FACE ANIMATION FOR HUMAN COMPUTER INTERFACES

Joern Ostermann, Axel Weissenfeld

Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Universität Hannover, Hannover, Germany

ABSTRACT

MPEG-4 based 3D face animation and image-based face animation are presented. The latter can produce animations indistinguishable from real videos. Subjective tests indicate that animated faces increase the trust that users have in the information presented on the computer. We predict that the use of animated faces will allow e-commerce and e-care web sites to enhance their effectiveness. Animated faces talk to the customer, give advice and suggest further actions in order to help the customer. Architectures for supporting efficiently the use of face animation in real time interactive services are currently available.

1. INTRODUCTION

Nowadays human-machine communication is dominated by text input and mouse clicks of the user, while the machine returns text and graphics. Recent advances in speech recognition, natural language understanding, speech synthesis and face animation give the opportunity to use talking heads as part of a modern human-machine interface. Talking heads combine text-to-speech synthesis (TTS) with facial animation. Due to the talking heads communication between humans and machines will be more natural and therefore increase the attention and trustfulness of humans toward machines [1][11]. Interactive services, such as e-commerce or e-care, will benefit from using talking heads as a modern interface.

In Figure 1 a typical application of facial animation is presented. Here an internet-based customer service integrates a talking head into its web site. Hence, companies can give e-commerce customers their own virtual shopping assistant, which communicates with potential customers in a natural way. Subjective tests showed, that Electronic Commerce Web sites with talking heads got a higher ranking from customers [2] [4].

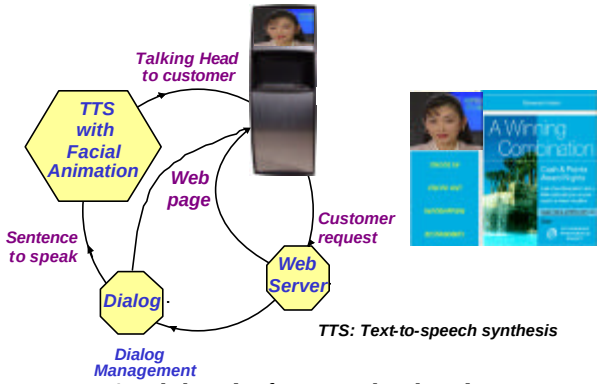

Figure 1: A web-based information kiosk and a customer service site that integrate a web site with a talking head

In the future the sales volume in e-commerce and e-care (customer management relationship) will significantly increase. For instance, e-care will achieve a worldwide market size of 30 billion dollars in 2008 (source: Frost&Sullivan). It is expected that facial animation will be widely used as a modern human-machine interface for interactive services.

Many researchers studied methods for modeling and animating faces based on different approaches ranging from animating 3D models to image-based rendering. In this paper the MPEG-4 facial animation will be introduced as the first official standard for facial animation. Moreover, facial animation with image-based rendering will be discussed, because this approach offers photo realistic talking heads.

2. MPEG-4 FACIAL ANIMATION

The facial animation in MPEG-4 is based on animating a 3D-mesh, which characterizes the shape of the face [3][8][9]. The 3D-mesh consists of nodes from which 84 are defined as feature points as presented in Figure 2.

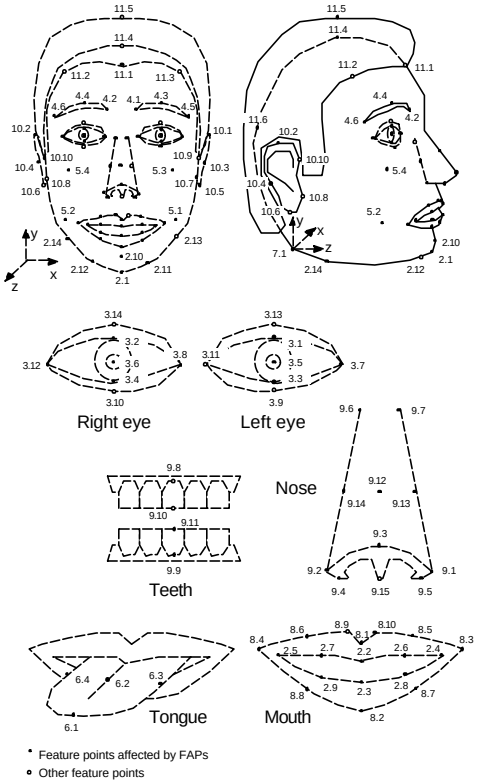

Figure 2: Feature points define the shape of a proprietary face model. The facial animation parameters (FAPs) are defined by motion of some of these feature points (from [3])

In order to animate this facial model 50-feature points can be moved by 68 different facial animation parameters (FAPs). Lowlevel FAPs move individual feature points and their surroundings, high-level FAPs define facial expressions (Figure 3) and visemes, which are mouth shapes that correspond to the phonemes of the spoken words.

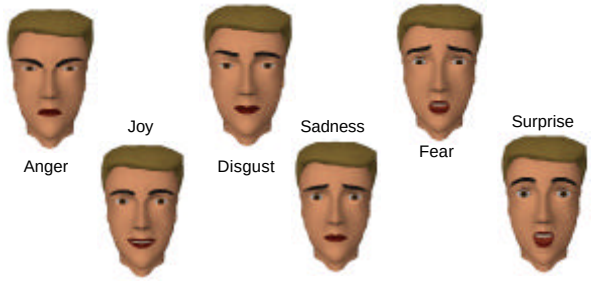

Figure 3: Facial Expressions [3]

For interactive services, face animation can be integrated with a text-to-speech synthesizer (TTS). The phonemes synthesized by the TTS are converted into mouth shapes that are shown synchronously with the synthesized speech thus creating the impression of a talking head (Figure 4).

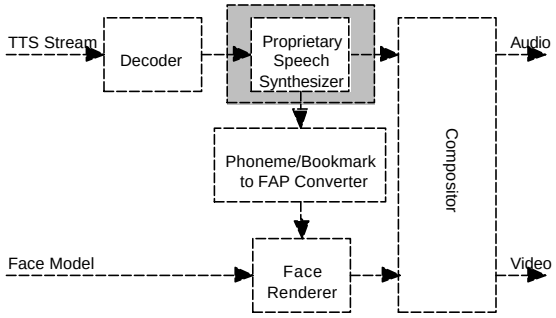

Figure 4: Block diagram showing the integration of a proprietary Text-to-Speech Synthesizer into the facial animation part of an MPEG-4 decoder [13]

The TTS stream (Figure 4) contains the text to be synthesized as well as bookmarks for controlling facial expressions. The encoder has to encode and send only the TTS stream, if a Phoneme/Bookmark-to-FAP-Converter derives the facial animation parameters from the phonemes and related timing information provided by the TTS synthesizer as presented in Figure 4. In this way a talking head can be animated with a data rate of less than 200 bit/s [3]. Thus, this low data rate can be transmitted over narrow channels. If the encoder also sends the FAPs to the decoder instead of the decoder deriving them from the text to be synthesized, then only the facial animation requires a data rate of up to 2000 bits/s [3].

For e-commerce applications, the encoder may transmit a face model with its Facial Animation Tables (FATs) defining the animation rules for FAPs into the receiver and animate this model with its defined look and behavior [14]. However, MPEG-4 requires the face model to be represented with a 3D mesh and one optional texture map. Therefore, the face animation will always look artificial.

3. FACIAL ANIMATION WITH IMAGE-BASED RENDERING

Facial animation with image-based rendering generates photo realistic animations of talking heads using a database of images from a recorded talent [5][10][12]. Therefore the image-based model looks like a 'real person' as shown in Figure 5.

Image-based rendering processes only 2D images and does not require a 3D model. Facial animation with image-based rendering consists of two main steps: Audiovisual analysis of the recording of the talent and

synthesis of photo realistic facial animation. In the analysis step a database with images of deformable facial parts of the talent is collected, while the audio file is segmented into phonemes. The database images are annotated with feature descriptions like mouth height and width as well as with the audio context when the image was recorded. The audio context includes the phonemes spoken when the image was recorded.

Figure 5: Image-based model

A face is synthesized by first synthesizing the audio from the text using a TTS. The TTS sends phonemes and their timing to the face animation engine. The face animation engine overlays facial parts corresponding to the generated speech [6][7] over a background video sequence with typical short head movements.

In Figure 6 the different steps of image-based facial animation are illustrated. During the analysis step the appropriate mouth posture 5.b is extracted from the recorded sequence 5.a and stored in a database. The synthesis of facial animation playbacks a background sequence 5.d. A mouth sequence has to be matched to the background sequence, so that the mouth posture fits to the spoken output. The appropriate mouth posture is retrieved from the database and the pose of the head of the background sequence is calculated. Then the mouth sample is matched to the previously calculated pose 5.c. Finally the mouth sample 5.c and the background sequence 5.d are combined and the animated face is obtained 5.e. Subjective tests show that the synthesized video cannot reliably be distinguished from recorded video [7].

People expect to interact with realistic faces in serious applications such as e-commerce or e-care. The advantage of image-based rendering is the photo realistic appearance of the talking head, so that customers can interact with 'real persons'. Consequently, e-commerce and e-care web sites, which integrate an image-based talking head, will be more successful.

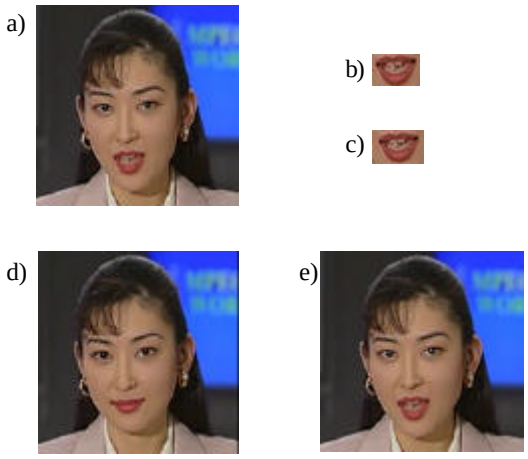

Figure 6: Step-by-step processing of analysis and synthesis for image-based facial animation

4. EXPERIMENTAL RESULTS

Subjective tests measuring the trust between a human user and a computer are shown in Fig. 7. The 'Social Dilemma' game is an established tool for measuring trust [11]. In the 'text' case, the computer used text to interact with the human, in the 'TTS' case the computer used text on the screen as well as synthesized speech, and in the talking head case the computer used a talking head consisting of synthesized speech with facial animation in addition to the text for communication with the user. The talking head was a cartoon-like face as shown in Fig. 3.

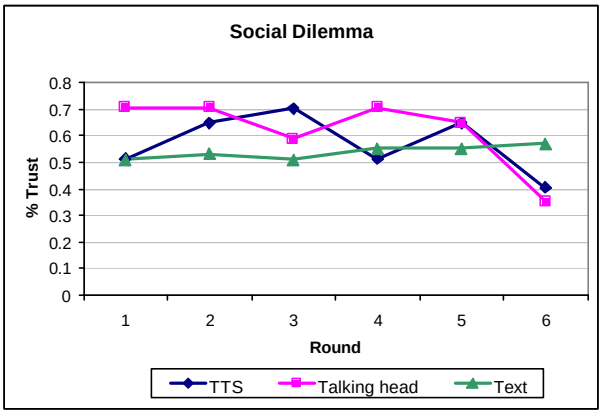

Figure 7: Cooperation rate of a human user as a function of the round of the game. Different modalities representing the computer achieve different cooperation rates

The evaluation of a questionnaire showed a weak trend that users prefer a human computer interface with face animation and TTS. While the results of this questionnaire might not justify the use of face animation, the change in trust from 52% for the text interface to 67% for the talking head interface justifies the use of face animation in e-commerce. Apparently, users are not aware on how they are influenced by a talking face.

5. CONCLUSIONS

MPEG-4 provides tools for face animation using 3D meshes and a texture map. Newer developments in face animation based on image-based rendering enable the synthesis of animated faces that cannot be distinguished from recorded faces. This opens new ways for many applications like automatic newsreaders, electronic commerce, customer service or information kiosks as well as video manipulation.

In subjective tests, we evaluated MPEG-4 face animation simulating a game geared towards measuring trust comparing the performance of interfaces using talking heads, text-to-speech and text. According to the results, the use of talking heads in the design of interactive services was favorably rated for most of the attributes in these experiments. An important result for E-commerce is that users cooperate more with a computer if it is represented with text-to-speech and facial animation instead of TTS only or text only. We measured an increase of the cooperation rate from 50% for text to 70% for a talking head in a first human computer interaction. The average cooperation rate over 5 consecutive interactions increased from 52% to 67%. Based on these results we expect that talking heads will increase the performance of web stores and web-based customer service.

6. REFERENCES

[1] Ostermann, D. Millen, “Talking heads and synthetic speech: An architecture for supporting electronic commerce,” Proc. ICME, pp. MA2.3, 2000.

[2] I. Pandzic, J. Ostermann, and D. Millen, "User Evaluation: Synthetic Talking faces for Interactive Services", accepted for publication in The Visual Computer, Special Issue on Real-time Virtual Worlds, 1999.

[3] J. Ostermann, “Face Animation in MPEG-4”, in MPEG-4 Facial Animation: The Standard, Implementation and Applications, Igor S. Pandzic (Editor), Robert Forchheimer (Editor), Wiley, Chichester, England, 2002, pp. 17-56.

[4] J. Ostermann, “E-COGENT: An electronic convincing agent”, in MPEG-4 Facial Animation: The Standard, Implementation and Applications, Igor S. Pandzic (Editor), Robert Forchheimer (Editor), Wiley, Chichester, England, 2002, pp. 253-264.

[5] E. Cosatto, H.P. Graf, “Sample-Based Synthesis of Photo-Realistic Talking heads,” Proc. IEEE Computer Animation, pp. 103-110,1998.

[6] E. Cosatto, H.P. Graf, “Photo-realistic talking heads from image samples”, IEEE Trans. on Multimedia, vol. 2, no. 3, pp. 152-163, Sept. 2000.

[7] E. Cosatto, J. Ostermann, H. P. Graf, J. Schroeter (2003), “Lifelike Talking Faces for Interactive Services,” Invited Paper, Proc. of the IEEE, Special Issue on Human-Computer Multimodal Interface, Vol. 91, No. 9, pp. 1406-1429, Sept. 2003.

[8] Gabriel Abrantes and Fernando Pereira, "MPEG-4 Facial Animation Technology: Survey, Implementation and Results", IEEE CSVT vol. 9, no. 2, pp. 290-305, 1999.

[9] F. Lavagetto and R. Pockaj, "The facial animation engine: Toward a high-level interface for the design of MPEG-4 compliant animated faces", IEEE CSVT vol. 9, no. 2, pp. 277-289, 1999.

[10] C. Bregler, M. Covell, and M. Slaney, “Video Rewrite: Driving Visual Speech with Audio”, Proc. ACM SIGGRAPH 97, in Computer Graphics Proceedings, Annual Conference Series, 1997.

[11] Van Mulken, S., Andre, E., & Muller, J. An Empirical Study on the Trustworthiness of Life-Like Interface Agents. *Proceedings o HCI International '99 - Volume 2*. (Munich, Germany, August 22-26, 1999), Lawrence Erlbaum, 153-156.

[12] T. Ezzat, T. Poggio, “MikeTalk: A Talking Facial Display Based On Morphing Visemes”, Proc. IEEE Computer Animation, pp. 96-102,1998.

[13] J. Ostermann and M. Beutnagel, A. Fischer, Y. Wang, "Integration of talking heads and text-to-speech synthesizers for visual TTS", ICSLP 99, Australia, December 99.

[14] J. Ostermann, E. Haratsch, "An animation definition interface: Rapid design of MPEG-4 compliant animated faces and bodies", International Workshop on synthetic - natural hybrid coding and three dimensional imaging, pp. 216-219, Rhodes, Greece, September 5-9, 1997.