Realistic Facial Animation System for Interactive Services

Kang Liu, Joern Ostermann

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

Abstract

This paper presents the optimization of parameters of talking head for web-based applications with a talking head, such as Newsreader and E-commerce, in which the realistic talking head initiates a conversation with users. Our talking head system includes two parts: analysis and synthesis. The audio-visual analysis part creates a face model of a recorded human subject, which is composed of a personalized 3D mask as well as a large database of mouth images and their related information. The synthesis part generates facial animation by concatenating appropriate mouth images from the database. A critical issue of the synthesis is the unit selection which selects these appropriate mouth images from the database such that they match the spoken words of the talking head. In order to achieve a realistic facial animation, the unit selection has to be optimized. Objective criteria are proposed in this paper and the Pareto optimization is used to train the unit selection. Subjective tests are carried out in our web-based evaluation system. Experimental results show that most people cannot distinguish our facial animations from real videos.

Index Terms: Talking Head, Unit Selection, Pareto Optimization, TTS (Text-To-Speech)

1. Introduction

The development of modern human-computer interfaces [1] such as web-based information services, E-commerce and E-learning will use facial animation techniques extensively in the future. Fig. 1 shows a typical application of a talking head for E-commerce. If the E-commerce web site is visited by a user, the talking head will start a conversation with the user. The user is warmly welcomed to experience the web site. The dialog system will answer any questions from the user and send the answer to the TTS (Text-To-Speech Synthesizer). The TTS produces the spoken audio track as well as the phonetic information and their duration which is required by the talking head plug-in embedded in the web site. The talking head plug-in selects appropriate mouth images from the database to generate a video. The talking head will be shown in the web site after the right download and installation of the plug-in and its associated database.

According to the underlying face model, talking faces can be categorized into 3D-model-based animation and image-based rendering of models [2]. The former usually does not enable a realistic talking head. Image-based approach [1] tries to achieve photo-realistic performance.

Here, we optimized our image-based facial animation system [3]. The training of the system to select the correct mouth images is time consuming and can only find one of the possible optimal parameters, such that the facial animation system can only achieve good quality for a limited set of sentences.

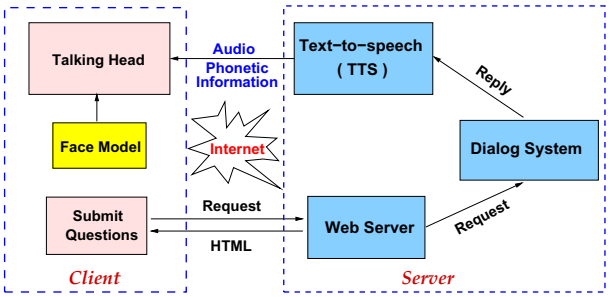


Figure 1: *Schematic diagram of web-based application with talking head for E-commerce.*

To better train the facial animation system, an evolutionary algorithm (Pareto optimization) [4] [5] is chosen. Pareto optimization is used to solve a multi-objective problem, which is to search the optimal parameters set in the parameter space efficiently and to track many optimized targets according to defined objective criteria. In this paper, objective criteria are proposed to train the facial animation system using Pareto optimization approach.

In the remainder of this paper, we describe the talking head system (Section 2), which will be optimized by Pareto optimization approach (Section 3). Experimental results and subjective evaluation are shown in Section 4. Conclusions are given in the last section.

2. Facial Animation System

2.1. Analysis

The audio-visual analysis of recorded human subjects results in a database of mouth images and their relevant features suitable for synthesis. The audio and video of a human subject reading text of a predefined corpus are recorded. The recorded audio and the spoken text are processed by speech recognition to recognize and temporally align the phonemes to the speech signal. Finally, the timed sequence of phonemes is aligned with the corresponding video. Therefore, for each frame of the recorded video, the corresponding phoneme and phoneme context are known. The phonetic context is required due to the co-articulation, since a particular mouth shape depends not only on its associated phoneme but also on its preceding and succeeding phonemes.

A 3D face mask is adapted to the first frame of the video using the calibrated camera parameters and some facial feature points. Motion estimation [6] is carried out to compute the rotation and translation parameters of the head movement in the later frames. These motion parameters are used to compen-

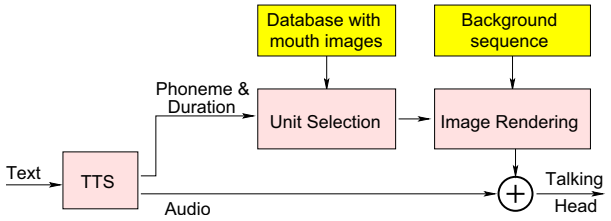


Figure 2: *The system architecture of the image-based talking head system.*

sate for head motion such that normalized mouth images can be saved in a database. Instead of PCA (principal component analysis) parameters, LLE (Locally Linear Embedding) parameters [7] are calculated to describe the texture of a mouth image. The geometric parameters, such as mouth corner points and lip position, are obtained by AAM based feature point detection [8]. All the parameters associated with an image are also saved in the database. Therefore, the database is built with a large number of normalized mouth images. Each image is characterized by geometric parameters, texture parameters (LLE parameters), phonetic context, etc.

2.2. Synthesis

The talking head system, also denoted as visual text to speech synthesizer (VTTS), is depicted in Fig. 2. First, a segment of text is sent to a TTS synthesizer. The TTS provides the audio track as well as the sequence of phonemes and their durations, which are sent to the unit selection. Depending on the phoneme information, the unit selection selects mouth images from the database and assembles them in an optimal way to produce the desired animation. The unit selection balances two competing goals: lip synchronization and smoothness of the transition between consecutive images. For each goal a cost function is defined, both of them are functions of the mouth image parameters. The cost function for lip synchronization considers the co-articulation effects by matching the distance between the phonetic context of the synthesized phoneme and the phonetic context of the mouth image in the database. The cost function for smoothness reduces the visual distance at the transition of images in the final animation, favoring transitions between consecutively recorded images. Then, an image rendering module stitches these mouth images to the background video sequence. Background videos are recorded video sequences of a human subject with typical head movements. Finally the facial animation is synchronized with the audio, and a talking head is displayed.

2.2.1. Unit Selection

The unit selection selects the mouth images corresponding to the phoneme sequence, using a target cost and a concatenation cost function to balance lip-synchronization and smoothness. As shown in Fig. 3, the phoneme sequence and audio data are generated by the TTS system. For each frame of the synthesized video a mouth image should be selected from the database for the final animation. The selection is executed as follows:

First, a search graph is built. Each frame is populated with a list of candidate mouth images that belong to the viseme corresponding to the phoneme of the frame. A viseme is a basic unit of speech in the visual domain, for example, the phonemes 'm', 'b', 'p' correspond to the closure viseme. Using a viseme

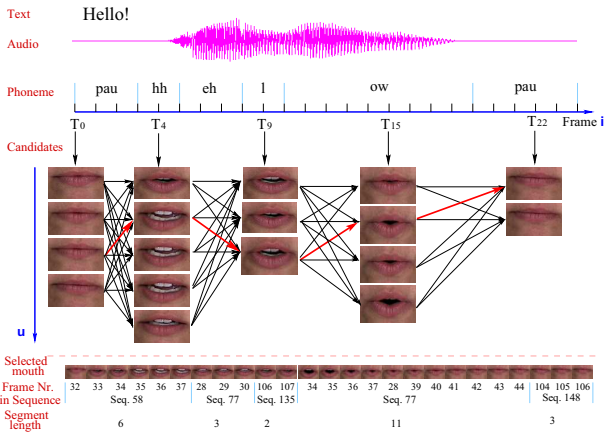


Figure 3: *Illustration of unit selection algorithm.*

instead of a phoneme increases the number of valid candidates for a given target, given the relatively small database. Each candidate is fully connected to the candidates of the next frame. The connectivity of the candidates builds a search graph as depicted in Fig. 3. Target costs are assigned to each candidate and concatenation costs are assigned to each connection. A Viterbi search through the graph finds the optimal path with minimal total costs.

The Target Cost (TC) is a distance measure between the phoneme at frame i and the phoneme of image u in the candidate list:

$$TC(i,u) = \frac{1}{\sum_{t=-n}^{n} v_{t+i}} \sum_{t=-n}^{n} v_{i+t} \cdot M(T_{i+t}, P_{u+t}) \quad (1)$$

where a target phoneme feature vector: $\overrightarrow{T_i} = (T_{i-n}, ..., T_i, ..., T_{i+n})$ with T_i representing the phoneme at frame i, a candidate phoneme feature vector: $\overrightarrow{P_u} = (P_{u-n}, ..., P_u, ..., P_{u+n})$ consisting of the phonemes before and after the u^{th} phoneme in the recorded sequence and a weight vector: $\overrightarrow{v_i} = (v_{i-n}, ..., v_i, ..., v_{i+n})$ with $v_i = e^{\beta_1 |i-t|}$, $i \epsilon [t-n, t+n]$, n is phoneme context influence length, depending on the speaking speed and the frame rate of the recorded video, β_1 is a negative constant. M is a phoneme distance matrix, which denotes visual similarities between phoneme pairs.

The Concatenation Cost (CC) is calculated using a visual cost (f) and a skip cost (g) as follows:

$$CC(u_1, u_2) = wccf \cdot f(U_1, U_2) + wccg \cdot g(u_1, u_2) \quad (2)$$

with the weights $wccf$ and $wccg$. Candidates, u_1 (from frame i) and u_2 (from frame $i-1$), have a feature vector U_1 and U_2 of the mouth image considering the articulator features including teeth, tongue, lips, appearance, and geometric features.

The Visual Cost f is defined as:

$$f(U_1, U_2) = \sum_{d=1}^{D} k_d \cdot ||U_1^d - U_2^d||_{L2} \quad (3)$$

$||U_1^d - U_2^d||_{L2}$ measures the Euclidean distance in the articulator feature space with D dimension. Each feature is given a weight k_d which is proportional to its discrimination.

The Skip Cost g is calculated as:

$$g(u_1,u_2)=\begin{cases} 0; |f(u_1)-f(u_2)|=1 \wedge s(u_1)=s(u_2) \\ w_1; |f(u_1)-f(u_2)|=0 \wedge s(u_1)=s(u_2) \\ w_2; |f(u_1)-f(u_2)|=2 \wedge s(u_1)=s(u_2) \\ \ldots; \\ w_p; |f(u_1)-f(u_2)|\geq p \vee s(u_1)\neq s(u_2) \end{cases} \tag{4}$$

with f and s describing the current frame number and the original sequence number that corresponds to a sentence in the corpus, respectively and $w_i = e^{\beta_2 i}$, β_2 and p are constant.

A path (p_1,p_2,...,p_i,...,p_N) through this graph generates the following Path Cost (PC):

$$PC = wtc\cdot\sum_{i=1}^{N} TC(i,S_{i,p_i}) + wcc\cdot\sum_{i=1}^{N} CC(S_{i,p_i},S_{i-1,p_{i-1}}) \tag{5}$$

with candidate S_{i,p_i} belonging to frame i. wtc and wcc are the weights of two costs.

Substituting Equ(2) in Equ(5) yields

$$PC = wtc\cdot C1 + wcc\cdot wccf\cdot C2 + wcc\cdot wccg\cdot C3 \tag{6}$$

with

$$\begin{aligned} C1 &= \sum_{i=1}^{N} TC(i,S_{i,p_i}) \\ C2 &= \sum_{i=1}^{N} (f(S_{i,p_i},S_{i-1,p_{i-1}})) \\ C3 &= \sum_{i=1}^{N} (g(S_{i,p_i},S_{i-1,p_{i-1}})) \end{aligned}$$

There are several weights that should be trained. But the global optimal weights cannot be found, only the local optimal weights are calculated in [9]. And this training is time consuming.

3. Unit Selection Training by Pareto Optimization

As discussed in section 2.2.1, several weights, influencing TC, CC and PC, should be trained. Generally, the training set includes several original recorded sentences (as ground truth) which are not included in the database. Using the database, an animation will be generated using the given weights for unit selection. We use objective evaluator functions as Face Image Distance Measure (FIDM). The evaluator functions are average target cost, average segment length, average visual difference between segments, and the similarity between the recorded sequence and the animated sequence, which is calculated by the cross-correlation of the mouth features. The average target cost indicates the lip-synchronization, the average segment length and average visual difference indicate the smoothness. The similarity measures the overall quality of the animations regarding to real sequences.

The average target cost and the average visual difference are computed as

$$TC_{avg.} = \frac{1}{N}\sum_{i=1}^{N} TC(i,S_{i,p_i}) \tag{7}$$

$$VC_{avg.} = \frac{1}{N}\sum_{i=1}^{N} (f(S_{i,p_i},S_{i-1,p_{i-1}})) \tag{8}$$

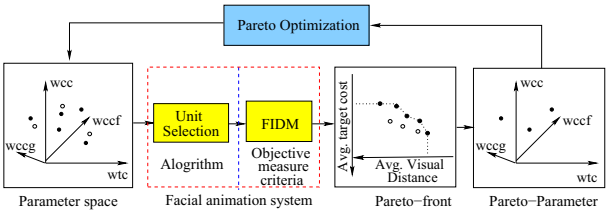


Figure 4: *The Pareto optimization for the unit selection.*

where (p_1,p_2,...,p_i,...,p_N) is the best path.

The average segment length is calculated as

$$SL_{avg.} = \frac{1}{L}\sum_{l=1}^{L} (SL_l) \tag{9}$$

where L is the number of segments in the final animation. For example, the average segment length of the animation in Fig. 3 is calculated as $SL_{avg.} = (6+3+2+11+3)/5 = 5$.

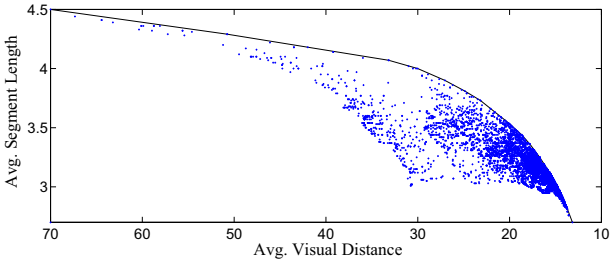
The similarity of the real sequence and the animated sequence is described by directly comparing the visual parameters of the animated sequence with the real parameters extracted from the original video. We use the cross-correlation of the two visual parameters as the measure of similarity. The visual parameters are the size of open mouth and the texture parameters.

FIDM is used to evaluate the unit selection and the Pareto optimization accelerates the training process. The Pareto optimization (as shown in Fig. 4) begins with thousand combinations of weights of the unit selection, where ten settings were chosen for each of the four weights in our experiments. For each combination, there is a value calculated using the FIDM criteria. The boundary of the optimal FIDM values is called Pareto-Front. The boundary indicates the animation with smallest possible target cost given a visual distance between segments. Using the Pareto parameters corresponding to the Pareto-Front, the Pareto optimization generates new combinations of the weights for further FIDM values. The optimization process is stopped as soon as the Pareto-Front is declared stable.

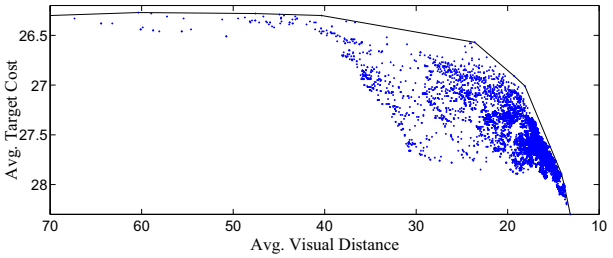
4. Experimental Results

The unit selection is trained by Pareto optimization with 30 sentences. The Pareto-Front is calculated and shown in Fig. 5. There are many weight combinations given results on the Pareto-Front, but only one combination of weights is determined as the best set of weights for unit selection. To evaluate the Pareto-Front, we conduct an informal subjective assessment in order to find the best animation with respect to the overall quality, naturalness, smoothness and synchronization. The weights corresponding to a selected point on the Pareto-Front are used for the unit selection. Animations generated by the optimal facial animation system are used for formal subjective test.

Assessing the quality of a talking head system becomes even more urgent as the animations become more lifelike, since improvements may be more subtle and subjective. A subjective test where observers give feedback is the ultimate measure of quality, although objective measurements used by the Pareto optimization can greatly accelerate the development and also increase the efficiency of subjective tests by focusing them on the important issues. Since we need a large number of observers, preferably from different demographic groups, we designed a web site for subjective tests.



(a) Evaluation space for $VC_{avg.}$ and $L_{avg.}$



(b) Evaluation space for $VC_{avg.}$ and $TC_{avg.}$

Figure 5: Pareto optimization for unit selection. The blue points are Pareto points. The curve is the Pareto-Front.

A Turing test [10] is performed to evaluate our talking head system on the web-based subjective test. We generate 8 facial animation sequences synchronized with accompanied real audio by the optimized unit selection. The real videos corresponding to the real audios are not part of the database. 8 recorded videos and its synthesized videos, totally 16 videos are collected to build a video database available for subjective test on our web site. 50 students and employees of Leibniz University of Hannover were invited to take part in the formal subjective test. All videos from the video database are presented to the participant randomly. The participant will decide whether it is a real or a synthesized video immediately after a video was presented.

The results of the subjective test are summarized in Table 1. The Turing test can be quantified in terms of the Correct Identifying Rate, which is defined as

$$CIR = \frac{Number \quad of \quad correctly \quad identified \quad utterances}{Number \quad of \quad testing \quad utterances} \tag{10}$$

Table 1 shows that on average 42% of the synthesized videos are detected as real videos, while 15% of the real videos are mistakenly considered as synthesized video. CIR 50% for the synthesized sequences is expected, which indicates a participant is never able to distinguish above chance between real and animated video. Comparing to the subjective test results in [10], the CIR for the synthesized sequences by our facial animations is closer to the chance level.

Based on the facial animation system, web-based interactive services such as E-shop and Newsreader are developed. The demos and related web site can be found at http://www.tnt.uni-hannover.de/project/facialanimation/demo.

5. Conclusions

We have presented an image-based talking head for web-based applications. To achieve a realistic talking head, the facial animation system has to be trained. The optimization criteria include lip synchronization, visual smoothness and others. The Pareto optimization is chosen to train the unit selection. Formal subjective tests show that synthesized animations generated by the optimized unit selection matches the corresponding audio naturally. More encouraging, 42% of the synthesized animations are so realistic that the viewers cannot distinguish them from real videos.

Table 1: *Results of the subjective test for a talking head (sample mean $\bar{x}$, standard deviation s and correct identifying rate CIR).*

Sequences	$\bar{x}$	s	total	CIR
synthesized	4.6	1.2	8	58%
real	6.8	0.8	8	85%

6. Acknowledgements

This work is funded by EC within FP6 under Grant 511568 with the acronym 3DTV. The authors would like to thank Holger Blume for his support with the Pareto optimization software.

7. References

[1] Cosatto, E., Ostermann J., Graf, H.P., and Schroeter, J. "Lifelike Talking Faces for Interactive Services", Proceedings of the IEEE, vol. 91, no. 9, pp. 1406-1429, September, 2003.

[2] Ostermann, J., and Weissenfeld, A. "Talking Faces - Technologies and Applications", Proceedings of ICPR04, vol. 3, pp. 826-833, August, 2004.

[3] Weissenfeld, A., Liu, K., Klomp, S., and Ostermann, J. "Personalized Unit Selection for an Image-based Facial Animation System", Proc. MMSP 05, Shanghai, China, October 2005.

[4] Zitzler, E., Laumanns, M., and Bleuler, S. "A Tutorial on Evolutionary Multiobjective Optimization", MOMH 2002, Springer Verlag, 2004.

[5] Von Livonius, J., Blume, H., and Noll, T.G. "Flexible Umgebung zur Pareto-Optimierung von Algorithmen - Anwendungen in der Videosignalverarbeitung", ITG 2007.

[6] Liu, K., Weissenfeld, A., and Ostermann, J. "Robust Rigid Head Motion Estimation based on Differential Evolution", Proc. ICME 06, pp.225-228, Toronto, Canada, July 2006.

[7] Weissenfeld, A., Urfalioglu, O., Liu, K., and Ostermann, J. "Parameterization of mouth images by LLE and PCA for image-based facial animation", Proc. ICASSP 06, Toulouse, France, May 2006.

[8] Liu, K., Weissenfeld, A., Ostermann, J., and Luo, X. "Robust AAM Building for Morphing in an Image-based Facial Animation System", Proc. ICME 08, Hanover, Germany, June 2008.

[9] Hunt, A., and Black, A. "Unit selection in a concatenative speech synthesis system using a large speech database", In Proceedings of ICASSP, pp.373-376, 1996.

[10] Xie, L., and Liu, Z.Q. "Realistic Mouth-Synching for Speech-Driven Talking Face Using Articulatory Modelling", IEEE Transactions on Multimedia, 9(3), pp.500-510, 2007.