select 229 facial actions of upper faces which are divided into 7 different upper face action patterns. Table 1 shows the pattern number, cropped and resized upper face image samples, AUs present in the pattern performed by the subject, and the number of pattern samples selected from the database.

Similarly, we select 282 facial actions of lower face that are divided into 7 different lower face action patterns, according to Table 2.

TABLE 1
UPPER FACE ACTION PATTERNS

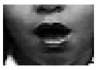
Pattern	Sample Image	Action Units	Number of Samples
1		0	70
2		1+2	26
3		1+2+5	43
4		4+7	14
5		4+6+7+9	12
6		4+7+9	20
7		6	44

A. *Average Neutral Face Difference*

The advantages of using the difference of the neutral face in face action and emotion recognition include robustness to illumination changes, removal of skin variations between subjects, and emphasis of the dynamic aspects of the image sequence [10].

Donato *et. al*. [7] have proposed the use of the difference between the last frame of a video sequence, in which the facial action is presented in its maximum intensity, and all the previous frames of the sequence, starting with the neutral frame. They demonstrated that such an approach can improve the performance of the recognition system. For such, it is necessary to perform a perfect aligning of the eyes and mouth for each frame of the sequence to adjust them in the same position, so that the subtraction technique becomes efficient.

TABLE 2
LOWER FACE ACTION PATTERNS

Pattern	Sample Image	Action Units	Number of Samples
1		0	55
2		25	39
3		26	49
4		27	45
5		12+25	53
6		15+17	24
7		20+25	17

Fasel and Luettin [11] have extracted the facial features by subtracting the maximum emotion intensity image from a sample of the neutral image of the subject. The database used has been recorded under controlled conditions, i.e. the face keeps a fixed position for the entire frame sequence without scale and rotation changes and/or lighting variations. They have also shown the importance of the use of the neutral face subtraction to characterize faces in expression recognition systems. However, in real-time systems, to estimate the neutral face of an unknown subject can be a very complex and time demanding task.

In this work we propose a new robust technique to subtract the image of maximum face action intensity from the neutral face average image (AU 0). Average images AU 0 for the upper and lower faces are estimated from overall available neutral face images of all subjects present in the database. This approach keeps the advantages of the neutral face subtraction technique while eliminating the previous analysis and the definition of the neutral face of a given unknown subject. This procedure reduces the computational complexity of the pre-processing stage. The neutral face average image can be obtained in the training phase of the recognition system.

In previous published approaches, a perfect face alignment has been needed so that the subtraction method can be efficient. In our approach an accurate alignment is no longer needed since the average neutral image contains the averaged information of non-aligned edges, thus making the system more robust to roughly cropped upper and lower images. Figure 2 shows the average neutral images for the upper and lower faces estimated from the Cohn-Kanade database.

(a) (b)

Fig. 2. Average neutral face images. (a) Upper face. (b) Lower face.

B. Gabor Wavelets

The proposed system uses a multi-scale set and multi-orientation Gabor wavelets extracted of the face images.

Let $I(\vec{x})$ represent an image, where $\vec{x}=(x,y)$ is the pixel coordinate, transform $\Gamma_i(\vec{x})$ is then defined as a two-dimensional convolution, given by

$$\Gamma_i(\vec{x}) = \int\int I(\vec{x}\,')\Psi_i(\vec{x}-\vec{x}\,')\,d\vec{x}\,', \qquad (1)$$

where the Gabor kernels $\Psi_i(\vec{x})$ are expressed as

$$\Psi_i(\vec{x}) = \frac{\left|\vec{k}_i\right|^2}{\sigma^2} e^{\frac{\left|\vec{k}_i\right|^2\left|\vec{x}\right|^2}{2\sigma^2}} \left(e^{j\vec{k}_i\vec{x}} - e^{-\frac{\sigma^2}{2}} \right). \qquad (2)$$

Each $\Psi_i(\vec{x})$ is a plane wave characterized by vector $\vec{k}_i$ enveloped by a Gaussian function with standard deviation σ. The center frequency of the *i-th* filter is given by the characteristic wave vector. Thus,

$$\vec{k}_i = \begin{bmatrix} k_{ix} \\ k_{iy} \end{bmatrix} = \begin{bmatrix} k_v \cos(\theta_\mu) \\ k_v \sin(\theta_\mu) \end{bmatrix}, \qquad (3)$$

where k_v and θ_μ characterize the scale and orientation, respectively. The first term in brackets in (2) determines the oscillatory part of the kernel and the second term compensates the DC value of the kernel. Thus, by subtracting the DC component, the Gabor filter becomes insensitive to the overall level of illumination.

In our implementation, we use $\sigma=\pi$ and 8 orientations given by $\theta_\mu = \mu\pi/8$ for $\mu = 1...8$. Two spatial frequencies $k_{v1}=\pi/2$ and $k_{v2}=\pi\sqrt{2}/4$ are considered.

The system has also been tested with lower frequencies and fewer orientations, but the recognition rate obtained has been poor.

C. Principal Component Analysis

The PCA produces an orthogonal basis from the filtered images, allowing a reduction of dimensionality by using the most representative components [12].

We perform separately the PCA of all 229 upper and 282 lower face images, cropped and resized for each Gabor kernel. Thirty principal components are selected from the projections of each set of filtered images.

D. Neural Network

For this stage, two 3-layer perceptron neural networks have been considered. The backpropagation algorithm is used for training both the upper and lower faces. The input layer of each network has dimension $30\times N$, where N is the number of of Gabor kernels. For each network, the hidden layer is composed of $\frac{2}{3}(\text{source nodes}+7)$ hidden neurons with sigmoid-type activation function. Finally, 7 output neurons are considered in each network, corresponding to 7 facial expression patterns for each upper and lower face expressions. The training set for the upper face consists of 371 processed image samples and the test set of 119 samples. For the lower face, the training set contains 287 image samples and the test set, 98 samples.

(a) (b)

Fig. 3. Examples. (a) Pattern 5: AU 4+6+7+9. (b) Pattern 6: AU 4+7+9.

The upper face pattern examples shown in Figure 3 indicate that it is a very hard task to distinguish one pattern from the other. AU 6, corresponding to a raised cheek, is very difficult to detect even to non-trained human beings when it appears jointly with other AUs. In order to improve the classification capability for these 2 patterns, a second ad-hoc neural network has been used. This neural network has $30\times N$ source nodes, $30\times N$ hidden neurons and 2 output neurons, which are fired as pattern 5 or 6 is presented. We apply the second network only when output neuron 5 is activated in the primary neural network. A similar problem occurs in lower faces for distinguishing patterns 1, 2, and 3 as shown in Figure 4.

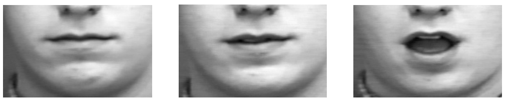

(a) (b) (c)

Fig. 4. Examples. (a) Pattern 1: AU 0. (b) Pattern 2: AU 25. (c) Pattern 3: AU 26.

We also apply a second ad-hoc neural network with $30\times N$ source nodes with equal neuron number in the hidden layer. The output layer has 3 neurons to discriminate patterns 1, 2, and 3. The second neural network is applied when one of those three patterns activates the primary neural network.

III. Experimental Results

We assess the performance of the proposed system considering eight upper-face experiments. The set-up of the experiments is presented in Table 3. For experiments 1, 2, and 3, 8 Gabor kernels with $k_v=\pi/2$ are used. In experiments 4, 5, and 6 we consider both $k_{v1}=\pi/2$ and $k_{v2}=\pi\sqrt{2}/4$, resulting in 16 Gabor kernels.

TABLE 3
SET-UP OF THE EXPERIMENTS

Experiment	Gabor Kernels	Frame difference
1	8	No
2	8	Neutral face
3	8	Neutral face average
4	16	No
5	16	Neutral face
6	16	Neutral face average
7	-	No
8	-	Neutral face average

Experiments 1, 4, and 7 use no frame difference. With experiments 2 and 5 the difference of the maximum facial action intensity image and a neutral image of the subject is carried out. Through experiments 3, 6, and 8 we test our approach, processing the difference of the image with maximum intensity and the average image of neutral faces present in the database. In experiments 1 to 6 the upper face image is directly applied to the Gabor filters and PCA. For experiments 7 and 8 the PCA is directly applied, without using Gabor filters. For the last two experiments, we select 140 principal components and conceive a neural network with 140 source nodes, 94 hidden neurons, and 7 output neurons.

TABLE 4
RESULTS FOR PRIMARY NEURAL NETWORK

Experiment	Recognition Rate
1	67.22 %
2	62.18 %
3	83.19 %
4	56.30 %
5	52.94 %
6	63.22 %
7	54.62 %
8	62.18 %

Experimental results using only the primary neural network for the upper face are shown in Table 4.

By comparing the obtained results of experiments 1, 2, 3 with those of experiments 4, 5, 6, we notice that the use of a reduced set of Gabor kernels and a higher frequency improves significantly the system performance. Donato *et. al.* [7] have also shown that by using a higher frequency it is possible to obtain a better recognition rate.

From Table 4 we can see that the use of the image difference obtained between the maximum intensity image and neutral face estimated from the same image sequence (experiments 2 and 5) does not lead to a satisfactory recognition rate. This problem also occurs when no image difference is used (experiments 1 and 4). This lower recognition rate is due to roughly cropped images and no careful eye and mouth alignments used in our system. To achieve better results a more accurate pre-processing stage is needed, which increases the computational burden. On the other hand, the use of neutral face average image difference (experiments 3, 6, 8) shows to be an efficient approach. The frame average use increases the robustness for not perfectly aligned images, keeping the advantages of using frame difference.

We assess the time required to perform the operations associated with experiment 3, which include: (1) re-scale to 30×45 pixels; (2) subtract the average neutral face image; (3) filtering with 8 Gabor kernels; (4) project each filtered image on 30 principal components; (5) apply the neural network. By using a Pentium IV processor with 1.6 GHz and 256 Mbytes of RAM memory, the time required is around 8 ms. This confirms the efficiency of the proposed system for real-time applications.

The confusion matrix obtained for experiment 3 is shown in Table 5, in which the rows contain the correct input patterns and the columns, the output patterns classified by the primary neural network. We can see that pattern 6 can be confused with pattern 5. Actually, they are very similar, as shown in Fig. 3, and so the neural network training could not correctly model pattern 6.

TABLE 5
CONFUSION MATRIX FOR THE UPPER FACE

	1	2	3	4	5	6	7
1	14	1	0	1	0	0	1
2	3	12	2	0	0	0	0
3	1	2	14	0	0	0	0
4	0	0	0	15	0	2	0
5	0	0	0	0	17	0	0
6	0	0	0	0	6	11	0
7	1	0	0	0	0	0	16

Then, we train a second neural network to distinguish only patterns 5 and 6. The confusion matrix results are shown in Table 6. Note that the ad-hoc neural network is now able to discriminate correctly pattern 6 from pattern 5, similarly as a trained FACS human. In this way, the recognition rate has been increased to 86.55%.

TABLE 6
CONFUSION MATRIX FOR UPPER FACE APPLYING THE SECOND NEURAL NETWORK

	1	2	3	4	5	6	7
1	14	1	0	1	0	0	1
2	3	12	2	0	0	0	0
3	1	2	14	0	0	0	0
4	0	0	0	15	0	2	0
5	0	0	0	0	17	0	0
6	0	0	0	0	2	15	0
7	1	0	0	0	0	0	16

For the lower face, we use the same structure as experiment 3 described in Table 3. Table 7 shows the obtained confusion matrix for the lower face experiment.

Figure 4 confirms the similarity existing between patterns 1, 2, and 3. We have applied an ad-hoc neural network to distinguish these 3 patterns. The obtained confusion matrix is shown in Table 8.

TABLE 7
CONFUSION MATRIX FOR LOWER FACE

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	5	6	0	1	0	2	0
2	1	12	1	0	0	0	0
3	0	6	7	1	0	0	0
4	0	0	3	11	0	0	0
5	0	0	0	0	14	0	0
6	0	0	0	0	0	14	0
7	1	0	0	0	0	0	14

TABLE 8
CONFUSION MATRIX FOR LOWER FACE APPLYING THE SECOND NEURAL NETWORK

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	6	5	0	1	0	2	0
2	0	13	1	0	0	0	0
3	0	5	8	1	0	0	0
4	0	0	3	11	0	0	0
5	0	0	0	0	14	0	0
6	0	0	0	0	0	14	0
7	1	0	0	0	0	0	14

Table 8 shows that the ad-hoc neural network is not capable to distinguish patterns 1, 2, and 3.

Table 9 presents the total number of tested samples. It includes the number of correctly classified samples, number of errors, and recognition rates obtained in experiment 3 for lower and upper faces.

TABLE 9
FINAL EXPERIMENT RESULTS

	Lower face	**Upper face**
#Test samples	98	119
#Correct	80	103
#Errors	18	16
Recognition rate	**81.63%**	**86.55%**

Donato *et. al.* [7] have reported a recognition rate of 95.5% to classify 6 upper and 6 lower face AUs using 40 Gabor kernels and perfectly aligned images with difference of the neutral frame of a subject. The classification technique used is based on the Euclidean distance.

Zhang *et. al.* [1] have obtained a recognition rate of 90.1% to recognize 7 emotion expressions (neutral, happiness, sadness, surprise, anger, disgust, and fear), combining 18 Gabor kernels and geometric features. The coefficients were extracted from 34 fiducial points selected manually and the database is composed of only female subjects. The neural network input data are 612 Gabor wavelet coefficients and more 612 coefficients for the geometric features.

In contrast, our system uses a heterogeneous database, including head motion, neither being necessary to define the neutral face of a subject nor the use of an accurate alignment of fiducial points. We obtain recognition rates of 86.55% and 81.63% for upper and lower faces, respectively, distinguishing very similar facial actions and with a lower complexity pre-processing stage.

IV. CONCLUSIONS

We have presented a recognition system for facial action detection by using Gabor wavelets followed by a PCA and an artificial neural network with neutral face average difference. We have used a set of 14 facial actions for upper and lower faces selected from a heterogeneous subject database with head motion. In addition, an ad-hoc neural network improves the recognition rates to 86.55% and 81.63% for upper and lower face actions, respectively, using 8 Gabor kernels with spatial frequency $k_v = \pi / 2$ and 30 principal components for each filtered image difference. Through the use of the average neutral face image difference we have improved the method's robustness, allowing a roughly cropped and aligned image to be processed successfully by a neural network-based classifier. This approach also reduces the complexity of the pre-processing stage, since the estimate of a neutral image of an unknown subject is no longer needed, making real-time implementations, such as video coders, easier than other previously published approaches.

REFERENCES

[1] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," *Int. Workshop on Automatic Face and Gesture Recognition*, pp. 454-459, 1998.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, Mar. 1991.

[3] P. Eckman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychologists Press*, 1978.

[4] K. D. Craig, S. A. Hyde, and C. J. Patrick, "Genuine, suppressed, and faked facial behavior during exacerbation of chronic low back Pain," *Pain*, vol. 46, pp. 161-172, 1991.

[5] I. Essa and A. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757-763, July 1997.

[6] M. Heller and V. Haynal, "The faces of suicidal depression (translation Les visages de la depression de suicide)," *Kahiers Psychiatriques Genevois (Medicine et Hygiene Ed)*, vol. 16, pp. 107-117, 1994.

[7] G. Donato, M. S. Bartlett, J. C. Hager, P. Eckman, and T. J. Sejnowski, "Classifying facial action," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974-989, Oct. 1999.

[8] Y. Tian, T. Kanade, and J. Cohn, "Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity," *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 229-234, May 2002.

[9] T. Kanade, J. Cohn, and Y. Tian. "Comprehensive database for facial expression analysis," *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 46-53, Mar. 2000.

[10] J. R. Movellan, "Visual speech recognition with stochastic networks," *Advances in Neural Information Processing Systems*, Eds. G. Tesauro, D. S. Touretzky, and T. Leen, vol. 7, pp. 851-858, Cambridge, Mass.: MIT Press, 1995.

[11] B. Fasel and J. Luettin, "Recognition of asymmetric facial action unit activities and intensities," *ICPR 2000*, Spain, pp. 5100-5103, 2000.

[12] I. T. Jollife, "Principal component analysis," *Springer-Verlag,* Berlin, 1986.