# Evaluating Face Models Animated by MPEG-4 FAPs

Jörgen Ahlberg, Igor Pandzic, Liwen You
{ahlberg, igor}@isy.liu.se
Image Coding Group, Dept. of Electrical Engineering
Linköping University, SE-58183 Linköping, Sweden
http://www.icg.isy.liu.se

## ABSTRACT

We want to investigate how well animated face models can express emotions when controlled by low level MPEG-4 FAPs reproducing the facial motion captured from real persons acting out the emotions. We propose a test procedure for evaluating the expressiveness of a face model and compare it to other face models as well as to real video. The test is reproducible, and the software and data used are publically available. We also propose one relative and one absolute measure by which the results can be measured and different face models be compared to each other.

## 1. INTRODUCTION

Creating animated human faces using computer graphics techniques has been an increasingly popular research topic the last few decades, and such synthetic faces, or virtual humans, have recently reached a broader public through movies, computer games, and the world wide web. Current and future uses include a range of applications, such as human-computer interfaces, avatars, video communication, and virtual guides, salesmen, actors, and newsreaders.

The MPEG-4 standard [4] on Face and Body Animation (FBA) specifies a set of Facial Animation Parameters (FAPs) used to control the animation of a face model. The FAP set contains two high level FAPs for selecting facial expressions and visemes, and 66 low level FAPs. Each low level FAP denotes a small facial motion, and they are closely related to minimal muscle actions. The low level FAPs are expressed as movement of feature points in the face, and MPEG-4 defines 84 such points. The feature points not affected by FAPs are used to control the static shape of the face.

The FAPs can be efficiently compressed and included in an FBA bitstream for low bitrate storage or transmission. An FBA bitstream can be decoded and interpreted by any MPEG-4 compliant face animation system [3, 5, 2], and a synthetic, animated face be visualised.

Despite the fact that the MPEG-4 standard has existed for a few years now, as have several MPEG-4 compliant face animation systems, there has been little effort made to evaluate the standard (or the face models used) in terms of subjective quality of reproduced facial motion. Even if it is clear that the actual feature point motion can be reproduced, it is not evident that the synthetic faces have the expressiveness of real, human faces. On the contrary, synthetic faces can rarely convey emotions as well as a real faces. Yet, this is of great importance for the deployment of synthetic faces if they are going to be widely used in any of the applications mentioned above. For example, experiments show that animated human faces add value to computer interfaces, provided that the face animation is good enough [6]. An unnatural-looking face animation might instead worsen the impression given by a computer interface.

We propose in the paper an evaluation procedure for measuring the subjective quality of a face model in terms of how well it can convey emotions (via facial expressions) to human observers. The procedure is reproducible, and all our test data  is publically available. In the following sections, the goal of the experiment is defined, the acquisition of the test data and the execution of the test are described, and a relative and absolute quality measures for face models defined. Since the author encourage others to repeat the test with their own face models, one section explains how to accquire our test data and reproduce the experiment.

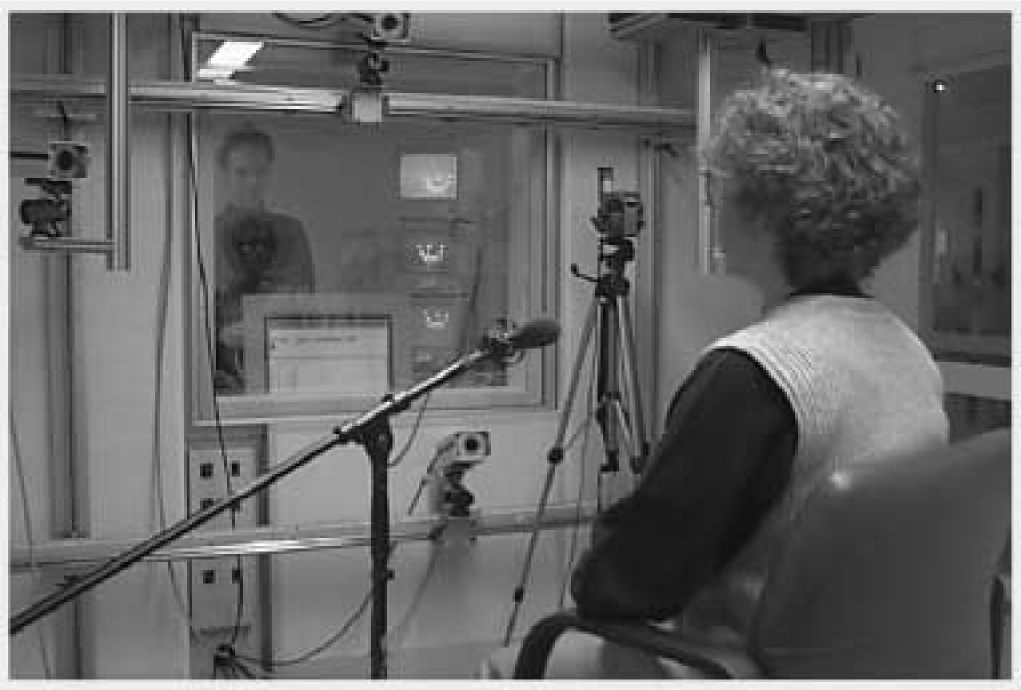
## 2. THE PURPOSE OF THE TEST

We want to investigate how well animated face models can express emotions when controlled by low level MPEG-4 FAPs reproducing the facial motion captured from real persons acting out the emotions. The expressiveness should be judged by the accuracy rate of human observers recognizing the facial expression being shown. We want to relate the result to the *ideal case*, the *random case,* and the *real case*. The ideal case is when all emotions are correctly recognized, and the random case when the recognition results are completely random, that is, drawn from a uniform distribution. The real case corresponds to the recognition rate for the real faces recorded on video, i.e., the faces whose motion the synthetic faces try to reproduce. The recognition rate for each face model is supposedly better than the random case but worse than the real case.

## 3. CREATING TEST DATA

In order to perform the test, it was necessary to record video sequences of persons acting to show different emotions. Additionally, the corresponding FAP sequences were required, so that synthetic sequences reproducing the facial motion from the real video sequences could be created. Thus, the 3D motion of the head and a subset of the MPEG-4 facial feature points had to be tracked.

For this tracking, we used the head tracking equipment at the Dept. of Computer and Information Science owned by Telia Research AB. To track the feature points with



igure 1: The recording setup. The person whose facial motion is to be recorded is seated in a chair. Four infrared cameras (oneis idden behind the person), an ordinary video camera, and a microphone are pointed at the person's face. In front of the person is a creen showing the text and emotion to be read/acted by the person.

the system, markers have been attached to the faces (at the feature points to be tracked) of a few actors. The markers are a few millimeters wide and reflect infrared light, making them visible and easily trackable by IR-sensitive cameras. The system uses four cameras to recover the full 3D motion of each feature point, operating at about 50 frames per second when tracking 30 markers, and somewhat lower for more markers. To recover the rigid 3D motion of the head, special glasses with five markers on were worn by the actor. The setup is illustrated in Figure 1.

First, one *calibration sequence* has been recorded for each actor. For this sequence, the actors wore markers corresponding to the 48 of the MPEG-4 feature points. The purpose was to measure the shape of each actors head, in terms of 3D coordinates for the facial feature points, and thus be able to compute the FAP Units (see below). For the calibration sequence, the actor was still for a few seconds and then turned his/her head slowly from left to right.

For the *animation sequences*, several markers corresponding to feature points not affected by FAPs were removed in order to improve the framerate of the tracking system. The remaining 27 feature points included feature points around the mouth, eyebrows, tip of the nose and on the glasses (for the rigid motion).

The actor then showed facial expressions corresponding to the following seven emotions: fear, anger, surprise, sadness, boredom, happiness, and neutral. The actor expressed each emotion a few seconds, interspaced with neutral facial expression. Also, the actor showed the above expressions while reading a sentence. The sentence chosen was “The small room was completely empty”. This sentence was carefully selected as one being easy to say in each of the above emotional states – it is easy to image being surprised as well as angry etc. due to the small room being empty.

Thus, fourteen animation sequences of head and feature point motion were recorded for each actor. Simultaneously, the actors were also filmed using an ordinary video camera, thus creating the *real sequences* for the test. The procedure was repeated with six actors, which should make 84 sequences. Unfortunately, it showed later when processing the data that the tracking system failed a few times, leaving us with around 50 sequences.

The people used to create the test data were not professional actors, but employees at the Dept. of Electrical Engineering. Thus, many sequences were found to be useless as they contained absolutely no recognizable expressions. After removing those, 21 sequences were left.

### 3.1 Creating FAPs

The 3D coordinates for each marker were, in each frame, identified with a facial feature point. From the five markers on the glasses, the 3D rigid head motion (3D rotation and translation) was computed. The head motion was removed from the other feature points' coordinates, leaving the relaitive, or local, motion. In the calibration sequences, where the local motion was close to zero, the feature point coordinates were used to create the FAP Units (FAPUs), which are the units in which the FAPs are to be expressed. For example, FAP #3, jaw_drop, is to expressed in the unit MNS, mouth-nose distance.

When the FAPUs have been computed for each actor, the feature point coordinates from animation sequences have been used to compute FAP sequences. Also, the head rotation is needed to compute FAPs #48 – #50.

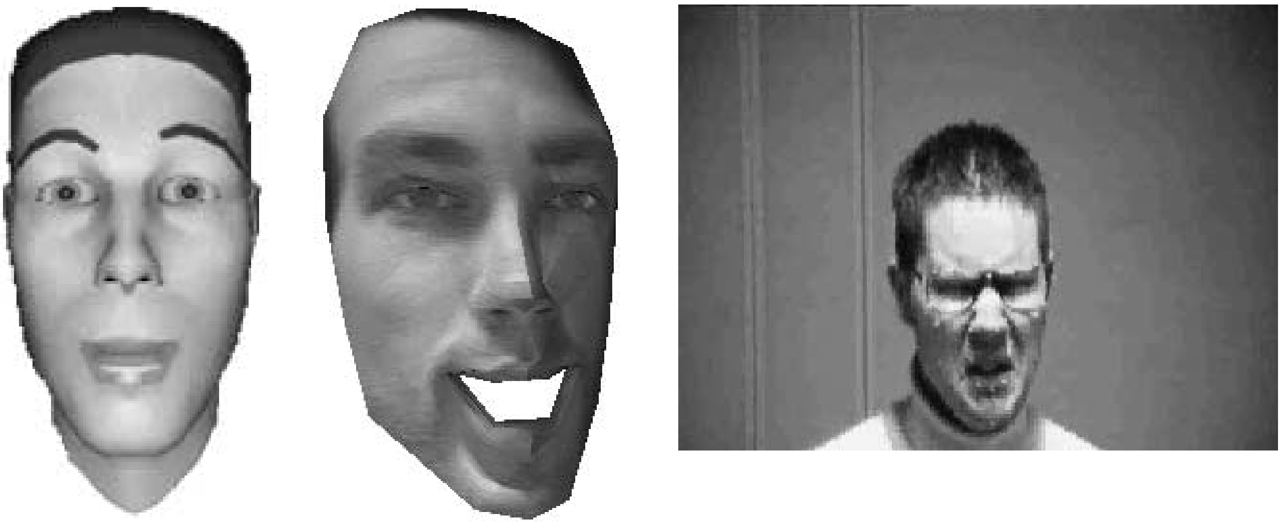


*Figure 2: The Oscar model (left), the Jorgen model (middle), and one frame from a real sequence (right).*

### 3.2 Creating video sequences

As mentioned above, two types of digital video sequences were created; real and synthetic ones. The real sequences were created by digitizing the video sequences recorded by he ordinary video camera while doing the face and facial feature tracking. The synthetic sequences were created by the two face animation systems to test with their two respective face models. For this experiment we have chosen the Facial Animation Engine (FAE) using the face model Oscar [3] and the MPEG-4 Facial Animation Applet (MpegWeb) using the face model Jorgen [5]. The FAP sequences created from the face and facial feature tracking were input into those two systems, producing 42 synthetic sequences.

## 4. Performing the Test

To evaluate the sequences, a number of subjects watched the real and the synthetic sequences, trying to recognize the facial expression the actors/face models were showing. The experiment was performed so that a group of subjects entered a room with a video projector and a canvas. Each test subject was given a (paper) form, where each video should be judged as showing one of the facial expressions. When ready (equipped with a form and a pencil), the subjects were shown the instructions and a few training video sequences. Then, for each video sequence, the following was shown:

1. A screen telling that video sequence number *n* is to be shown.
2. The video sequence one or more times. The number of times was determined so that the total playing time for each sequence is approximately 10 seconds.
3. A screen telling that video sequence number *n* was shown and that it is time to fill in the form at row *n*.

To make the experiment practical, it was automated as much as possible. Using a SMIL script [8], playable by RealPlayer [7], all the instructions and video sequences to be shown to the subjects were ordered and timed in advance. The SMIL scripts were generated automatically by a Matlab-program, taking as input the number of models to be tested, the number of subject groups etc.

Each video was watched by more than 100 subjects in this way.

## 5. Evaluating the Results

The results of the face models' ability to convey emotions should be measured in an absolute way as well as be put in relation to the performance of the real videos. We have thus chosen the error measure computed as described below.

First, compute the dispersion matrix for each model and for the real videos. Those three test matrices (the two synthetic ones and the real one) should be compared to the ideal dispersion matrix and the random matrix, corresponding to perfect recognition and totally random recognition.

The distance from a test matrix to the ideal matrix is computed as the $L1$-norm of the difference, and is then scaled so that the random matrix get the performace value zero and the ideal matrix performance 100. We call this measure the Abolute Expressive Performance (AEP), expressed as

$$AEP(X) = 100 \cdot \frac{\|R - I\|_1 - \|X - I\|_1}{\|R - I\|_1} \qquad (1)$$

where $X$ is the (synthetic or real) test matrix, $R$ is the random matrix, and $I$ is the ideal matrix. The $L1$-norm of a matrix is defined as

$$\|X\|_1 = \sum_{i,j} |X_{ij}| , \qquad (2)$$

where $X_{ij}$ is the element at row *i* and column *j* in the matrix *X*.

The relative measure of a face models' expressive performance is defined as

$$REP(X) = 100 \cdot \frac{AEP(X)}{AEP(X_{\text{real}})}. \qquad (3)$$

The results for our experiment shown in Table 1.

*Table 1: Absolute and Relative Expressive Performances of the models and the real video.*

| Animation | AEP | REP |
|---|---|---|
| Oscar model (FAE) | 9.1 | 15.6 |
| Jorgen model (MpegWeb) | 9.4 | 16.2 |
| Real video | 58.1 | n/a |

### 5.1 The significance of the measurement

To evaluate the validity of the results, a dispersion matrix has also been computed for each subject watching the sequences. Thus, the standard deviation can be estimated and a t-test for statistical significance [1] be done. It is found that the difference between the two face models' performances is not statistically significant on any level, but that the differences between the models and the random, real and ideal case are signifcant with a very high level of confidence (more than 99%).

## 6. Reproducing the Test

It is the authors' intent that this experiment should be easily reproducible by anyone wanting to test their face model and/or face animation system, thus offering a standardized way of measuring and comparing its subjective quality. Thus, all the files necessary for the performing the test are publically available, together with instructions how to use them. Included in the package are all the video files, the real as well as the synthetic ones, the FAP-files (in both ASCII form and binary form) for generating new synthetic sequences, and the scripts for generating SMIL files and the instruction screens (shown before and between the videos).

To reproduce the test, the package will be available for downloaded from the website of the Image Coding Group at Linköping University. Then, new synthetic videos should be generated (using the FAP files) using the face model to be evaluated, as well as new SMIL files. The SMIL files are generated by a Matlab-script, and input parameters are the number of groups of subjects, how long time for each test, and if a new model (new synthetic sequences) is added to the ones included in the package. Detailed instructions are included in the package.

## 7. Conclusion

From our experiments, it is clear that the face models we have evaluated have a far worse expressive performance than the real sequences, but no significant difference could be measured between the two models. Our main result is a the description of a reproducible test that anyone can perform to evaluate their face models.

## Acknowledgement

The authors thankfully acknowledge the financial support from the national Swedish SSF project VISIT (VISual Information Technology). The authors would also like to thank Telia Research AB and the Natural Language Processing Lab at the Dept. of Computer and Information Science at Linköping University for using their time and equipment, particularly Mustapha Skhiri for his help in acquireing and processing the raw data. We would also like to acknowledge Jacob Ström, formerly at Linköping University, now at Ericsson Media Lab, for his initial involvement in this project and contribution of ideas (in particular the "small room"-sentence).

## References

[1] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition, Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale, NJ, USA, 1988.

[2] M. Escher, I. S. Pandzic, N. Magnenat-Thalmann, "Facial Deformations for MPEG-4," *Proc. Computer Animation 98*, Philadelphia, USA, pp. 138-145, IEEE Computer Society Press, 1998.

[3] F. Lavagetto, R. Pockaj, "The Facial Animation Engine: towards a high-level interface for the design of MPEG-4 compliant animated faces," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 9, No. 2, March 1999. Software and face model available at www-dsp.com.dist.unige.it/~pok/RESEARCH/MPEG/fae.htm.

[4] Moving Pictures Expert Group, *ISO/IEC 14496, International Standard on Coding of Audio-Visual Objects (MPEG-4).* www.cselt.it/mpeg

[5] I. S. Pandzic, "Talking Virtual Characters for the Internet", *Proc. ConTel*, Zagreb, Croatia, June 2001. MPEG-4 Facial Animation Applet is available at www.icg.isy.liu.se/~igor/MpegWeb.htm.

[6] I. S. Pandzic, J. Ostermann, D. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer Journal*, Vol. 15, No. 7-8, pp. 330-340, Springer Verlag, 1999.

[7] RealPlayer, www.real.com

[8] Synchronized Multimedia Integration Language, W3C Recommendation REC-smil-19980615. www.w3.org/TR/REC-smil/

[9] A. M. Tekalp, J. Ostermann, "Face and 2-D mesh animation in MPEG-4," *Signal Processing: Image Communication*, Vol. 15, No. 4-5 (Tutorial Issue on the MPEG-4 Standard), pp. 387-421, January 2000.