Interactive Analysis for MPEG-4 Facial Models Configuration

Gabriel Antunes Abrantes, Fernando Pereira
Instituto Superior Técnico - Instituto de Telecomunicações
Av. Rovisco Pais, 1096 Lisboa Codex, PORTUGAL
E-mail: Gabriel.Abrantes@lx.it.pt; Fernando.Pereira@lx.it.pt

Abstract

Facial animation has been exhaustively studied for the past 20 or 30 years [1]. However, so far, no standard animation and coding methodologies, allowing to achieve interoperability between different facial animation implementations have been defined. Since the integration of natural and synthetic content is one of the major objectives of the forthcoming MPEG-4 content-based audiovisual coding standard, facial animation has finally seen some degree of standardisation.

The purpose of this paper is to describe the work done at Instituto Superior Técnico (IST) towards the extraction of facial data allowing the adaptation, by the sender, to a particular face, of a generic 3D facial model available at the receiver, using the syntactic elements specified in MPEG-4. This data can be obtained in a fully automatic way or by means of user interaction, refining the automatic results. This paper will shortly describe the interactive analysis scheme implemented and will also present some configuration results.

1. Context

The emerging MPEG-4 standard will be the first audiovisual coding standard understanding an audiovisual scene as a composition of audiovisual objects. The visual objects can be of natural or synthetic origin, 2D or 3D. These objects will have specific characteristics, notably spatial and temporal behaviour. The scene composition approach supports new functionalities, such as content-based coding and interaction [2], since the objects can be independently processed and accessed. Due to its relevance in terms of foreseen applications, 3D facial animation has been one of the synthetic elements addressed by MPEG-4. The applications, which should benefit from the standardisation of some basic tools in this area, range from videotelephony, and tele-learning, to storyteller on demand, and facial agents in kiosks. To provide facial animation capabilities, MPEG-4 decided to standardise two sets of parameters [3]:

i) Facial Animation Parameters (FAP) representing a set of basic facial actions related to some key facial features, such as the lips, the eyes, the head, the jaw, allowing the animation of a 3D facial model. The way FAP are supposed to be interpreted is described in the standard, in order that similar animation results are obtained independently of the 3D facial model to be used at the receiver (no facial model is standardised). FAP are intended to be used all along the animation.

ii) Facial Definition Parameters (FDP) allowing to configure/adapt any 3D facial model (at the receiver) to a particular face (by the sender) (see Figure 1). FDP consist in a set of 3D feature points, defining the basic geometry of a face, and a texture with the associated feature points, if texture

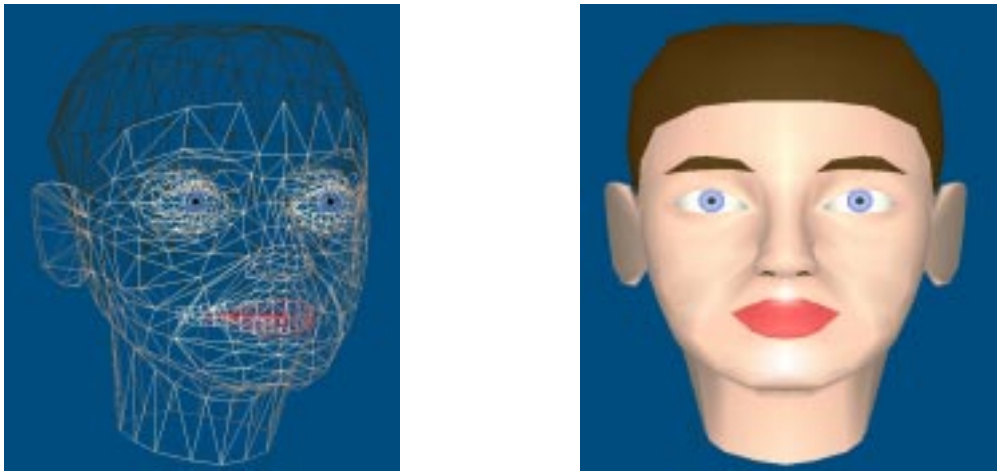

Figure 1 - IST facial model: (a) only with polygons; (b) with a simple texture

mapping is to be performed. Feature points can also be sent without any texture information if a less realistic animation is acceptable. FDP are supposed to be sent only once at the beginning of a session. This calibration data can be obtained in two ways: i) by extraction from real images, and ii) by synthetic generation.

This paper describes a way to obtain a set of 3D feature points and the corresponding texture information, allowing the receiver model to be calibrated in a realistic way by the sender.

2. Automatic Extraction of Feature Points

Depending on the application, facial analysis may need to be fully automatic or allow human interaction, giving guidance or refining the automatic results. While some applications like videotelephony will hardly allow interaction, requiring an automatic system with real time analysis, other applications like facial agents in kiosks allow off-line analysis and thus human interaction in order that better calibration and more realistic animations are obtained.

The automatic analysis system implemented extracts only the feature points corresponding to the most important facial features, namely those associated to the head, the mouth, the eyes, the nose and the chin. This decision was made considering that not all feature points have the same importance and that it would be very difficult to automatically obtain the complete set of feature points from a single image, specially for the teeth and the back of the neck. The current selection allows achieving a good and realistic adaptation of a 3D facial model.

The first step in this facial analysis scheme consists in locating the head within the image. For this purpose, the well-known snake algorithm was implemented [4], a method used to find objects in an image. A snake is an energy minimising spline that is pulled towards the edges of an object. To apply the snake algorithm, an edge image with the most important edges is obtained using the Sobel's edge detection method, after a morphological open filter is applied to simplify the original image. Other edge detection methods, like Canny's algorithm, are being studied in order to improve the edge extraction process. The morphological opening operation is applied to each component of the image *R*, *G* and *B* individually, and the individual results are weighted and summed to obtain the final image. The same procedure is taken for the Sobel's edge detection method. Moreover, in order that the image is clean from weak intensity spots due to colour variations, the Otsu's threshold selection method is applied [5]. The snake algorithm is then applied to the edge image, so that the corresponding contour is pulled towards the head boundary. However, in order to obtain good results, some constraints are required. First of all, the background should have as few objects that can originate strong edges as possible, to avoid obstructing the snake from contracting itself on the head boundary, and should have a good contrast with the face and the hair. The scene should also be well illuminated, so that the edges are well defined.

After locating the head, the automatic analysis proceeds to the detection of the remaining facial elements. For locating the mouth, the eyes, the nose and the chin, the assumption that these are areas of high contrast and thus originate strong edges, more specifically horizontal edges if the head is not rotated,

is made. Using *a priori* knowledge for the location of these facial elements, obtained from the available test material, the horizontal edge image is first selectively amplified and then the position of the facial elements is determined. This method is very sensitive to bad facial illumination, since this generates weak edges, and it also does not perform well if a bad head detection was made. There are other different methods used for facial analysis, like template matching [6], neural nets [7] or the use of Eigenfaces [6], which however are more suitable for face recognition and need large training sets of face images.

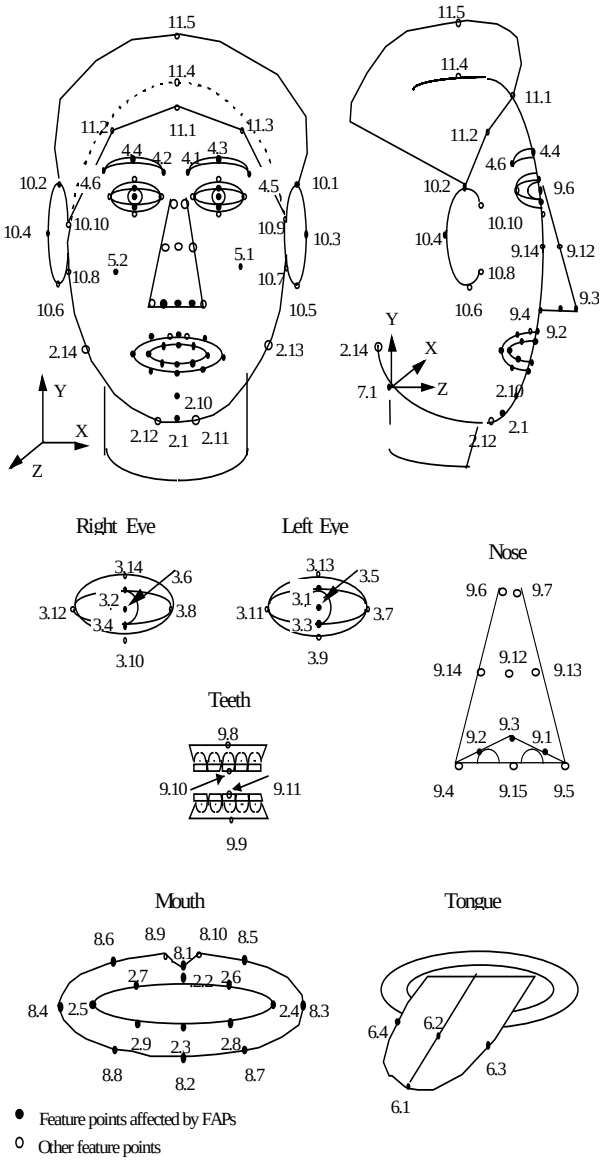

Figure 2 – Feature points grouping [3]

Once all the specified facial features are obtained, the feature points can be determined since they are closely associated to the position of the detected facial elements (see Figure 2). These features points are then used to map a projection of the 3D facial model on the texture image. After adapting the projection of the face model to the obtained feature points, the remaining feature points are extracted by getting the coordinate values of the corresponding vertices. With this process (using a single facial view), only the feature points on the texture are obtained, but not 3D feature points that can define the 3D geometry of the head. As there is no side-view information, the size of the face model in the *z* direction is resized to be harmonious with the other two directions, by determining the *z* coordinate values for the feature points. First an orthogonal projection on a 2D plane of the IST face model [6], which corresponds to an average face, is made. The projection is then adapted to the extracted 2D texture points. The *x* and *y* coordinate values of the vertices are extracted from the adaptation, and the relations between the original *x* and *y* coordinates of the model vertices and the ones of the adapted model are obtained. The 3D face model is then resized in the *z* direction according to the average of these two relations, and the *z* coordinate of the pretended 3D feature points are extracted from the corresponding vertices on the 3D face model. The resulting profile may be different of the real face, but differences are acceptable, since they only become noticeable when the face model is very rotated.

3. User Refinement of Feature Points

As it has been previously explained, it is very difficult to have an automatic analysis system that works well under very different conditions. Acknowledging that many applications do not mandatory need fully automatic analysis, it was decided to implement a module that allows adjusting/refining the less correct automatically extracted data. In this work, two types of interaction targeting the adjustment of the automatically detected set of feature points have been considered:

i) **2D interaction** - the user may vertically and horizontally (*x,y*) adjust the position of each detected texture feature point, displayed over the facial texture;

ii) **3D interaction** - the user may adjust each of the 3 coordinates for the 3D feature points, determined

as above indicated, shaping the model geometry. For both cases, a refined set of features points is generated.

4. Some Results

Due to space constraints, adaptation results are only showed for one image, using the IST 3D facial model. Figure 3 shows the various steps for the extraction of the specified set of feature points. Although, for this case, the automatic analysis already provides acceptable results, there is still room for interactive improvement, notably by refining the position of the eyebrows, mouth and chin, achieving better adaptation results.

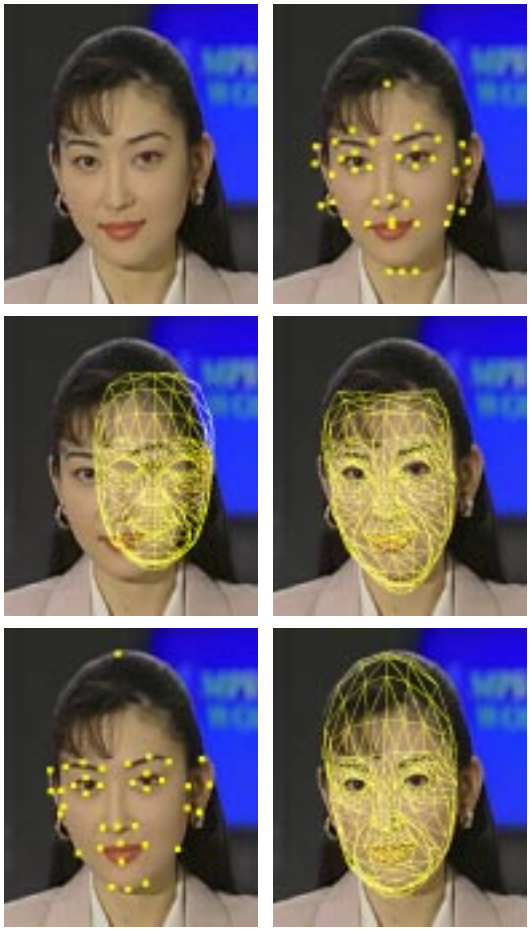

Figure 3 – (a) original image (b) automatically extracted feature points (c) model before adaptation (d) model adapted using the automatically extracted data (e) feature points after refinement by user interaction and (f) corresponding adapted model

In conclusion, this paper describes an interactive analysis system able to extract calibration data to perform the configuration of 3D facial models in the context of the emerging MPEG-4 audiovisual coding standard. More results and a user friendly software application implementing the analysis system here described will be available at the time of the conference.

References

[1] F.Parke, K.Waters, "Computer facial animation", A. K. Peters

[2] R.Koenen, F.Pereira, L.Chiariglione, "MPEG-4: context and objectives", Image Communication Journal, vol. 9, n°4, May 1997

[3] MPEG Video & SNHC, "Final Text of ISO/IEC FCD 14496-2 Visual", MPEG Tokyo Meeting, March 1997

[4] J.B.Waite, W.J.Welsh, "Head boundary location using snakes", British Telecom Technology Journal, Vol.8, n°3, July 1990

[5] N.Otsu, "A threshold selection method from grey-level histograms", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-9, n°1, January 1979

[6] M.Kampmann, J.Ostermann, "Autmatic adaptation of a face model in a layered coder with an object-based analysis-synthesis layer and a knowledge-based layer", Signal Processing: Image Communication, Vol.9, pp. 201-220, 1997

[7] C.Nightingal, R.Hutchinson, "Artificial neural nets and their application to image processing", British Telecom Technology Journal, Vol.8, n°3, July 1990

[8] M.Turk, A.Pentland, "Eigenfaces for recognition", Journal of Cognitive Neuroscience, Vol.3, n°1, 1991

[9] G.Abrantes, F.Pereira, "An MPEG-4 SNHC Compatible Implementation of a 3D facial animation system", IWSNHC3DI'97, Rhodes-Greece, September 97