ROBUST AAM BUILDING FOR MORPHING IN AN IMAGE-BASED FACIAL ANIMATION SYSTEM

Kang Liu, Axel Weissenfeld, Joern Ostermann *

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

Xinghan Luo†

Institute of Signal Processing,
Tampere University of Technology,
P.O. Box 553 FI-33101 Tampere, Finland

ABSTRACT

Facial animation has been combined with text-to-speech synthesis to create innovative multimodal interfaces, such as web stores and web-based customer services. This paper presents an image-based facial animation system using Active Appearance Models (AAM) for precisely detecting feature points in the human face, which are required for selecting mouth images from the database of the face model. In order to minimize the impact of human error when creating the training data, a new optimization method for building an AAM is produced. Optimized training set reduces the average feature point location error from 1.15 pixels to 0.17 pixel. The feature points are suitable for automatic morphing between mouth images with large visual differences. Subjective tests show that morphing improves the visual quality of the animation from "fair" to "good".

Index Terms— Human Machine Interface, Facial Animation, AAM, Morphing, Unit Selection

1. INTRODUCTION

Realistic face animation is still challenging, especially when we want to automate it to a large degree. Faces are the focus of attention for any audience, and the slightest deviation from normal faces is immediately noticed, especially for the mouth part. Talking faces play an important role in modern human computer interfaces, such as virtual agents, newsreaders, or a virtual assistant as part of an e-commerce, e-learning or e-care web site. Subjective tests show that the trust and attention of humans towards machines increases by 30% if humans are communicating with talking heads instead of text-only [1].

The image-based facial animation [1] [2] is divided into two parts, namely analysis and synthesis. The audio-visual analysis of recorded human subjects is depicted in Fig.1 The analysis results in a database of mouth images and their relevant features suitable for synthesis. The recorded audio and the spoken text are processed by speech recognition and labeled with phoneme and time information.

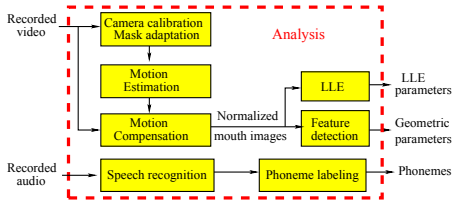


Fig. 1. An exemplary flowchart of the analysis of facial animation. The analysed parameters is to build the database.

The recorded videos are calibrated and a face mask is adapted to the first frame using the calibrated camera parameters and some facial feature points. Then motion estimation is carried out to compute the motion parameters of the head movement in the later frames. These motion parameters are used to compensate for the head motion such that normalized mouth images can be saved in the database. Texture parameters are calculated by LLE (locally linear embedding) [3] instead of PCA parameters in the reference system [1]. The geometric parameters, such as mouth corners and lip contour, are obtained by feature detection. 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, phonetic context, etc.

The synthesis architecture of a talking head is shown in Fig.2. First, a segment of text is inputted to a text-to-speech synthesizer (TTS). The TTS provides the audio track as well as the sequence of phonemes and their durations, which are sent to the unit selection engine. 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 com-

*This paper is funded by EC within FP6 under Grant 511568 with the acronym 3DTV.

†Part of the work was performed during Xinghan Luo's visit (10.2006-05.2007) under the 3DTV student exchange program.

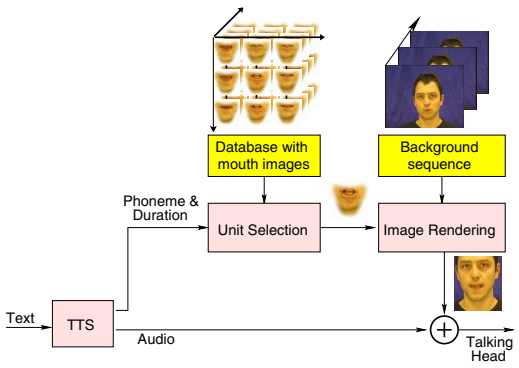


Fig. 2. The system architecture of a talking head.

peting goals: lip synchronization and smoothness of the transition between consequent images. For each goal a cost function is defined, both of them are functions of the mouth image parameters defined above. Then, an image-rendering module stitches these mouth images to the background video sequence. Background sequences are recorded videos of the human subject with typical short head movements. Finally the facial animation is synchronized with audio, and a talking head is displayed.

In the reference system [1], the feature detection is implemented by two steps. First, the face is located by color based segmentation. Second, feature points are detected using manually predefined feature templates, such as mouth corners with different open size. One of the deficiencies is that it is time consuming and very sensitive to the skin color and lighting of the subject. If the color contrast between lip and skin is small, the threshold cannot be found correctly. Second, the feature templates are selected manually. The feature points are not always consistent in all the templates. In this paper, the facial features are detected by AAM [4] [5], which uses not only the color information but also the shape of the subject such that the approach can locate the feature points very precisely and robustly. Moreover, a new approach to reduce the manual error when creating the training data is proposed in the paper.

The second improvement in our facial animation system is to smooth the transitions between segments by morphing instead of blending [1]. A segment is a chunk of consecutively recorded mouth images. Morphing [6] requires several corresponding feature points between two images. The color based method can only detect the features points such as mouth corners, which are not enough for Morphing. However, the results from AAM can satisfy this need.

The remainder of this paper is organized as follows. Section 2 summarized the basic AAM algorithm. Section 3 describes the optimized AAM with subpel accurate feature detection. Section 4 addresses the smoothness of animations is achieved by morphing of transitions between segments. Experimental results are reported in section 5. Finally, section 6 concludes the paper.

2. ACTIVE APPEARANCE MODELS (AAM)

The active appearance models [4] were introduced by Edwards et al. a few years ago, and have since been the subject of much research. An appearance model is a joint statistical model of the shape and the texture of an object, while the active appearance model is an appearance model combined with a directed search algorithm for adapting the model to an image. This section will summarize the traditional AAM algorithm.

To our mouth image database, several different mouth images are selected and landmarked. Each image is related to a landmark vector. The landmark vectors can be used to build the shape model for the object (mouth, in our context). The statistical shape is parameterized according to

$$x = \overline{x} + P_s b_s \tag{1}$$

$\overline{x}$ is the standard (average) shape of the model, the columns of P_s are the shape eigenvectors, and the b_s contains the shape parameters.

The shapes of all images are aligned by Generalized Procrustes Analysis (GPA). The textures under the landmarked shapes are extracted and normalized to the mean shape. The normalized texture data vector are parameterized by PCA, a linear model is obtained as

$$g = \overline{g} + P_g b_g \tag{2}$$

$\overline{g}$ is the mean texture, the columns of P_g are the texture eigenvectors, and the b_g contains the texture parameters.

In order to recover the correlation between the shape and the texture, the parameters b_s and b_g are combined in a third PCA space. A combined parameter c is computed as

$$b = \begin{pmatrix} W_s b_s \\ b_g \end{pmatrix} = \begin{pmatrix} W_s P_s^T (x - \overline{x}) \\ P_g^T (g - \overline{g}) \end{pmatrix} = \begin{pmatrix} P_{cs} \\ P_{cg} \end{pmatrix} c = P_c c \tag{3}$$

where W_s is a diagonal matrix of weights for each shape parameter, allowing for the difference in units between the shape and texture model. P_c is the combined eigenvectors.

The Appearance model is parameterized by c as

$$\begin{cases} x = \overline{x} + P_s W_s^{-1} P_{cs} c \\ g = \overline{g} + P_g P_{cg} c \end{cases} \tag{4}$$

A synthesized image can be generated by warping the texture of g into the shape of x for a given c and applying the current pose parameters $p = [t_x \quad t_y \quad s \quad \theta]$ where t_x, t_y and θ denotes in-plane translation and rotation, and s denotes the shape size.

The goal of AAM search is to find the optimal adaptation of the model to the input image, that is to find the parameter

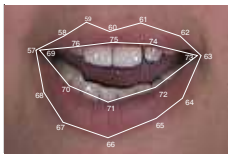
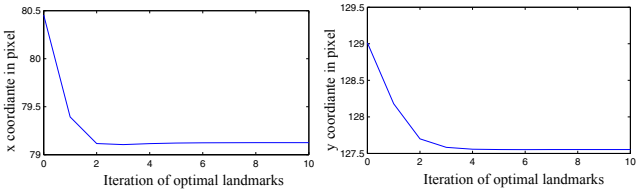


Fig. 3. 20 Landmarks are defined for mouth part.



(a) x coordinate (b) y coordinate

Fig. 4. Manual errors of landmarks are reduced iteratively by AAM optimization.

c and p that minimizes the distance between the model and the image using L_2-norm as a cost function. The iterative updating scheme is based on a fixed Jacobian estimate or a principle component regression [4].

3. FEATURE DETECTION BY AAM

Facial features are defined as landmarks on the mouth part (Fig.3) in our system. The landmarks are divided into three groups: anatomical-, mathematical- and pseudo-landmarks [5]. For example, the mouth corners are anatomical landmarks, the lowest points of inner and outer lip are mathematical landmarks, and the points on the outline or between landmarks are pseudo landmarks.

An appropriate training set is required for AAM building. According to different mouth appearance such as lip open size, teeth and tongue, about 40 mouth images are selected from the database with 20000 mouth images. The training set includes the mouth images and their landmarks.

3.1. Optimization of AAM Model

Because the manual work of placing landmarks is a subjective task, in general, the landmarks are not consistent in the training set. Improving the accuracy of the landmarks in the training set should be carried out, which is also called a repeatability and reproducibility study. Stegmann [5] has proposed that by letting a set of operators annotate the data set several times, the average landmarks are estimated as the optimal landmarks for the data set. However, this method needs a lot of manual work and it is tedious. In this part, we will propose a new approach, which can reduce the manual error of landmarks in the training set.

The optimization process is described as follows:

Step 1: The training set is used to build the AAM model.

Step 2: Features of each image from the training set are detected by the AAM model. If the features did not change from the last iteration, the optimization process will stop. Otherwise, do next step.

Step 3: The landmarks in the training set will be updated by the detected feature points. Go to step 1.

Generally the landmark will be converged to a stable position after 3 to 5 times iteration.

Fig.4 shows an optimization process of a mouth corner. The x coordinate of the mouth corner is shown in Fig.4 (a), the y in Fig.4 (b). The optimization starts from the manual landmark. In this case, 80.5 is the x coordinate of mouth corner. After 4 iterations, the x coordinate is converged to 79.13. The position of the landmark moves about 1.37 pixels in x, 1.46 pixels in y. The average correction of the landmarks is about 0.78 pixel in x and 0.89 in y.

3.2. Sub-pel accurate feature detection

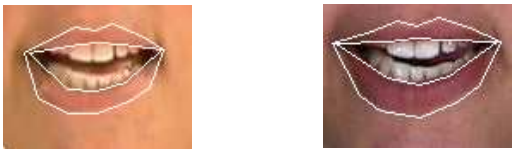
The optimized AAM is used to detect feature points of the mouth images, resulting in precise positions. We have also defined an objective criteria to measure the accuracy of the feature points.

The most popular method to measure the accuracy is to mark a test sequence manually, the average difference between the detected features and the manual marked features is calculated as the error. This method is time consuming and cumbersome. The manual marked features are assumed to be ground truth. But the manual mark may not be consistent in different images. Therefore, consistent feature points are needed as ground truth.

We start with an optimized AAM based on 40 training images with landmarks. We manually mark landmarks in the 100 test images. Next we build an extended optimized AAM using training and test images. This extended AAM detects the ground truth in the test images. The position error is measured by the difference between the feature points detected using the AAM based on the training images and the ground truth.

The average error of the manually marked landmarks of the test images is about 1.15 pixel. Error becomes 0.44 pixel using the non optimized AAM model built by the training images. The error is reduced to 0.17 pixel by the optimized AAM. The results of feature detection under different conditions such as lighting and color changes are shown in Fig.5.

The mouth corners are detected by optimized AAM as accurately as by template base method [2], while AAM based detection saves much manual work to train the model and it is more robust to the lighting and color changes. The lip contour can also be detected with high accuracy by optimized AAM, but the template based method cannot. The detected mouth features are further used to control the morphing for anima-



(a) low contrast (b) high contrast

Fig. 5. Results of feature detection by optimized AAM.

tion.

4. MORPHING FOR ANIMATION

As introduced in section 1, the selected mouth images are made of segments of consecutively recorded image samples. The Viterbi search has done its best to minimize visual differences at the junction of these segments; however, because the database has a limited size, many junctions remain for which a large visual difference exists. These abrupt transitions are generally noticeable on the final animation and greatly lower its overall quality. Instead of blending, field morphing is implemented when the junctions have a large visual difference, although the appropriate mouth images are selected by the unit selection algorithm from a limited database. Morphing requires the feature points of the mouth part. The correspondences of the feature points in the mouth images have been detected by optimized AAM approach. Therefore, morphing can be implemented in our system, and the whole system runs also real time, because the mouth part is small and morphing is calculated only in the junctions with large visual difference.

5. EXPERIMENTS

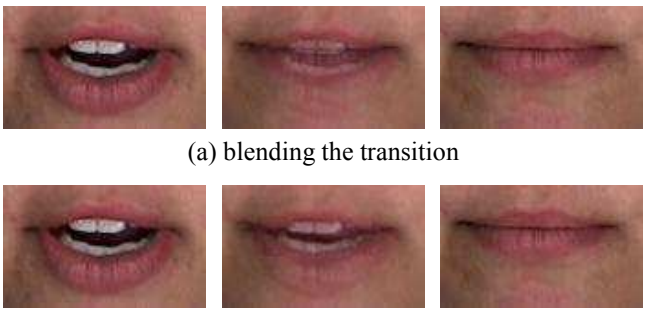
We use the optimized AAM and morphing techniques to generate different animations, comparing the reference approach. 20 students from Leibniz University of Hannover were asked to score the animations in terms of naturalness and smoothness, namely scoring test. The results of the scoring test showed that the average quality increasing from 3 to 4.5, where a five-point assessment is used that 5 means "excellent" and 1 means "bad".

An animation sequence (25 fps) is produced for the utterance "We are working on facial animation. " Only three of 17 transitions need morphing. On average 25% of transitions are morphed in animations. The mouth part of animations with blending and morphing the transitions are shown in Fig.6.

Some facial animations can be found at http://www.tnt.uni-hannover.de/project/facialanimation/demo.

6. CONCLUSIONS

This paper presents an improved training method for an Active Appearance Model (AAM), which can reduce the average



(a) blending the transition

(b) morphing the transiton

Fig. 6. Snapshots of mouth part of facial animations. The middle image is blended in (a), morphed in (b). The visual quality of mouth part is improved by morphing. The blending factor is set to 0.5.

position error of landmarks from 1.15 pixels to 0.17 pixel iteratively. In contrast with the reference method, the feature detection using our improved AAM model performs more accurately and robustly. The detected geometric features are used for the synthesis of facial animation. Furthermore, the feature points are also used for morphing, which smoothes the transition between segments with large visual difference automatically. Since the two improvements are built in our system, the overall quality of animations is better than these by the reference method as evaluated by subjective tests.

7. REFERENCES

[1] E. Cosatto, J. Ostermann, H. Graf, and J. Schroeter, "Lifelike talking faces for interactive services," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1406-1429, Sept. 2003.

[2] A. Weissenfeld, K. Liu, S. Klomp, and J. Ostermann, "Personalized unit selection for an image-based facial animation system," *IEEE-MMSP 05*, Shanghai, Oct. 2005.

[3] K. Liu and A. Weissenfeld and J. Ostermann, "Parameterization of mouth images by LLE and PCA for image-based facial animation", Proc. ICASSP 06, pp. 461-464, Toulouse, France, May 2006.

[4] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 23(6):681-685, 2001.

[5] M.B. Stegmann, B.K. Ersboll, and R. Larsen, "FAME - a flexible appearance modelling environment," *IEEE Trans. on Medical Imaging*, 22(10):1319-1331, 2003.

[6] T. Beier and S. Neely, "Feature-Based Image Metamorphosis", Computer Graphics, Vol.26,No.2,pp.35-42, 1992.