A New Model and Process Architecture for Facial Expression Recognition

Ginés García Mateos[1], Cristina Vicente Chicote[1*]

[1] Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Murcia,
30.070 Campus de Espinardo, Murcia, Spain
Tel.: +(34) 968 36 46 08 / +(34) 968 36 46 41
Fax: +(34) 968 36 41 51
(ginesgm, cristina)@dif.um.es

Abstract. In this paper we address the problem of facial expression recognition. We have developed a new facial model based only on visual information. This model describes a set of bidimensional regions corresponding to those elements which most clearly define a facial expression. The problem of facial gestures classification has been divided into three subtasks: face segmentation, finding and describing relevant facial components and, finally, classifying them into one of the predefined categories. Each of these tasks can be solved independently using different techniques already applied to a wide range of problems. This have led us to the definition of a modular, generic and extensible process architecture. A prototype has been developed which makes use of different simple solutions for each module, using a controlled environment and a low-cost vision system. We report the experimental results achieved by the prototype on a set of test images.

Keywords. Facial expression recognition, facial modeling, feature location, facial segmentation, facial components.

1 Introduction

Non-verbal communication plays a basic role in human interaction. Face or hand gestures and voice tone add substantial meaning to communication. In particular, we can associate each human emotion to one or more facial expressions. Actually, this implicit information is an essential feedback for speakers to know whether the audience feel interested, surprised, amused or indifferent to his or her words. Hence, depending on this feedback, the speaker will guide his/her words in a different way to obtain the desired effect on the audience.

Research in facial expression recognition attempts to provide automatic systems with emotional information from their users. Tele-teaching is a good example to

* C. V. C. thanks Fundación Séneca (C.A.R.M.) for a grant.

show how this techniques can help computers to be more friendly and useful. Providing a computer with a camera and a face expression recognition software, the system can follow the student reactions while he or she is studying a lesson and thus guide the teaching process to keep a high attention level. Other examples of applications are information browsing, VR, games, home safety and eldercare [1].

The recognition system presented in this paper proposes a generic and extensible solution to the following problem: given an image or a sequence of images obtained from a human face placed in the foreground in front of a camera, classify its gesture as the most likely one among a set of predefined ones.

2 Related Research

Earlier relevant results in the field of automatic analysis of facial expressions, are due to Ekman and Friesen [2] who in the seventies developed their Facial Action Coding System (FACS), widely accepted nowadays. They envisaged that facial movements are produced by the activation of 43 Action Units (AUs) which can be measured directly from the electrical activity of some muscular regions. Ekman and Friesen defined a set of six basic emotion expressions: surprise, fear, disgust, anger, happiness and sadness. The completeness of this set of emotions is still an open debate.

Most of the ongoing projects are founded on the theoretical principles presented in FACS. For example, the Integrated System for Facial Expression Recognition (ISFER) developed at the Delft University of Technology [3] includes a set of modules, each one implemented using different techniques. Researchers from the M.I.T. have designed an alternative coding method called FACS+ which avoids the use of heuristics by characterising facial movements in a probabilistic way. An on-line description of this system can be found in [4].

Neural networks are commonly applied to the classification phase of the problem. An example is described in reference [5]. Some other approaches should also be mentioned as the one proposed at the Osaka University based on elastic networks [6], the one based on facial flexible models with probabilistic adjust [7] or those based on optical flow analysis [8], among many others.

3 Facial Modeling

The way a generic face is modelled is an essential question when trying to detect facial expressions. Usually, extracting visual information requires to make the model fit the input images. Therefore, the model description determines the way the information is extracted and its structure. Nowadays, it is possible to find in the literature a wide variety of facial models such as tridimensional models of generic faces or those based on flexible grids, eingenspaces, muscular activations or characteristic points or regions.

The model presented in this paper is entirely based on visual information, i.e. biological and environmental causes of image formation have not been considered explicitly. The proposed model defines a set of bidimensional regions representing those frontal facial elements essential for expression recognition. Some possible input images for the recognition problem are shown in Fig. 1.

Fig. 1. Possible input images for facial expression classification. Key information for human face expression analysis can be extracted from the shape and relative position of the eyes, eyebrows and mouth.

Given the previous problem formulation, the model of a human face can be expressed in terms of six main components: eyebrows, eyes, nose and mouth. No matter how this characteristics are detected, their location and appearance are the only information a human being requires to accurately classify a facial expression. Moreover, it is clear that the nose can be removed from the model since it doesn't contribute much to facial gestures.

Some *a priori* information is used to complete the model since we know that relative positions between the five considered elements (nose is not included) follow some fixed rules. For example, the eyebrows will always be placed over the eyes, and the mouth under and between them. Fig. 2 shows a hand-made graphical representation of the resulting model. It would be possible to find an automatic way to build this model by calculating a mean face from a set of examples, but our approximation is enough to reveal the underlying ideas presented here.

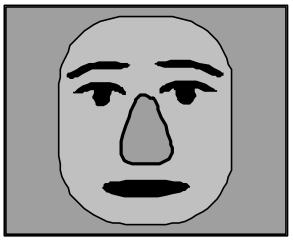

Fig. 2. Graphical representation of the proposed model. Stripped zones correspond to uninteresting regions while non-stripped ones are used to represent special shapes in certain relative positions.

Another important point to be considered is the noise, defined as anything disturbing the model. Apart from acquisition distortions, it is possible to define different kinds of noise depending on the problem they cause. The following classification shows some of them.

- **Component grouping.** Low resolution, bad segmentation or poor illumination

which may cause shadows, can lead to non-separable components.
- **Component distortion.** Shadows produced by facial components, moustache or beard have not been explicitly considered in the model. This could blur the shape and/or the location of the five relevant components.
- **Component occlusion.** Facial components might also be partially or totally hidden by other elements (fringe, dark glasses, etc).

4 Performance Criteria

Usually, within the recognition context, the utmost performance criterion is to maximise the percentage of recognition success and subsidiarily, to minimise the involved resources. Unfortunately, this is not of much help when trying to design a solution.

We could consider the problem of facial expression recognition divided into a subtask for extracting information from the images followed by a classification subtask. Thus, it is possible to define separate performance criteria for each subproblem.

The feature extraction task gives a measure of how input images fit the model. Therefore, we define the following criteria for this step of the recognition process.

- **"Good location" criterion**. The five selected facial components must be detected in positions as close as possible to the real ones. This classical performance criterion was introduced by Canny in his edge detection works. However, in our case, accurate location is not referred to points but to regions corresponding to eyebrows, eyes and mouth.
- **"Good shape description" criterion.** The chosen shape descriptor should be as simple as possible but complete enough to pick up all significant information from the facial components. For example, in order to describe an eyebrow it could be enough to know its height, width, inclination and global curvature.

These criteria are quite difficult to quantify, but they can help us to design an adequate feature extractor for the proposed facial model.

For the classification phase we use the well known criteria of maximising the number of correctly labelled examples.

5 Design of the Facial Expression Recognition Process

The design of the recognition process is based on the model and the criteria previously defined. The aim is to build a modular, extensible and generic architecture which will be evaluated by implementing a prototype.

We consider an architecture to be generic when only the responsibilities of its modules and not their explicit contents, are defined. Thus, modules can be grouped or isolated as their responsibilities converge or differ, respectively.

Most current approaches implicitly divide the recognition process into three main tasks: face segmentation, feature location and extraction, and classification. We explicitly consider this decomposition and propose the schema shown in Fig 3.

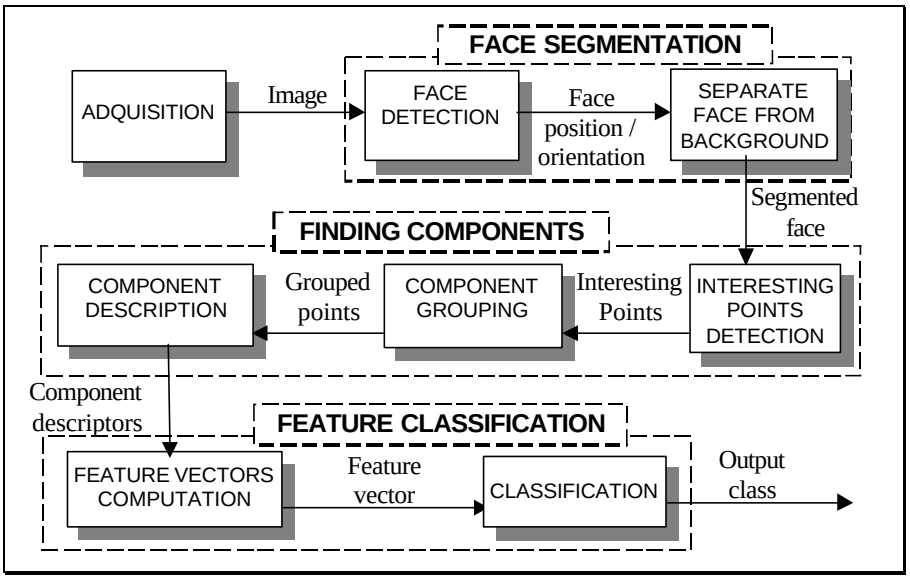

Fig. 3. The facial expression recognition process. Notice the dotted boxes used to group the tasks into three main processes: segmentation, detection and classification.

5.1 Face Segmentation

Segmentation consists of removing non-interesting areas from images in order to simplify the problem. Thus, given the model shown in Fig. 2, we must draw out all stripped regions from the input images. Segmentation comprises the following phases: find the position, orientation and size of the face in the image, and remove non-interesting regions from the input images.

Finding a face in an arbitrary image is not a trivial task. This problem has been widely investigated and some solutions have been proposed. Most of them make use of colour segmentation techniques, as the one presented in [9].

After that, removing non-interesting pixels can be accomplished effortlessly applying a mask to the input image. This mask corresponds to the stripped areas of the model and must be properly translated, scaled and rotated to fit the detected face. Alternatively, we could normalise the input images and use a fixed mask.

5.2 Detection of Facial Components

Once the face has been segmented from the image, it is necessary to find those regions corresponding to the eyebrows, the eyes and the mouth and to represent them using proper shape descriptors. On the one hand, it is known that those five facial components are located in certain relative positions in the face. On the other hand, after segmentation, the position, size and orientation of the face are known. Given these two facts, the regions where to seek for this components are restricted and approximately known.

The proposed method is based on the extraction of some points of interest using low level techniques and afterwards grouping them into the five facial components. A possible solution to find the points of interest could be to threshold the image. However, as uniformity of intensity is not granted through the face, it seems more appropriate to use an edge detection technique which can make use of the fact that the five components and only them are darker than the rest of the face.

Once the points of interest have been extracted, to group them into components is an easier task, although not trivial. From the model, it is obvious that five and only five components must be found and their most likely positions are known. Thus, the grouping process is mainly guided by the *a priori* information extracted from the model. Nevertheless, some problems as spurious points or non-separable components must be carefully taken into account.

After having grouped the points into the components, adequate shape descriptors must be computed to reduce the input information for the classifier. A simple solution could be to describe each region with one or more characteristic points.

An alternative to the extraction of the points of interest could be the use of deformable contours or snakes [10]. They would be initially located on the *a priori* position of the components and would incrementally fit the areas of maximum intensity gradient. Thus, the tasks of component grouping and description will be accomplished simultaneously.

5.3 Feature Classification

Different kinds of output could be considered as the result of the recognition process, such as the opening degree of the eyes or of the mouth, the probability of a certain facial expression, etc. Here, we shall consider the output of the classifier, i.e. the class to which the input image belongs, according to the classification criteria.

Classification is a generic and widely studied problem in A.I. Linear discriminants, nearest neighbours methods and neural networks are the mechanisms most commonly applied to solve it, although there exist lots of others [11]. All of them, include a vast set of techniques. Nevertheless, they all need to work with a set of features representative enough of the selected classes. Selecting proper features as the input to the classifier is as important as the quality of the classifier itself.

For the set of features to be representative enough, a huge amount of data might be needed. Some classifiers can cope with high dimension inputs, but if the number of shape descriptors parameters is too high, a problem of undertraining can arise. To

avoid it, it is possible to introduce an intermediate stage for reducing the extracted information to a small and meaningful computed set of features. Data mining techniques could be applied to automatically perform this extraction.

6 Prototype Implementation

The implemented prototype is a fulfilment of the architecture described in 5, showing its practical viability. It is not our objective to build a complex system, but to select different simple solutions for each module. The prototype has been tested in a controlled environment.

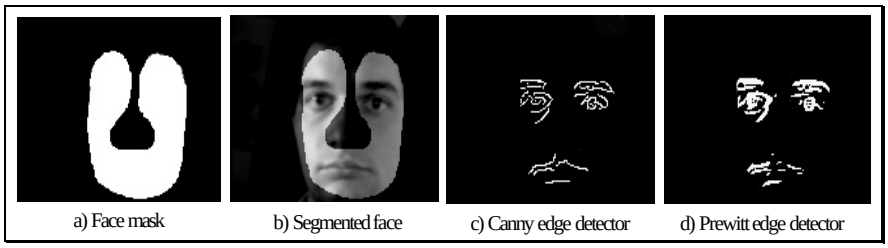

Fig. 4. Mask and edge detection examples. a) Predefined segmentation mask, b) Segmented face using (a), c) Edges found in (b) using the Canny operator, d) Edges found in (b) using the Prewitt operator.

The prototype makes use of a simplified definition of the problem given by the assumption of fixed position of the face within the image: the position, size and orientation of the face are considered to be fixed and known through all the images taken from the same individual. This supposition makes trivial the segmentation of the face. Removing non-interesting regions is carried out using always the same *a priori* defined mask. Figure 4a shows the mask applied in the tests.

Facial component detection comprises two phases. Firstly, the points of interest are searched for. Then, they are grouped into components and characterised using proper shape descriptors. Edge points have been used as points of interest as its validity was justified in 5.2. Figures 4c and 4d compares the edges retrieved by two operators. Although Canny's edge detector is the most adequate in most applications, in our case a simpler edge detector has been applied. We have chosen Prewitt's operator, since it is simpler to compute and we are interested in detecting regions instead of finding edges.

A flexible and elegant way to solve the problem of grouping the points of interest is to apply a mixture model using the EM algorithm to adjust its parameters. We assume that edge points correspond to a random sample from a bidimensional probability function made up by a mixture of five gaussian probability functions, one for each facial component. Using bidimensional gaussians as basic components of the mixture is equivalent to use elliptical shape descriptors. This choice achieves a good ratio between the information they supply and the descriptor complexity. Given the

edge points, to determine the gaussian parameters, i.e. mean and covariance matrix, we have applied the iterative EM algorithm. Our implementation of the algorithm includes the noise treatment proposed in [12].

For the classifier, a simple solution has been chosen: a nearest neighbour method using Mahalanobis distance. Before classification itself is carried out, a feature vector is calculated in order to reduce the amount of information the classifier has to deal with. Six features have been empirically defined by analysing the way the five facial components change through the different facial gestures. Specifically, the system computes from the gaussian parameters previously obtained, the mouth width and opening degree, eyes opening degree, eyes-eyebrows distance, eyes-mouth distance and eyebrows angle.

7 Tests and Results

For the acquisition of test images we have used a low-cost videoconference camera. These kind of devices are the most commonly used in tele-teaching environments and its quality is good enough for our purposes. Recorded images have a 160x120 pixels resolution using 256 grey levels.

A set of six basic facial expressions has been defined: normal, sleeping, smiling, surprised, yawning and angry. This expressions are expected not to be forced or exaggerated but natural. Despite of this fact, they must be non-ambiguous for a human observer.

From each class, we have recorded 10 examples under the same lighting conditions and 5 more in a different moment and conditions, all of then from the same individual. Half of the first ones have been used to train the system, while the other 10 have been used as test images. Faces are supposed to be located in a fixed position within the images, but in practice there exist a 5% mean deviation in face position in the first set of examples and an 8% in the second set, with respect to face height.

Global classification results are shown in Table 1. In each cell two values are shown: the left ones correspond to the rate achieved on the set with same illumination conditions, while the right correspond to the different conditions set.

Table 1. Results of classification tests

	Output Class											
Real Class	**Normal**		**Sleeping**		**Yawning**		**Laughing**		**Surprised**		**Angry**	
Normal	3	2				1		2			2	
Sleeping			5	4						1		
Yawning					4	5					1	
Laughing							5	1				4
Surprised							1		4	5		
Angry	2										3	5

As shown in Table 1, the system achieves an 80% of successfully classified examples from the first set, while a 73% on the second set of images. In the first set

classes are well differentiated and higher confusion appears between ‘normal’ and ‘angry’ expressions. The second set performs worse, and errors are more irregularly distributed. Some other simplified tests have been performed over sequences of images and colour images. These tests exhibited similar results. Three examples of execution of the prototype are shown in Fig. 5.

Although the prototype was not optimised in execution time, it is appropriate to emphasise the high speed achieved. Thus, a complete execution, including reading the input image from disk and its graphical representation on the screen, takes about 0.35 seconds in a K6 processor working at 350 MHz.

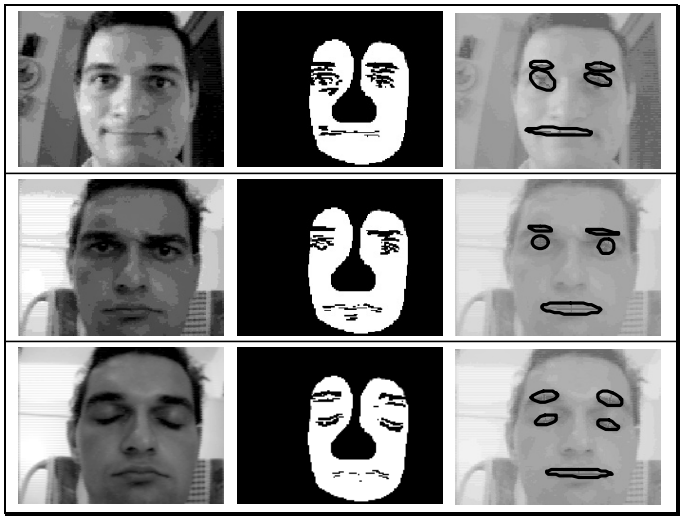

Fig. 5. Three prototype execution examples. Left: input images; middle: face mask and edge points; right: resulting shape descriptors. Top to down: laugh, anger, sleeping. All examples were successfully classified by the prototype.

8 Conclusions

In this paper we have addressed the problem of facial expression recognition. We propose a facial model entirely based on visual information and not on biological or tridimensional models of the face. As a result, we have designed a model based on 2D characteristic regions corresponding to the five most relevant facial components.

The developed process is highly based on this model. First, the division between interesting and non-interesting regions leads to a segmentation phase which extracts the exact position, size and rotation of the face in the image. Secondly, the model enumerates a reduced set of components which unequivocally determine the facial gesture. This second stage comprises the location and description of these compo-

nents. Finally, the classification process uses shape descriptors to make its final decision about the facial expression.

The result is a modular and generic process definition. The modules pursue clearly differentiated, non-overlapped and generic objectives which comprehend a wide range of research areas more than a reduced set of techniques. Consequently, the prototype implementation allowed us to face the problem from a more practical point of view and to experiment with the designed process. The prototype is a fulfilment of this architecture. Implementation decisions were based on the *a priori* defined criteria for a good characterisation of the input images: component location and shape description.

The final percentage of successfully classified examples is over 75%. It undoubtedly reflects the degree of simplicity/complexity of the set of test images used. Nevertheless, the fact that the system achieves very similar results for both sets of input images, using same illumination conditions and different ones, is a good evidence of the viability of the implemented scheme.

Acknowledgements and Thanks

We would like to thank Dr. Alberto Ruiz, who helped us to design the process and made a global review of this article. Thank you to Begoña Moros, who kindly participated in the testing of the prototype.

This work has been partially supported by CICYT project TIC1998-0559.

References

1. Pentland, A.: Looking at People: Sensing for Ubiquitous and Wearable Computing, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, (January 2000)
2. Ekman, P., Friesen, W.V.: Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action. Manual, Consulting Psychologists Press, Palo Alto, CA (1978)
3. Rothkrantz, L.J.M., van Schouwen, M.R., Ververs, F., Vollering, J.C.M.: A multimedial tool for facial expressions, Euromadia '98, Leicester (1998)
4. Essa, I.A., Pentland, A.: Dynamic Facial Analysis System: Coding, Analysis, Interpretation, and Recognition of Facial Motions, http://www-white.media.mit.edu/~irfan/DFACE.demo /Dface.demo.html, Massachusetts Institute of Technology (1996)
5. Pantic, M., Rothkrantz, L.J.M.: Automated Facial Expression Analysis, ASCI'98 Proceedings of the fourth annual conference of the Advanced School for Computing and Imaging, Lommel, Belgium (1998)
6. Kimura, S., Yachida, M.: Facial Expression Recognition and Its Degree Estimation, Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR-97), Puerto Rico (1997)
7. Lanitis, A., et alt.: Automatic Interpretation and Coding of Face Images Using Flexible Models, IEEE Transactions on Pattern Analysis and Machine Intell., Vol. 19, No. 7, (July 1997)
8. Mase, K.: Recognition of facial expressions by optical flow, IEICE Transactions, Special Issue on Computer Vision and its Applications, E 74(10) (1991)
9. Yang, J., Waibel, A.: A Real-Time Face Tracker, Proceedings of WACV'96, Saratosa,

Florida, pp. 142-147 (1996)
10. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contours Models, Proceedings First International Conference of Computer Vision, London, pp. 259-269 (1987)
11. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, (January 2000)
12. Lopez-de-Teruel, P.E., Ruiz, A.: On-Line Probabilistic Learning Techniques for Real Time Computer Vision, WorkShop Learning 98, Getafe, Spain (1998)