Analysis-Synthesis Cooperation for MPEG-4 Realistic Clone Animation

J.-L. Dugelay, A. C. Andrés del Valle

Institut Eurécom. 2229, route des Crêtes
06904 Sophia Antipolis - France
Email: {andres, dugelay}@eurecom.fr

ABSTRACT

This article presents a complete analysis-synthesis scheme for realistic face animation. By analyzing real-time video sequences we obtain Face Animation Parameters (FAP) to animate a highly realistic 3D head model. We describe how to introduce a tight cooperation between analysis and synthesis for the face movement analysis to improve the realism in the results.

1. INTRODUCTION

Face cloning has become a need for many multimedia applications where human interaction with virtual and augmented environments enhances the interface. Its promising future in different areas such as mobile telephony, the Internet, etc. has made of it an important subject of research. Proof of this interest is the increasing appearance of companies offering their customers the creation of customized synthetic faces and the government support through public grants like the European Project INTERFACE [1].

We can classify synthetic faces in two major groups: avatars and clones. Avatars generally are a rough or symbolic representation of the person. Their appearance is not very accurate. They are speaker-independent because their animation follows general rules independently of the person that they are assumed to represent. Most of the current commercial synthetic faces fall in this category. In some applications, avatars do not completely please people because they create a feeling of mistrust [2]. Clones are more realistic and their animation takes into account the nature of the person; they are speaker-dependent. Opposite to avatars, which are generated from more or less simple mathematical models, realistic models are difficult to build without using specialized equipment (e.g. Cyberware) and equally complicated to deal with in real time due to their complexity.

There exist several ways to synthesize face animation. Currently two major trends are found in the literature: image or feature reconstruction (model-based) and 3D-model movement synthesis. Systems that synthesize face animation from recorded video images belong to the first group [3][4]. The second group includes all those animation techniques applied on 3D head meshes to replicate human movements. We can obtain animation by simulating the face tissue and muscles [5][6], or we can define more or less complex parametric models, where each Action Unit or Face Animation Parameter (FAP) is related to the movement of some specific nodes of the head mesh [7].

Face expression analysis is mostly performed to obtain semantic information, not giving any indication of how this information should be synthesized. A complete detachment between analysis and synthesis does not lead to realistic results. Some analysis techniques utilize a 3D face model to provide feedback for movement understanding [8]. Although those techniques seem to use an analysis-synthesis cooperation, this cooperation is very limited because they do not explicitly synthesize the results. Piat and Tsapatsoulis [9] have developed a system to analyze face expressions along the time, roughly deducing how the expression is generated in terms of FAP. In [3] L. Yin uses on-line input to generate the expression synthesis. These approaches show some promising analysis-synthesis cooperation techniques.

In this paper we discuss a complete analysis-synthesis face animation system. From real video input we analyze some specific features to generate on-line FAP that we apply onto a realistic MPEG-4 compliant head model. This speaker-dependent system does not yet generate articulated animation

This work has been partially supported by FRANCE TELECOM R&D.

but provides realistic face movements extracted from video sequences. The video-to-FAP converter can be used to automatically generate MPEG-4 compliant face animation streams; therefore other clones or avatars can synthesize the expressions.

2. OUR ANALYSIS-SYNTHESIS COOPERATION

We propose analysis-synthesis cooperation techniques that lead to algorithms that deduce face parameters from video sequences in real time. These algorithms work under any lighting conditions and analyze faces that do not have any kind of special makeup or markers on. Our techniques study the video on the image domain, differing from those techniques that need 3D information while analyzing. Since the analysis generates FAP, we can apply them onto the synthesis system to use the resulting animation as feedback.

This analysis-synthesis cooperation is made possible thanks to the highly realistic head models we use. Our first 3D models were generated from Cyberware^{TM} scanned data. At the moment, a less costly system for face data acquisition is in progress [10].

To generate a fast analysis-synthesis cooperation, we need a parametric, easy and flexible synthesis module. We are developing an MPEG-4 compliant face animator. MPEG-4 introduces the concepts of object and scene into video coding. The real or synthetic objects are the elements that compose a scene. In general, the mesh of a synthetic object is enough to completely define it, and then we animate it by sending some coded actions that transform its nodes. Due to the importance that face synthesis may have on some multimedia applications, the standard includes a concrete definition for the face object. The norm specifies the decoding of FAP and some predefined nodes (FDP) that the synthetic head model must contain. The face models we use already contain these required nodes and our animation is described in terms of face animation parameters. These parameters may be directly part of the 68 FAP specified by MPEG-4 or they can be decomposed in a combination of them.

MPEG-4 suits our system because the decoding of the animation has been specified in such a way that leaves complete freedom for the animation design. It allows us to build an efficient cooperation analysis-synthesis because we define the animation from the analysis results and we are able to use the synthesis to improve the analysis. The standard permits customized animation per clone. It also ensures compatibility with other animation systems and the proper integration of our animated clones in virtual or hybrid multimedia environments.

3. FACE MOVEMENT ANALYSIS

3.1. General scheme

We consider face expression comprehension from a video sequence as a function of the general pose of the face on the sequence, the illumination conditions under which the video is recorded and the expression movement. To obtain FAP from video frames, we first study the illumination conditions of the face in the sequence; this information will enable our algorithms to work under any lighting. Then, we estimate the pose of the face obtaining translation and rotation parameters. Finally, we extract some specific features from the face and we apply on them some dedicated analysis techniques on them to obtain face animation parameters.

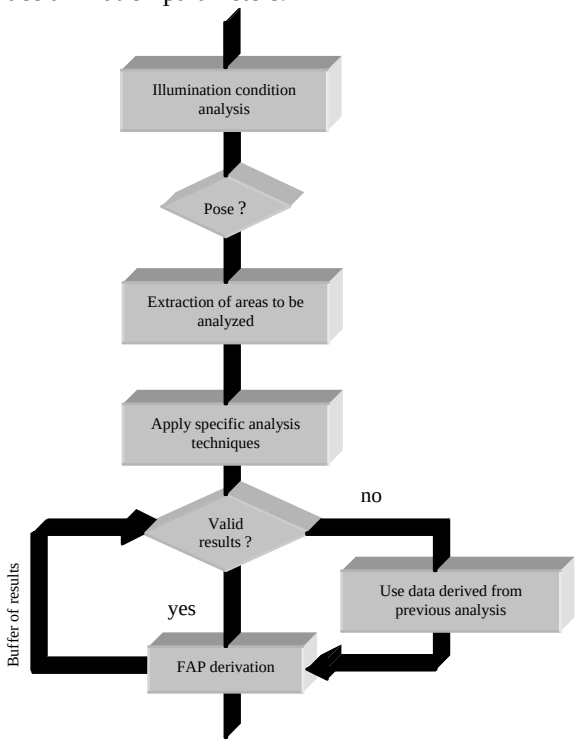

Figure 1: Diagram of the analysis procedure.

Synthesis cooperation can be applied at different stages of the analysis chain. Currently, we use the synthetic output to predict the next face pose in the video sequence; we utilize the synthetic clone to generate expression databases and clones also substitute people in any possible training steps that the system may need.

3.2. Real-time face analysis techniques

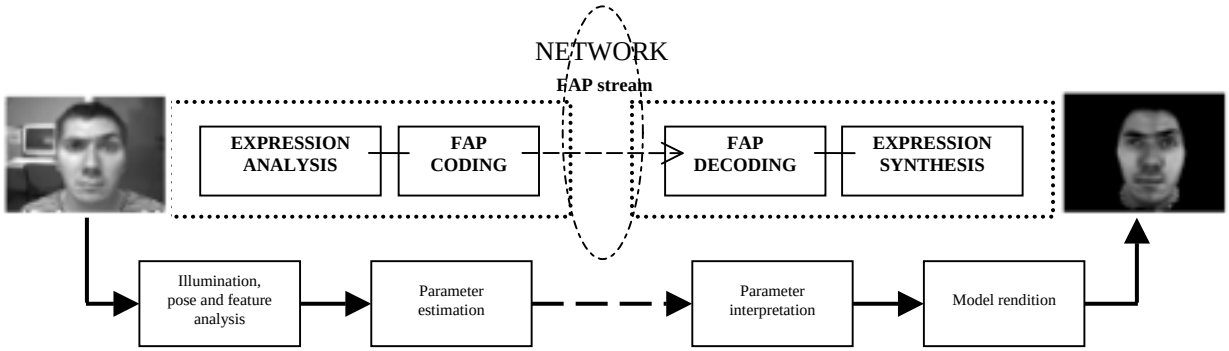

Figure 2: This scheme shows the complete transmission process that includes the analysis and the synthesis of expressions.

To obtain the global pose of the synthetic model we have developed a tracking algorithm that profits from a high analysis-synthesis cooperation. This algorithm utilizes a feedback loop. The synthesized image from the clone is compared to the image of the face in the sequence to extract some 2D information. We feed a Kalman filter with this information and the filter predicts the translation and rotation parameters to apply onto the synthetic clone, whose image is again compared with the following frame. This algorithm analyzes the model and the video sequence at the image level therefore we have to previously perform some light compensation on the synthetic model to adapt it to the lighting conditions of the video sequence. More details about this part can be found in [11]. The algorithm also tracks the location of the most interesting features (eyes, eyebrows and mouth) on the video.

Our first approach for the expression analysis of the features was an algorithm that compares the image of the feature with a database of images of that feature recorded under different lighting conditions. The realism of our clones permits generating such a database from the model. We perform some training to relate the comparison scores with combination of FAP. During the training, the clone synthesizes some expressions and we build an estimator that relates the FAP we use to synthesize them with the score they produce when compared with the images of the database. To ensure lighting independence we perform some pre-processing on the images we compare. This technique has proved to give fair results but the performance depends on the complexity and the size of the database. Storing the images of all possible lighting conditions, global pose situations and FAP combinations becomes unbearable for features like the mouth and the eyes where expressions can become quite complex. Nonetheless, this approach is very suitable for eyebrow movement analysis for example [12].

Only considering one technique shows not to be optimal for the analysis of the expressions of any feature. At the moment we are developing other analysis techniques to suit each feature: extracting the eye movements based on the energy distribution of the image we obtain from the tracking, possible lip-reading algorithms for the mouth, and so on. We envisage the use of the synthetic clone to calibrate all those parameters of the analysis (λ) that may be person dependent. This way, general FAP (μ) could be customized via some transformation (denoted as α in Fig. 3).

These algorithms have not been yet developed considering the possible coupling interference that the expression can have upon the global pose estimation. Global tracking and expression analysis have been tested separately. Future experiments will perform the tracking and the expression analysis together to estimate the influence of the feature changes on the global pose detection algorithm.

4. REALISTIC SYNTHESIS

As stated in Section 2, a high analysis-synthesis cooperation is possible thanks to a realistic face synthesis. Our clones are not only a precise physical representation but also try to replicate the exact movement of the speaker. Opposite to how avatars behave, the FAP that we provide from the video analysis will create customized animation for the clone. The customization is always done through a training process. This training is designed to minimize the user interaction. The proposed system (Fig. 3) will "learn" the specific movements and customize the FAP values during an on-line transmission session (possibly with a delay).

This realism-based scheme for animation generation already acts in our head tracking; eventually will prove to be helpful when dealing with the pose and expression coupling. The analysis-synthesis cooperation takes part before transmission, right when parameters are estimated, see Fig. 2. Needless to say that generated FAP,

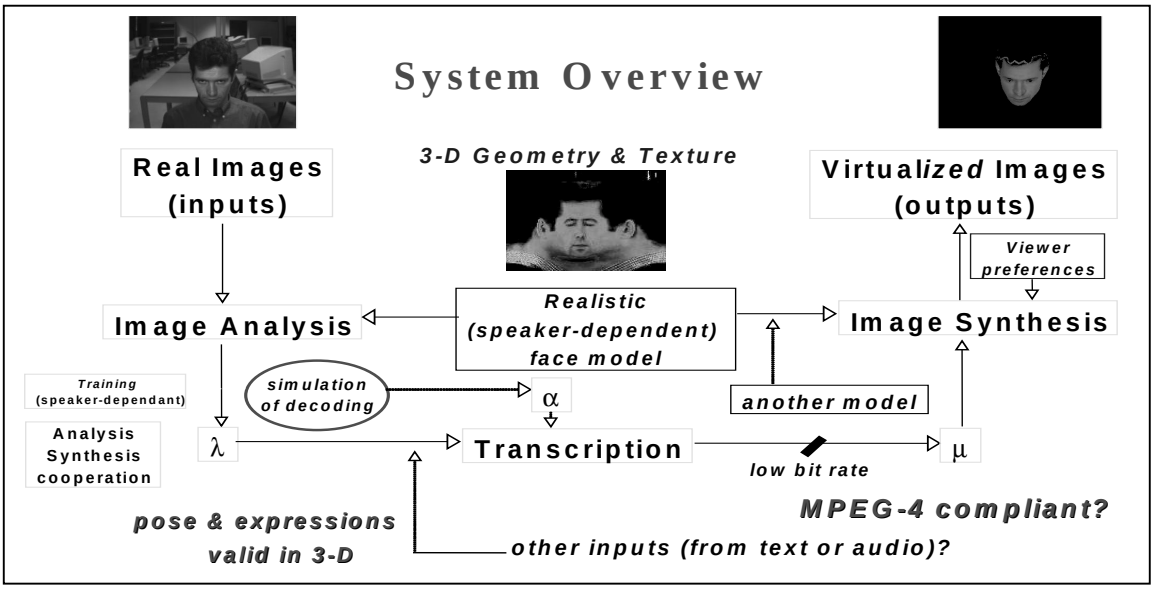

Figure 3. Overview of the analysis-synthesis cooperation in our complete system.

whether they have been customized or not, can be used as an input stream for any other clone or avatar. The requirement of highly realistic synthesis is kept for the encoding part of the system.

4. CONCLUSION

Realistic clone animation needs a strong speaker dependent cooperation between analysis and synthesis. In this article we have presented a complete system that obtains face animation parameters (FAP) from video sequences for MPEG-4 compliant face animation. Our algorithms exploit the use of highly realistic clones for an enhanced analysis-synthesis cooperation where they substitute the real user in all of the training stages.

5. REFERENCES

[1] IST-European Project: INTERFACE IST-1999-10036. http://www.cordis.lu/ist/projects/99-10036.htm

[2] J. Ostermann and D. Millen. "Talking Head and synthetic speech: an architecture for supporting electronic commerce". In *ICME 2000*, New York City, NY, August 2000.

[3] E. Cosatto, G. Potamianos and H.P. Graf. "Audio-visual selection for the synthesis of photo-realistic Talking-Heads". In *ICME 2000*, New York City, NY, August 2000.

[4] L. Yin and A. Basu. "Partial update of active textures for efficient expression synthesis in model-based coding". In *ICME 2000*, New York City, NY, August 2000.

[5] D. Terzopoulos and K. Waters. "Physically based facial modeling analysis, and animation". *J. Visualization and animation* 1(2), pp. 73-80, 1990.

[6] G. Breton, C. Bouville and D. Pelé. "FaceEngine, un moteur d'animation faciale 3D destiné aux applications temps réel". In *CORESA 2000*, Poitiers, France, October 2000.

[7] F. Lavagetto and R. Pockaj. "The facial engine: toward a high-level interface for design of MPEG-4 compliant animated faces". *IEEE Transaction on circuits and systems for video technology*, Vol. 9, No. 2, pp. 277-289, March 1999.

[8] P. Eisert and B. Girod. "Analyzing facial expressions for virtual teleconferencing". *IEEE Computer Graphics and Applications*, pp. 70-78, September 1998.

[9] F. Piat and N. Tsapatsoulis. "Exploring the time course of facial expressions with a fuzzy system". In *ICME 2000*, New York City, NY, August 2000.

[10] E. Garcia, J.-L. Dugelay, H. Delingette. "Low Cost 3D Face Acquisition and Modeling". In ITCC2001, Las Vegas, 2-4 April 2001,

[11] S. Valente and J.-L. Dugelay. "Face tracking and realistic animations for telecommunicant clones". *IEEE Multimedia Magazine*, pp. 34-43, February 2000.

[12] S. Valente and J.-L. Dugelay. "A visual analysis/synthesis feedback loop for accurate face tracking". *Image Communication*, Vol. 16, No. 6, pp. 585-608, February 2001.