

Real-Time Facial Expression Recognition for Natural Interaction

Eva Cerezo¹, Isabelle Hupont¹, Critina Manresa², Javier Varona²,
Sandra Baldassarri¹, Francisco J. Perales², and Francisco J. Seron¹

¹Departamento de Informática e Ingeniería de Sistemas, Instituto de Investigación en
Ingeniería de Aragón, Universidad de Zaragoza, Spain
{ecerezo, 478953, sandra, seron}@unizar.es

²Departament de Matemàtiques i Informàtica, Universitat de les Illes Balears, Spain
{crisrina.manresa, xavi.varona, paco.perales}@uib.es

Abstract. The recognition of emotional information is a key step toward giving computers the ability to interact more naturally and intelligently with people. This paper presents a completely automated real-time system for facial expression's recognition based on facial features' tracking and a simple emotional classification method. Facial features' tracking uses a standard webcam and requires no specific illumination or background conditions. Emotional classification is based on the variation of certain distances and angles from the neutral face and manages the six basic universal emotions of Ekman. The system has been integrated in a 3D engine for managing virtual characters, allowing the exploration of new forms of natural interaction.

Keywords: real-time features tracking, emotional classification, natural interfaces.

1 Introduction

Human computer intelligent interaction is an emerging field aimed at providing natural ways for humans to use computers as aids. It is argued that for a computer to be able to interact with humans it needs to have the communication skills of humans. One of these skills is the ability to understand the emotional state of the person, and the most expressive way humans display emotions is through facial expressions. Nevertheless, to develop a system that interprets facial expressions is difficult. Two kinds of problems have to be solved: facial expression feature extraction and facial expression classification. Related to feature extraction, and thinking in interface applications, the system must be low-cost with real-time, precise and robust feedback. Of course, no special lighting or static background conditions can be required. The face can be assumed to be always visible, however, difficulties can arise from in-plane (tilted head, upside down) and out-of-plane (frontal view, side view) rotations of the head, facial hair, glasses, lighting variations and cluttered background [1]. Besides, when using standard USB web cams, the provided CMOS image resolution has to be taken in account. Different approaches have been used for non invasive face/head-based interfaces, mainly for the control of the head's position analyzing

facial cues such as color distributions [2], head motion [3] or, recently, by means of facial features' tracking [4,5]. From the extracted facial features, emotional classification has to be performed. Three different classification methods are usually used for expression recognition: patterns, neuronal networks or rules [6]. Most of them follow the emotional classification of Ekman [7] that describes six universal basic emotions: joy, sadness, surprise, fear, disgust and anger.

The aim of this work is to show how a non-invasive robust face tracking system can feed an effective emotional classifier to build a facial expression recognition system that can be of great interest in developing new multimodal user interfaces. As it will be shown, the system developed has been successfully integrated in a character-based interface, allowing the exploration of new forms of affective interaction.

2 Real-Time Facial Feature Tracking

The computer vision algorithm is divided into two steps: initialization and tracking. The initialization step is responsible of learning the user's facial characteristics such as its skin color, its dimensions and the best face features to track. This process is totally automatic and it can also be used for system's recovering when a severe error occurs, adding the robustness necessary so that it can be used in a human-computer interface.

First of all, the algorithm automatically detects the user's face by means of a real-time face detection algorithm [8]. The face will not be considered as found until the user sits steady for a few frames and the face is detected in the image within those frames. A good detection of the features is very important for an effective performance of the whole system and the user must start the process with the so called neutral face: the mouth is closed and the gaze is directed perpendicular to the screen plane, the eyes are open and the eyelids are tangent to the iris. Then, it is possible to define the initial user's face region to start the search of the user's facial features. Based on anthropometrical measurements, the face region can be divided into three sections: eyes and eyebrows, nose, and mouth region. In the nose region, we look for those points that can be easily tracked, that is, those whose derivative energy perpendicular to the prominent direction is above a threshold [9]. This algorithm theoretically selects the nose corners or the nostrils. However, the ambient lighting can cause the selection of points that are not placed over the desired positions; this fact is clearly visible in Fig. 1 (a). Ideally, the desired selected features should be at both sides of the nose and should observe certain symmetrical conditions. Therefore, an enhancement and a re-selection of the features found is carried out taking into account symmetrical constraints. Fig. 1 (b) shows the selected features when symmetry respect to the vertical axis is considered. This reselection process achieves the best features to track and contributes to the tracking robustness. Fig. 1 (c) illustrates the final point considered, that is, the mean point of all the final selected features; due to the reselection of points it will be centered on the face.

Finally, in order to learn the user's skin color and complete the initialization step, the pixels inside the face region are used as a learning set of color samples to find the

parameters of a Gaussian model in 3D RGB density using standard maximum likelihood methods.

The aim of the tracking step is to control the position of the face in order to detect and constraint the search region of the 10 face features used in the expression recognition stage. The detected and enhanced features of the initialization step are tracked by using the spatial intensity gradient information of the images in order to find the best image registration [10]. As it was mentioned before, for each frame the mean of all nose features is computed and it is defined as the *face tracking point* for that frame. The tracking algorithm is robust for handling rotation, scaling and shearing, so that the user can move in a more unrestricted way.

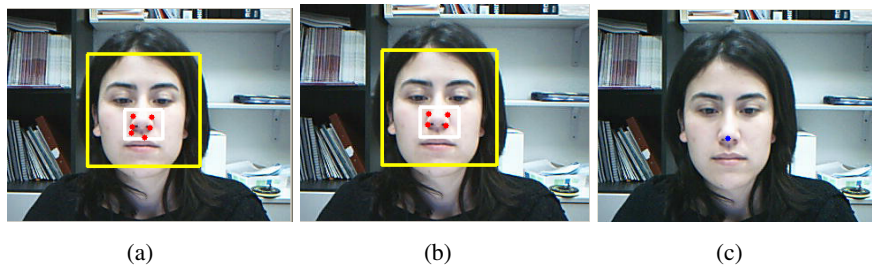


Fig. 1. (a) Automatic face detection and initial set of features. (b) Best feature selection using symmetrical constraints. (c) Mean of all features: face tracking point.

The *face tracking point* is used to constrain the image region to process and the color probability distribution, both computed in the initialization step, is used to calculate the probability of a face pixel being skin so that “skin mask” of the user’s face can be created. Using this mask the system can detect, as a result of their non-skin-color property, the user’s eyebrows, eyes and mouth bounding boxes and due to their position related to the *face tracking point*, the system can label the zones. One problem can appear if the user has got his eyes a little bit sunk, then due to the shadow in the eyelid, most probably the eyebrow and eye will be found as a single blob. In that case, we divide this bounding box assuming that the eyebrow has been detected together with the eye. Finally, from the bounding boxes positions, 10 face features are extracted. These 10 feature points of the face will later allow us to analyze the evolution of the face parameters (distances and angles) used for expression recognition. Fig. 2 shows the correspondence between these points and the ones defined by the MPEG-4 standard.

3 Classification of Emotions

3.1 General Method Description

Our classification method works with the emotional classification of Ekman and it is based on the work of Hammal et al [11]. They have implemented a facial classification method for static images. The originality of their work consisted, on the

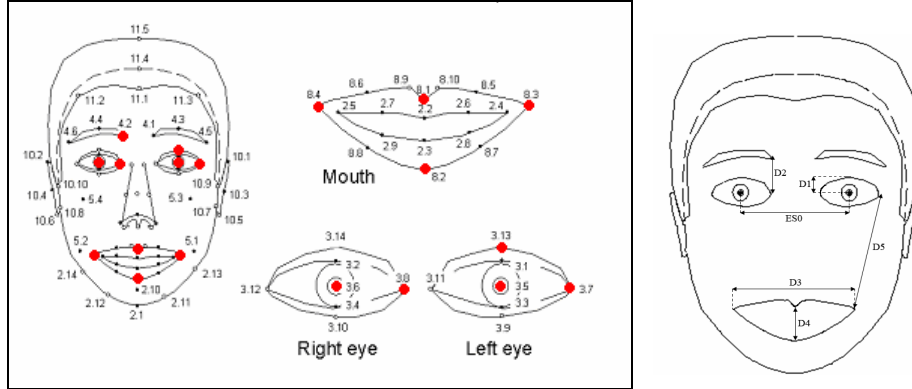


Fig. 2. Facial feature points extracted and used for expression recognition according to the MPEG-4 standard (left). Characteristic distances used in our method (right).

one hand, in the supposition that all the necessary information for the recognition of expressions is contained in the deformation of certain characteristics of the eyes, mouth and eyebrows and, on the other hand, in the use of the Belief Theory to make the classification. Our method studies the variation of a certain number of face parameters (basically distances and angles between some feature points of the face) with respect to the neutral expression. The characteristic points, shown in section 2, are used to calculate the five distances also shown in Fig. 2. All the distances are normalized with respect to the distance between the eyes, which is a distance independent of the expression. In addition to the five distances our system works with additional information about the mouth shape (from the four feature points two angles and the width/height relationship is extracted).

The objective of our method is to assign a score to each emotion, according to the state acquired by each one of the parameters in the image. The emotion (or emotions in case of draw) chosen will be the one that obtains a greater score.

Each parameters can take three different states for each of the emotions: C^+ , C^- and S. State C^+ means that the value of the parameters has increased with respect to the neutral one; state C^- that its value has diminished with respect to the neutral one; and the state S that its value has not varied with respect to the neutral one. First, we build a descriptive table of emotions, according to the state of the parameters, like the one of the Table 1 (left). From this table, a set of logical rules tables can be built for each parameter (right), in which a score is assigned to each state for each emotion, depending on the degree in which this state of the parameter is characteristic of the emotion. Once the tables are defined, the implementation of the identification algorithm is simple. When a parameter takes a specific state, it is enough to select the vector of emotions (formed by the scores assigned to this state for each emotion) corresponding to this state. If we repeat the procedure for each parameter, we will obtain a matrix of as many rows as parameters we study and 6 columns, corresponding to the 6 emotions. The sum of the scores present in each column of the matrix gives the total score obtained by each emotion. If the final score does not surpass a certain threshold, the emotion is classified as “neutral”.

Compared to the method of Hammal, ours is computationally simple. The combinatory explosion and the number of calculations to make are considerably reduced, allowing us to work with more information (more parameters) of the face and to evaluate the six universal emotions, and not only four of them, as Hammal does.

Table 1. Proposed table of one parameters' states for each emotion (left) and logical rules table for that parameter

	Pi						
Joy	C-						
Surprise	C+						
Disgust	C-						
Anger	C+						
Sadness	C-						
Fear	S/C+						

Pi	E1 joy	E2 surprise	E3 disgust	E4 anger	E5 sadness	E6 fear
C+	0	3	0	2	0	1
C-	1	0	2	0	2	0
S	0	0	0	0	0	1

3.2 Tuning the Method: The FG-NET Database

In order to define the emotions in terms of the parameters states, as well as to find the thresholds that determine if a parameter is in a state or another, it is necessary to work with a wide database. In this work we have used the facial expressions and emotions database FG-NET of the University of Munich [12] that provides images of 19 different people showing the 6 universal emotions from Ekman plus the neutral one. From these data, we have built a descriptive table of the emotions according to the value of the states (Table 2).

Table 2. Proposed table of the states for the parameters used by the classification method. Some features do not provide any information of interest for certain emotions (squares in gray) and in these cases they are not considered.

	D₁	D₂	D₃	D₄	D₅	Ang 1	Ang 2	W/H
Joy	C-	S/C-	C+	C+	C-	C+	S/C+/C-	S/C-
Surprise	S/C+	S/C+	S/C-	C+	S/C+	C-	C+	C-
Disgust	C-	C-	S/C+/C-	S/C+	S/C-	S/C+/C-	S/C+	S/C-
Anger	C-	C-	S/C-	S/C-	S/C+/C-	C+	C-	C+
Sadness	C-	S	S/C-	S	S/C+	S/C+/C-	S/C-	S/C+
Fear	S/C+	S/C+/C-	C-	C+	S/C+	C-	C+	C-

3.3 Validation

Once the states that characterize each emotion and the value of the thresholds are established, the algorithm has been tested on the 399 images of the database. In the evaluation of results, the recognition is marked as "good" if the decision is coherent with the one taken by a human being. To do this, we have made surveys to 30 different

people to classify the expressions shown in the most ambiguous images. Related to classification success, it is interesting to realize that human mechanisms for face detection are very robust, but this is not the case of those for face expressions interpretation. According to Bassili [13], a trained observer can correctly classify faces showing emotions with an average of 87%. The obtained results are shown in Table 3. The method has also been tested with other databases different from the one used for the threshold establishment, in order to confirm the good performance of the system.

Table 3. Classification rates of Hammal [11] (second column) and of our method with five distances (second column) and plus the information about the mouth shape (third column)

EMOTION	% SUCCESS HAMMAL METHOD	% SUCCESS FIVE DISTANCES	% SUCCES MOUTH SHAPE
Joy	87.26	36.84	100
Surprise	84.44	57.89	63.16
Disgust	51.20	84.21	100
Anger	not recognized	73.68	89.47
Sadness	not recognized	68.42	94.74
Fear	not recognized	78.95	89.47
Neutral	88	100	100

3.4 Temporal Information: Analysing Video Sequences

After having tuned and validated the classification system with the static images, the use of the automatic feature extraction has enabled us to track video sequences of user's captured by a webcam. Psychological investigations argue that the timing of the facial expressions is a critical factor in the interpretation of expressions. In order to give temporary consistency to the system, a temporary window that contains the emotion detected by the system in each one of the 9 previous frames is created. A variation in the emotional state of the user is detected if in this window the same emotion is repeated at least 6 times and is different from the detected in the last emotional change.

The parameters corresponding to the neutral face are obtained calculating the average of the first frames of the video sequence, in which the user is supposed to be in the neutral state. For the rest of the frames, a classification takes place following the method explained in the previous sections.

4 Application: New Input Data for Natural Interfaces

To demonstrate the potential of our emotional tracking system, we have added it to Maxine [14], a general engine for real-time management of virtual scenarios and characters developed by the group. Maxine is a tool that has been created with the aim of making it easy the use of character-based interfaces in different application

domains. The general vision is that if a user's emotion could be recognized by computer, human interaction would become more natural, enjoyable and productive. The system presented here has been configured as a new multimodal input to the system. The system recognizes the emotion of the user and responds in an engaging way. The features extraction program captures each facial frame and extracts the 10 feature points which are sent to the emotion classifier. When an emotional change is detected, the output of the 7-emotion classifier constitutes an emotion code which is sent to Maxine's character. For the moment, the virtual character's face just mimics the emotional state of the user (Fig. 3), accommodating his/her facial animation and speech.

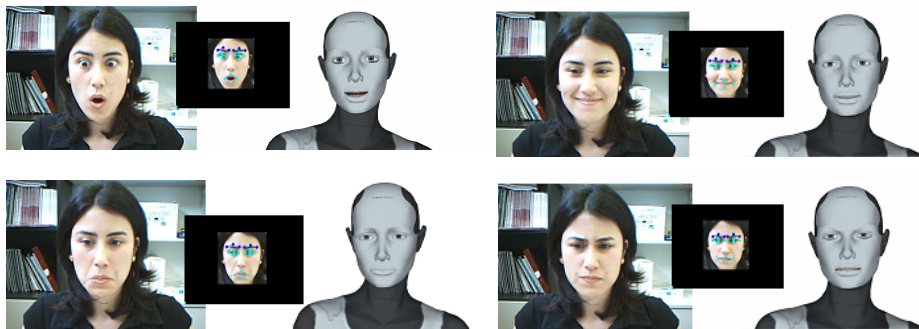


Fig. 3. Examples of the integrated real-time application: detection of surprise, joy, sadness, anger. For each example, images captured by the webcam, small images showing automatic features' tracking and synthesized facial expressions are shown. The animated character mimics the facial expression of the user.

5 Conclusions and Future Work

We have presented a simple and effective system for the real-time recognition of facial expressions. In opposition to other systems that rely on the use of wearable detectors, the system developed is non-invasive and is based on the use of a simple low cost webcam. The automatic features extraction program allows the introduction of dynamic information in the classification system, making it possible the study of the time evolution of the evaluated parameters, and the classification of user's emotions from live video.

To test its usefulness and real-time operation, the system has been added to the Maxine system, an engine developed by the group for managing 3D virtual scenarios and characters to enrich user interaction in different application domains. For the moment, and as a first step, the emotional information has been used to accommodate facial animation and speech of the virtual character to the emotional state of the user. More sophisticated adaptive behaviour is now being explored. As it has been pointed out, recognition of emotional information is a key step toward giving computers the ability to interact more naturally and intelligently with people.

Acknowledgments

We would like to thank Sergio Garcia Masip for his work in Maxine. This work has been partially financed by the Spanish "Dirección General de Investigación", contract number N° TIN2004-07926 and by the Aragon Government through the WALQA agreement (ref. 2004/04/86).

J. Varona acknowledges the support of a Ramon y Cajal fellowship from the Spanish MEC.

References

1. Turk, M., Kölsch, M.: Perceptual Interfaces. In: Medioni, G., Kang, S.B. (eds.) *Emerging Topics in Computer Vision*, Prentice Hall, Englewood Cliffs (2005)
2. Bradski, G.R.: Computer Vision Face Tracking as a Component of a Perceptual User Interface. In: *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 214–219 (1998)
3. Toyama, K.: "Look, Ma – No Hands!" Hands-Free Cursor Control with Real-Time 3D Face Tracking. In: *Proceedings of the Workshop on Perceptual User Interfaces*, pp. 49–54 (1998)
4. Gorodnichy, D.O., Malik, S., Roth, G.: Nouse 'Use Your Nose as a Mouse' – a New Technology for Hands-free Games and Interfaces. *Image and Vision Computing* 22, 931–942 (2004)
5. Betke, M., Gips, J., Fleming, P.: The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access for People with Severe Disabilities. *IEEE Transactions on neural systems and Rehabilitation Engineering*, vol. 10 (2002)
6. Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions: The State of the Art. *Pattern Analysis and Machine Intelligence. IEEE Transactions* 22(12), 1424–1445 (2000)
7. Ekman, P.: *Facial Expression, the Handbook of Cognition and Emotion*. John Wiley et Sons, Chichester (1999)
8. Viola, P., Jones, M.: Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 137–154 (2004)
9. Shi, J., Tomasi, C.: Good Features to Track. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 593–600 (1994)
10. Baker, S., Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision* 56, 221–225 (2004)
11. Hammal, Z., Couvreur, L., Caplier, A., Rombaut, M.: Facial Expressions Recognition Based on the Belief Theory: Comparison with Different Classifiers. In: *Proc. 13th International Conference on Image Analysis and Processing* (2005)
12. <http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html> (Reviewed in February 2006)
13. Bassili, J.N.: Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology* 37, 2049–2059 (1997)
14. Seron, F., Baldassarri, S., Cerezo, E.: MaxinePPT: Using 3D Virtual Characters for Natural Interaction. In: *Proc. WUCAmI'06: 2nd International Workshop on Ubiquitous Computing and Ambient Intelligence*, pp. 241–250 (2006)