REALISTIC FACIAL EXPRESSION SYNTHESIS FOR AN IMAGE-BASED TALKING HEAD

Kang Liu and Joern Ostermann

Institut für Informationsverarbeitung, Leibniz Universität Hannover
Appelstr. 9A, 30167 Hannover, Germany
kang@tnt.uni-hannover.de, ostermann@tnt.uni-hannover.de

ABSTRACT

This paper presents an image-based talking head system that is able to synthesize realistic facial expressions accompanying speech, given arbitrary text input and control tags of facial expression. As an example of facial expression primitives, smile is used. First, three types of videos are recorded: a performer speaking without any expressions, smiling while speaking, and smiling after speaking. By analyzing the recorded audio-visual data, an expressive database is built and contains normalized neutral mouth images and smiling mouth images, as well as their associated features and expressive labels. The expressive talking head is synthesized by an unit selection algorithm, which selects and concatenates appropriate mouth image segments from the expressive database. Experimental results show that the smiles of talking heads are as realistic as the real ones objectively, and the viewers cannot distinguish the real smiles from the synthesized ones.

Index Terms— Talking head, image-based animation, facial expression, unit selection

1. INTRODUCTION

The development of modern human-computer interfaces and their applications such as E-Learning, web-based information services and video games has been the focus of the computer graphics community in recent years [1]. These applications will use facial animation techniques combined with dialog systems extensively in the future [2].

The current image-based talking heads [2] [3] are so realistic that the people cannot distinguish them from real videos. However, the talking head is inexpressive, such that the user will be bored, if the talking head has no expressions for a long time conversation.

Facial expressions reflect one's motivation or emotional states, which is important for communicating. Methods [4] [5] for modeling facial expressions have been largely investigated with few attempts to achieve realistic expressions synchronized with speech.

Unit selection synthesis can generate realistic image-based animations by concatenating recorded mouth images in an appropriate order. Using a large database with a large number of units available with different appearance, shape and expression, it is possible to synthesize more natural looking facial animations than 3D-based approaches, because the dynamics of the lips and tongue are difficult to be modeled parametrically. Experiments on the expressive synthesis indicate that the animations depend on the audio-visual database. Facial expressions [2] are represented as template behaviour (expression patterns), for example a welcome smile. These patterns are appended to an expressionless facial animation, which results in an unrealistic talking head, because the expressions cannot be modeled by the patterns only. Moreover, expressions accompanying speech are not investigated.

Facial expression synthesis is not simply defined as mouth animation. Smiles have sometimes an impact on the deformation of the eye parts, although a smile is formed by flexing the muscles near both ends of the mouth [6]. In this paper, we are focusing on the expressive mouth animation, since eye animations, such as eye gazing and eye blinding, are developed in [7], which can be directly integrated in our system. Furthermore, according to perceived smile meanings [8], smiles are categorized in three perceptually distinct types: amused, polite, and embarrassed/nervous. To simplify the facial expression synthesis, the smile type of polite is used in this paper. Other smile types could be synthesized in the same way.

The main contribution of this paper is to synthesize a smiling talking head, which is based on the unit selection synthesis. The expressive unit selection is used to select and concatenate mouth image segments from an expressive database in an optimal way. The expressive database consists of a large number of recorded neutral and smiling mouth images.

The paper is organized as follows: Section 2 presents a system overview. Section 3 introduces the technique to synthesize expressive talking heads. Section 4 gives experimental results. The paper is concluded in Section 5.

2. SYSTEM OVERVIEW

Based on our previous work [3], the talking head system is extended with facial expressions as shown in Fig. 1. The talk-

This work is funded by German Research Society (DFG Sachebeihilfe OS295/3-1). This work has been partially supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

978-1-61284-350-6/11/$26.00 ©2011 IEEE

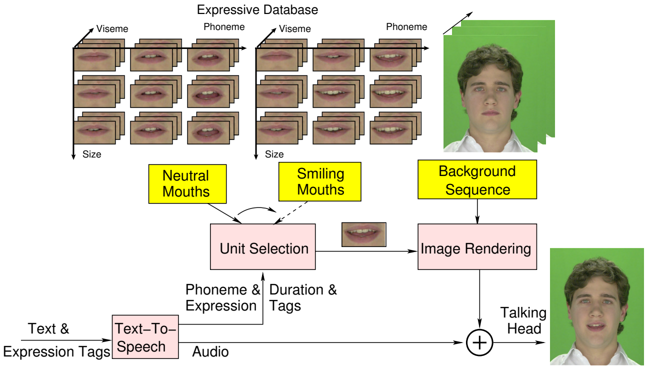


Fig. 1. System overview of expressive talking head synthesis.

ing head system in [3] can generate only neutral faces without any facial expression.

First, a segment of text and expression control tags are sent to a Text-To-Speech (TTS) synthesizer. The TTS provides the audio track as well as the sequence of phonemes and their durations, which are sent to the expressive unit selection. Depending on the phonetic information and input expression tags, the unit selection selects mouth images from the database and assembles them in an optimal way to produce the desired animations with and without expressions. In the context of this paper, units are mouth images in the database. The expressive database contains normalized neutral images and smiling mouth images, as well as their associated features and expressive labels. The unit selection balances two competing goals: lip synchronization and smoothness of the transition between consecutive images. An image rendering module stitches these mouth images to the background video sequence. The mouth images are first wrapped onto a personalized 3D face mask and rotated and translated to the correct position on the background images. Background videos are recorded videos of a human subject with typical head movements. Finally the facial animation is synchronized with the audio, and a smiling talking head is displayed.

3. FACIAL EXPRESSION SYNTHESIS

The general image-based approach to model facial expressions is to build an expressive database, which is composed of different expressions, such as smile, angry, and surprising. In this paper, we are animating smiles as an example, since other facial expressions, like sadness and anger, will be straightforward to integrate.

3.1. Creation of Expressive Database

In our studio, a native speaker is recorded while reading a corpus of 150 sentences. These sentences are designed to find a trade-off between the English phoneme coverage and the size of the corpus. In our experiment, each sentence is recorded for three times with different expressions, i.e. speaking without any expression (neutral speaking), smiling after speaking, and smiling while speaking. The video format is 576×720 pixels at $50\ fps$. The audio signal is sampled at $48\ kHz$. A total of 450 utterances are recorded to build the database, which contains $78,797$ normalized mouth images with a resolution of 288×304 pixels.

After recording, the audio data and the texts are aligned by an aligner, which segments the phonemes of the audio data. Each frame of the recorded video is labeled with the corresponding phoneme and its phoneme context. The phoneme context is required in order to capture coarticulation effects. In our system, we use the American English phoneme and viseme inventory to phonetically transcribe the input text [3].

Using a 3D head scan of the recorded speaker, the recorded videos are processed by model-based motion estimation, which estimates the head motion parameters for each frame. Using the motion parameters, all frames are normalized to a reference position and the mouth regions of all frames are cropped to build an expressive database.

Normalized mouth images are transformed into the PCA space, so that few parameters are needed to describe the appearance of the mouth. The shape of the mouth images is extracted by Active Appearance Models, from which geometric parameters such as mouth width and height are derived.

The database is built with a large number of normalized mouth images. Each image is labeled with the following parameters: an expression tag, appearance parameters (PCA), geometric parameters (mouth width and height), phoneme and phonetic context, original sequence and frame number.

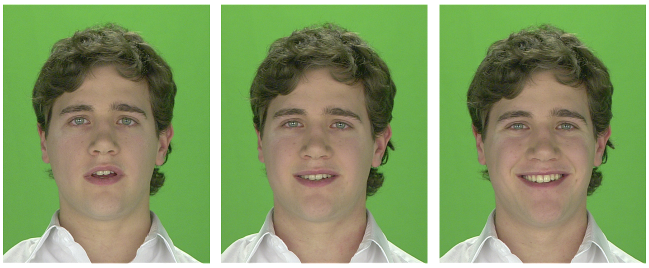
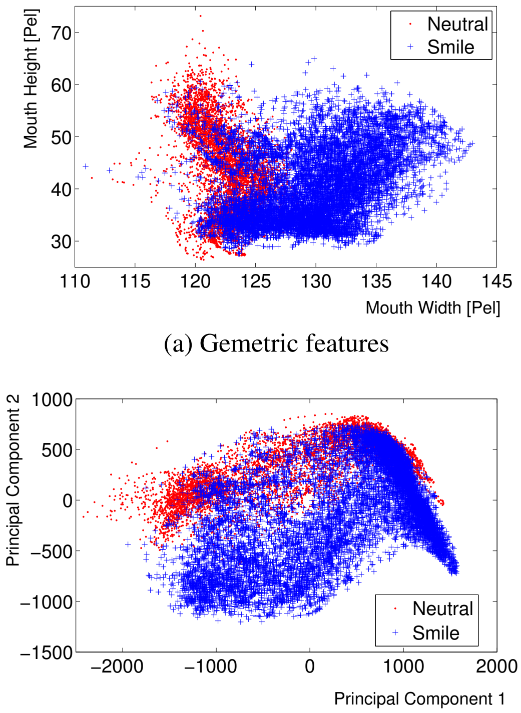


Fig. 2. Smiles with different degree. The left image is a neutral speaking face, the middle is a speaking face with a small smile, and the right is a big smile without speaking.

3.2. Analysis of Viseme Transitions while Changing Expressions

In order to analyze viseme transitions while changing expressions, we firstly analyze the natural expression change of humans, then analyze viseme transitions in the expressive database. The analytical results are used in the expressive unit selection.



(a) Gemetric features

(b) Visual features

Fig. 3. Plots of geometric and visual features for viseme 1 (silence). (a) Distribution of mouth width and height. (b) Distribution of first two components of PCA parameters. Red points represent the neutral mouths, blue points for smile.

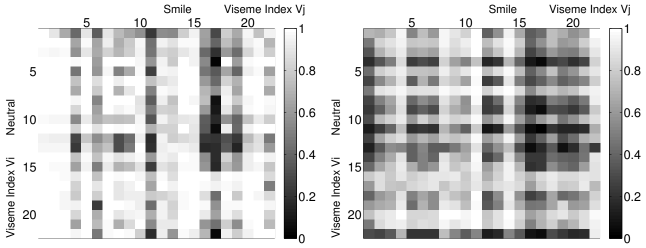
3.2.1. Natural Expression Change of Humans

In order to control facial expressions, the behavior of humans has to be analyzed. In this paper, we conduct facial expression “smile” as an example, and for this reason, we analyze natural smile of humans from recorded videos. The goal of this analysis is to measure when human begins and ends a smile and to which extent human smiles in which situation.

Our facial animation is aiming to generate realistic facial expressions. For this purpose, the recorded video requires natural facial expression while speaking. The recorded videos for database building are unsuitable for analysis of expressions, since the speaker is aware of the recording and their expression is controlled. Therefore, videos of spontaneous speech and conversations are required.

Three video sets are collected to analyze natural smiles:

1. recording of special texts (about 10 minutes): these special texts are able to make the speaker laugh when he is reading, the special texts are like “schuettelverse”, i.e., poems of two lines embedding spoonerisms;
2. spontaneous speech and conversation (about 10 minutes): the most natural facial expression is from the spontaneous conversation, where the speaker is not aware of expression change;



(a)p_{v_i,v_j} from neutral to smile(b) p_{v_j,v_i} from smile to neutral

Fig. 4. Switching matrix between neutral and smile visemes. (a) Viseme switching ability p_{v_i,v_j} from neutral to smile; (b) Viseme switching ability p_{v_j,v_i} from smile to neutral.

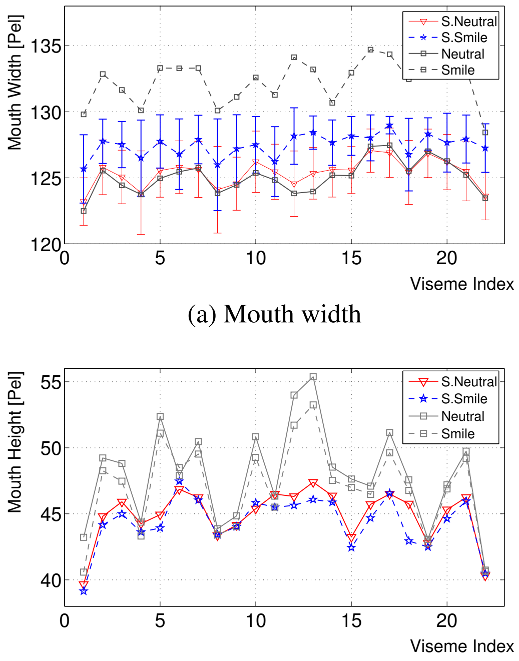
3. videos of news reporter (10 minutes): news reporter is trained to give facial expression, and they knows when and how they smile. News reporter is a typical application for talking head.

Based on the collected videos, we are able to find some simple rules for facial expression. In the first video set, the speaker cannot smile at the begin of the sentence. For example, sometimes the speaker begins to smile at the end of the schuettelverse sentence, because the speaker understands the meaning after reading the sentence. In the second video set of conversation, facial expression depends tightly on current topic and partner’s reactions. Even though for the same topic, different partners have different facial expressions. It is also very difficult to be determined when a smile begins while speaking. To start a smile is really individual, since some one will laugh earlier and some one later. However, several rules are also general in the conversational video. Smile begins before speaking and ends after the whole sentence, which is happened when the speaker tries to answer some funny questions. In some cases, a smile begins somewhere in the sentence, when the meaning changes from neutral to funny. Generally, these changes are almost with a short pause in the sentence. These pauses are always between words or clauses. Someone smiles when he is listening, which gives the partner some information whether the story is funny or he has understood the words very clearly. In the third video set, the news reporter smiles in most cases at the end of the sentence, which gives audience some hints or make the news report friendly.

Another issue is the degree of smile. A smile or a neutral speaking appears at the end of a news, which depends on the speaker. Generally, smiles from the news speaker videos are small and begin from the last sentence of a news. Depending on the content, a big smile or a small one could be at the end of speaking. We have observed that these smiles always start before the last syllable. In most cases, big smiles happen at the listening state without speaking in the conversational videos. Smiles with different degree are shown in Fig. 2.

Towards above mentioned different kinds of smiles, the spontaneous conversational video is used to analyze the statistics. 43% of smiles appear at the last syllable of a sentence. 24% of smiles are accompanying with speaking. 20% of smiles happen between clauses. Big smiles are always happening without any speaking. In a few cases (5% of smiles), smiles are occurring between words.

In order to generalize our smiling talking head to all other facial expression synthesis, we have observed that the transition between different facial expressions, like smile and angry, are through neutral state. In this way, our smiling talking head is easy to be extended to synthesize all other facial expressions. Transition of any two facial expressions is realized by finding a smooth path through neutral mouths.



(a) Mouth width

(b) Mouth height

Fig. 5. Geometric features of the switchable mouths and the whole mouths of each viseme. The red line and the blue dash line is the average width and height of the switchable neutral images (S.Neutral) and the switchable smile images (S.Smile), respectively. The gray curves are the corresponding features of the whole database.

3.2.2. Viseme Transitions of Expressive Database

In order to generate smooth transitions between the neutral and the smiling mouths by using our expressive database, the relationship between the neutral and the smiling mouths has to be analyzed. The goal of this analysis is to show how well the viseme transition can be done for different visemes.

The neutral and smiling mouths have different shapes and appearances. Fig. 3 shows the distributions of the geometric and appearance features of the database for viseme 1 "silence". Given Fig. 3(a), the average width of the neutral mouths is clearly smaller than the average width of smiling mouths, even though the average heights of both viseme 1 are similar. Either the distribution of geometric features of Fig. 3(a) or the distribution of appearance features of Fig. 3(b), both have the same property that the area covered by smiling mouths appears to be a superset of the area of the neutral mouths. Hence, smooth transitions between visemes are possible.

The switching ability from viseme v_i to viseme v_j, is measured in the following way: $p_{v_i,v_j} = \frac{m_{i,j}}{N_{v_i}}$ and $p_{v_j,v_i} = \frac{m_{j,i}}{N_{v_j}}$, where N_{v_i} is the number of mouth images of viseme v_i and N_{v_j} is the number of mouth images of viseme v_j. $m_{i,j}$ is the number of mouth images in v_i, which can find a neighbor from v_j so that the distance is smaller than a predefined threshold. The threshold is chosen so that transitions between two images are smooth, if the distance between two images is smaller than the threshold. Switchable images are images with one expression, which have a similar neighbour with another expression, and their distance is less than the threshold. In our experiments, the threshold is manually determined in the PCA space. Fig. 4 shows the switching matrix of viseme transitions. Due to the higher p_{v_i,v_j} value, switching from neutral to smile is easier than reverse.

Evaluating the whole database, we have measured that 65% of the neutral mouth images are able to find at least one similar smiling mouth to switch to, while only 25% of the smiling mouth images can switch to the neutral mouths smoothly. For each viseme the geometric features of these mouth images are plotted in Fig. 5. In Fig. 5(a), the average mouth widths of the switchable neutral and smiling mouths approach closely. Furthermore, the large overlap intervals of the average mouth width with plus/minus standard deviation indicate that the smiling mouth images with small widths can be switched to the neutral mouths smoothly. In Fig. 5(b), we can see that the average mouth height of the switchable neutral mouths and the switchable smiling mouths is very low, compared to the average mouth height of the whole neutral and smile mouths (gray curves). Therefore, the switch tends to happen at the boundary of a viseme, when the switchable neutral and smiling mouths are almost closed.

3.3. Expressive Unit Selection

The expressive unit selection uses weighted target costs and concatenation costs. The target cost measures the lip synchronization of the mouth image. The concatenation cost measures the smoothness of the transition from one image to another. The target cost is computed by calculating the distance between the phoneme context of the mouth image and the input phoneme context. The concatenation cost of two images is computed by calculating the weighted geometric and appearance distance of these two mouth images. Thus the larger the database is, the more units the unit selection can choose from, which makes it easier to find matching units for a given phoneme sequence. A Viterbi search finds optimal units from the database by minimizing the two types of costs.

Based on the results of analyzing natural expression change, four possible natural talking heads can be synthesized by the expressive database for a given sentence:

- Case 1: talking head speaks without any expression;
- Case 2: smile is accompanying speaking;
- Case 3: facial expression is changed from neutral to smile during speaking;
- Case 4: a smile pattern is appended to an animation for non-verbal communication.

For case 1, the unit selection selects mouth images only from the neutral mouths and for case 2, from the smiling mouths only with no need for switching the database. Case 4 is for non-verbal communications, which is easily done by appending one of the smile patterns to an animation. The focus of the expressive unit selection is on case 3, where the database has to be switched from the neutral mouths to the smiling ones while speaking, or vice versa.

For case 3, mouth images are selected from the expressive database by switching between neutral and smiling mouth images. Based on the switching matrix of viseme transitions, the viseme transition with lowest value around the input expression tag is determined as an initial switching position. In order to achieve a very smooth transition while speaking and changing expressions, units from both expressions with small concatenation costs have to be found. With low concatenation cost, the unit selection can generate smooth animations.

Fig. 6 shows a sequence of mouth images that the unit selection considers for the mouth animation. The time line is labeled with phonemes. According to the target cost, a segment is selected from the database. According to the concatenation cost, optimal transitions between segments are determined. Depending on the input expression tag, the unit selection selects appropriate mouth images from the database containing the corresponding expression. We have observed that viseme switching from neutral to smile is always occurring when the mouth is almost closed.

4. EXPERIMENTAL RESULTS

4.1. Objective Evaluation

Animations are driven by the phonemes and durations of the real audio, so that the comparison between real and synthesized sequences is possible. Objective evaluation is performed by directly comparing the geometric (mouth height) and appearance (the first significant PCA component) parameters of animated sequence with the real one. The first PCA component of the mouth image database represents the mouth height. However, the mouth height cannot replace the first PCA component, because different mouth textures could have the same mouth height.

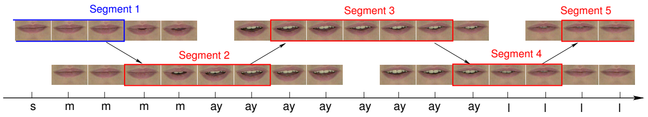


Fig. 6. Part of the unit selection for the word “smile”. Segment 1 is from the neutral mouths, the other segments are from smiling mouths. In this example, switching occurs at the phoneme ’m’, when the mouth is closed.

Experiments indicate that human viewers are very sensitive to closures, and get the closures at the right time may be the most important criterion for providing the impression that lips and sound are synchronized. Closures are easy to identify visually, simply by finding whether the inner lips are touched. The precise shapes of the openings are less important for the perceived quality of the articulation [2].

We select 9 recorded sentences with smile from the expressive database as ground truth. The talking head smiles at the end of each sentence before the last syllable. These 9 sequences are not included in the expressive database. For each real sequence, an animated sequence is generated by the expressive unit selection. The closures of the 9 real and animated sentences are measured. Experimental results show that the closures between animated and real sequences are always matched. The trajectories of geometric and appearance parameters for the animated and the real sequences of a sentence are shown in Fig. 7. These 9 real and animated sequences are further used for subjective tests.

The finding of the Advanced Television Systems Committee(ATSC) states that lip sync errors become noticeable if the audio is early by more than $15ms$ or late by more than $45ms$ [9]. We have measured the maximal frame offset of the closures between the synthesized and real sequences are 2 frames (audio delayed video $40ms$ at $50Hz$). Hence, the audio and video of animations are synchronized.

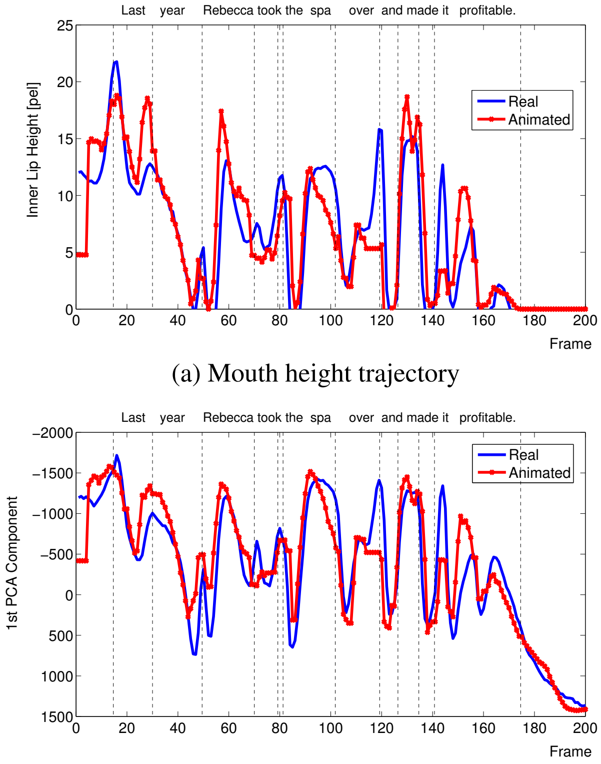
4.2. Subjective Evaluation

The standard approach [10] to assessing naturalness of a talking head is to conduct subjective tests where viewers score animations on a scale from 1 (bad) to 5 (excellent), which is also used in the first visual speech synthesis challenge [1].

30 students and staff from different departments are involved in the subjective tests. The viewers should give scores to the synthesized videos and the corresponding real ones.

Even though only the mouth is replaced by the new mouth, the quality of the whole face is influenced by colour and pose. Therefore, naturalness tests are to evaluate the overall quality of talking heads.

The 9 video pairs of random order are shown to each viewer only once. Each video pair has also a random order, so that the viewer does not know which one is played first. After showing a video pair, the viewer should give scores im-



(a) Mouth height trajectory

(b) PCA trajectory

Fig. 7. Trajectories of the animated sequence and the recorded sequence of a sentence. (a) shows the mouth height trajectories, (c) shows the 1^{st} PCA component trajectories. The sentence is "Last year Rebecca took the spa over and made it profitable.". The dashed lines are word boundaries.

mediately. The average mean opinion scores (MOS) and the confidence interval are plotted in Fig. 8. The overall MOS score of the 9 animated videos is 3.9 and the overall MOS score of the 9 recorded videos is 4.1.

In addition, we have asked the viewers, which part of face decreases the quality, such as eye, mouth, or unknown. The answers are insignificant. The viewers indicate that they cannot distinguish the smiling animations from the real ones.

Animations can be downloaded from http://www.tnt.uni-hannover.de/project/facialanimation/demo/emotion.

5. CONCLUSIONS

This paper presents an unit selection approach to generate a smiling talking head from input text and expression control tags. The unit selection selects appropriate units from the expressive database containing a large number of neutral and smiling mouth images. Viseme transitions are analyzed and optimal switching positions are found by a Viterbi search. Experimental results show that the expressive talking head is synchronized with speech, which is similar to the real one.

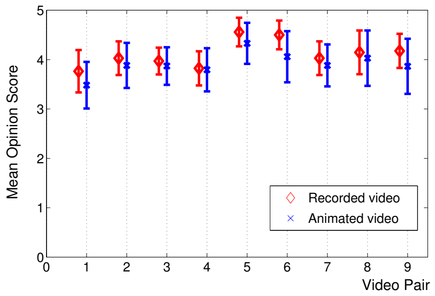


Fig. 8. Naturalness test of recorded and animated videos.

Subjective tests show that the synthesized smiles are indistinguishable from the real ones.

6. REFERENCES

[1] B. Theobald, S. Fagel, G. Bailly, and F. Elsei, "Lips2008: Visual speech synthesis challenge," in *Proceedings of Interspeech 2008*, 2008, pp. 2310–2313.

[2] E. Cosatto, J. Ostermann, H.P. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," in *Proceedings of the IEEE*, 2003, vol. 91, pp. 1406–1429.

[3] K. Liu and J. Ostermann, "Optimization of an image-based talking head system," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[4] Y. Cao, W.C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283–1302, 2005.

[5] Z. Deng and U. Neumann, *Data-driven 3D facial animation*, Springer-Verlag, 2008.

[6] P. Ekman, W.V. Friesen, and M. O'Sullivan, "Smiles when lying," *Journal of Personality and Social Psychology*, vol. 54, pp. 414–420, 1988.

[7] A. Weissenfeld, K. Liu, and J. Ostermann, "Video-realistic image-based eye animation via statistically driven state machines," *The Visual Computer, International Journal of Computer Graphics*, 2009.

[8] Z. Ambadar, J.F. Cohn, and L.I. Reed, "All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous," *J. Nonverbal Behav*, vol. 33, pp. 17–34, 2009.

[9] Advanced Television Systems Committee(ATSC), *ATSC Implementation Subcommittee Finding: Relative Timing of Sound and Vision for Broadcast Operations Advanced Television*, Doc. IS-191, 2003.

[10] ITU-R BT.500 11, "Methodology for the subjective assessment of the quality of television pictures," 2002.