PARAMETERIZATION OF MOUTH IMAGES BY LLE AND PCA FOR IMAGE-BASED FACIAL ANIMATION

Kang Liu, Axel Weissenfeld and Joern Ostermann, Fellow IEEE

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

ABSTRACT

This paper describes parameterization of mouth images for an image-based facial animation system. The analysis part of the facial animation system produces a face model, which is composed of a personalized mask as well as a large database of mouth images and their related phonetic and visual information. Then, a photo-realistic talking head is synthesized by rendering a personalized mask textured with a mouth image, which is selected from the database. The selection is driven by a unit selection algorithm, which finds the appropriate mouth images from the database such that they match the words spoken by the talking head. The selection of mouth images is based on parameters describing the mouth images. Therefore, the parameterization of mouth images is the key part for creating a photo-realistic facial animation. Hereby the visual parameterization of mouth images by LLE (Locally Linear Embedding) is investigated comprehensively and compared with PCA (Principal Component Analysis). Experimental results show that the parameterization of mouth images by LLE performs better for an image-based facial animation system than by PCA.

1. INTRODUCTION

Image-based rendering techniques have recently been introduced to the field of facial animation. Unlike model-based facial animation, image-based animation can achieve photo-realistic animations. Facial animation [1] can be applied in many applications, such as an information kiosk, customer service, e-commerce, e-learning etc. Subjective tests [2] show that a talking head embedded in a human computer interface can increase the trust of humans to computer.

Image-based facial animation systems [1] consist of two parts. One is the audiovisual analysis of recorded human subjects, the other is the synthesis of facial animations. In the analysis part the face model is created. The input of the visual analysis process is a recorded video sequence and a 3D face mask adapted to the shape of the recorded human subject. The face model consists of a personalized mask and a large database. Mouth images are normalized and stored in the database. Each image is labeled with its phonetic information, which is retained from the recorded audio by a speech-labeling software. The synthesis of facial animation involves an unit selection module, which selects appropriate mouth images from the database such that they match the words spoken by the talking head. The unit selection depends on geometric features (mouth width and height) and pixel features (the whole image information such as teeth, tongue, etc). Among them the impact of the pixel feature parameters of mouth images on the realism of the talking head is dominant. In [3] we have mainly discussed the weights of phonetic and visual costs for unit selection of facial animation system. Dimensionality reduction of mouth images is useful for faster and computationally more efficient animation. In [1], pixel features of the mouth images are parameterized by PCA. PCA is very suitable for data, which contain a linear or near linear structure in a high dimensional space. However, mouth images form a high dimensional data space that is neither linear nor a convex set.

To better parameterize the mouth image set, a non-linear algorithm, called LLE, is chosen. In [5] LLE has been introduced to describe the mouth data in a low dimensional space. However, their results were not tested for audiovisual speech synthesis. In this paper we thoroughly compare the visual parameterization method of mouth images in the database using PCA and LLE.

The remainder of this paper is organised in four parts. Section 2 introduces the synthesis of talking head. Section 3 explains parameterization of mouth images using PCA and LLE. Section 4 presents how to evaluate the two parameterization methods and some experimental results are given. The last section concludes the paper.

2. SYNTHESIS OF TALKING HEAD

The system architecture (Fig. 1) shows the block diagram of the synthesis of a talking head, which is also defined as a vi-

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

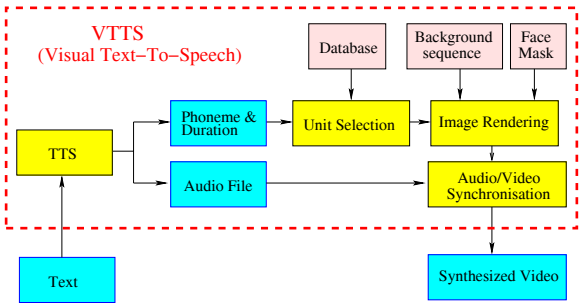

Fig. 1. Architecture of the synthesis part.

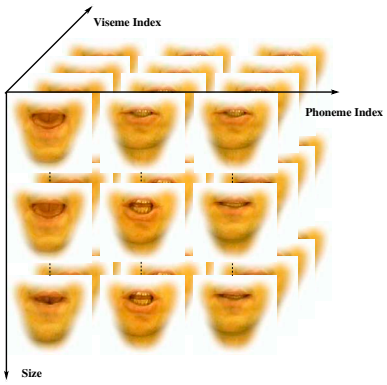

Fig. 2. Database of mouth images.

sual text to speech synthesizer (VTTS). First, a segment of text is sent to TTS (Text-To-Speech). The TTS provides the audio track as well as the phonetic information and their duration, which are sent to the unit selection engine. The unit selection chooses appropriate mouth images from the database. Then, an image rendering module stitches these mouth images to the background video sequence. Background videos are recorded video sequences of the human subject with typical short head movements. Finally the facial animation is synchronised with audio, and a talking head is displayed.

The database contains a large number of mouth images as shown in Fig. 2. Mouth images are labeled with viseme and phoneme indices. There are many mouth images associated with each phoneme. The size axis indicates the mouth size in the mouth image. For each image, we store related information like geometric features, pixel features, phonetic context, etc. Among them pixel features are very important for the unit selection algorithm. Unit selection [3] balances the two costs: phonetic and visual costs so as to find the best sequence of mouth images. Both of the two costs are functions of the pixel features of mouth images, which are used to measure the visual distance in the Euclidean space. Therefore, the quality of facial animations is related to pixel feature parameters of mouth images directly.

3. PARAMETERIZATION OF MOUTH IMAGES BY PCA AND LLE

The details of PCA and LLE algorithm are presented in [4] and [5]. Here we summarize the algorithms and comment on the advantages and disadvantages of both methods.

PCA and LLE are dimensionality reduction methods. They treat each mouth image as a data point in a high dimensional space. The high dimensional space of mouth images is transformed into a low dimensional space by PCA or LLE.

PCA algorithm contains an analysis and a reconstruction part.

PCA analysis:

Step 1: Calculate mean image $\overrightarrow{m}$ of data $\overrightarrow{X_i}$, $\overrightarrow{m} = \sum_{i=1}^{N} \overrightarrow{X_i}$, subtract it from each image $\overrightarrow{z_i} = \overrightarrow{X_i} - \overrightarrow{m}$ and get matrix $\mathbf{A} = [\overrightarrow{z}_1, \overrightarrow{z}_2, ..., \overrightarrow{z}_N]$.

Step 2: Calculate the covariance matrix $cov(\mathbf{A}) = \frac{1}{n}\mathbf{A}\mathbf{A}^T$ and its eigenvalues λ_n and eigenvectors ν_n.

Step 3: Select d eigenvectors ν_n with the first d largest eigenvalues λ_n as principal components.

Step 4: Compute coordinates $\overrightarrow{C_i}$ by projecting $\overrightarrow{X_i}$ on these d principal components.

PCA reconstruction:

$$\overrightarrow{X}_i^{'} = [\nu_1, \nu_2, ..., \nu_d] \cdot \overrightarrow{C_i} + \overrightarrow{m} \tag{1}$$

MSE (Mean Square Error) of the original data and its reconstructed data is calculated as :

$$E\left[\frac{1}{N}\sum_{i=1}^{N}|\overrightarrow{X}_i - \overrightarrow{X}_i^{'}|^2\right] = \frac{1}{N}\sum_{n=d+1}^{N}\lambda_n \tag{2}$$

Eq. (2) calculates the relationship between the dimensionality and distortion. If the mouth image data is transformed to the PCA space with the same dimensionality as the original high dimensional space, the MSE is zero. PCA can be used to estimate the intrinsic dimensionality of high dimensional data. The structure of the mouth image space is not changed, because PCA only rotates and translates the space.

LLE algorithm also contains an analysis and a reconstruction part.

LLE analysis:

Step 1: Find K nearest neighbors of each data point $\overrightarrow{X_i}$.

Step 2: Solve for weights W_{ij} that best construct each data point $\overrightarrow{X_i}$ from its neighbors, minimizing $\varepsilon(W)$.

$$\min \quad \varepsilon(W) \stackrel{def}{=} \sum_{i=1}^{N}\left|\overrightarrow{X_i} - \sum_{j=1}^{K} W_{ij}\overrightarrow{X_{ij}}\right|^2 \tag{3}$$

with constrained weights $\sum_{j=1}^{K} W_{ij} = 1, \forall i$.

Step 3: Compute embedding vectors $\overrightarrow{Y_i}$ using the weights W_{ij}, minimizing $\Phi(Y)$.

$$\min \quad \Phi(Y) \stackrel{def}{=} \sum_{i=1}^{N}\left|\overrightarrow{Y_i} - \sum_{j=1}^{K} W_{ij}\overrightarrow{Y_{ij}}\right|^2 \tag{4}$$

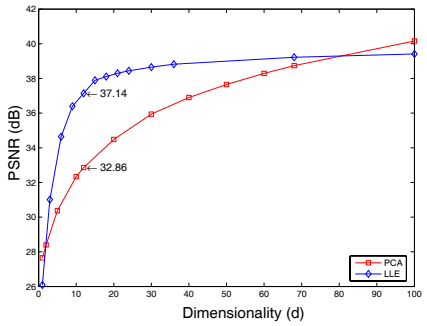

Fig. 3. PSNR performance of LLE and PCA.

subject to embedding vectors $\overrightarrow{Y}_i$ having unit covariance and $\frac{1}{N}\sum_i \overrightarrow{Y}_i \otimes \overrightarrow{Y}_i = I$, where I is the $d \times d$ identity matrix.
LLE reconstruction:
Step 1: Find K nearest neighbors of each data point $\overrightarrow{Y_i}$ in embedded space Y.
Step 2: Solve for weights W'_{ij} that best construct each data point $\overrightarrow{Y_i}$ from its neighbors.
Step 3: Reconstruct points in space X using the neighbors of $\overrightarrow{X}_i$ and the weights W'_{ij} of $\overrightarrow{Y}_i$.

$$\overrightarrow{X}'_i = \sum_{j=1}^{K} W'_{ij} \cdot \overrightarrow{X_{ij}} \tag{5}$$

LLE is nonlinear dimensionality reduction method. It transforms the mouth image data in three steps. First LLE finds the K nearest neighbours of every mouth image. Second, it calculates the optimal weights overall for every mouth image. Finally, using the derived weights the new coordinates of the mouth image are reconstructed in a low dimensional space. The structure of mouth image data is changed, but their topological relation retains the same before and after LLE transformation. Especially, inverse LLE transform must use the original mouth images, but the weights from Y subspace. LLE, which is not able to estimate the intrinsic dimensionality of mouth image data like Eq. (2) by PCA, calculates distortion and dimensionality only by the original data and its reconstructed data. K, the number of neighbours, is a free parameter that cannot deduce the relationship between distortion and dimensionality. LLE does not improve the reconstructed mouth image quality so much with increasing K. Typical values for K are integers from 6 to 24. Furthermore, LLE cannot reconstruct the high dimensional data losslessly like PCA, because the minimisation of Eq. (3)(4) is global and some particular data points do not involve local minima.

For facial animation, dimensionality reduction of mouth images and the best parameters representing the mouth images are two factors to be traded off. The first is for fast calculation, the latter is for accurate selection. Fig. 3 plots reconstruction quality of mouth images by PCA and LLE over

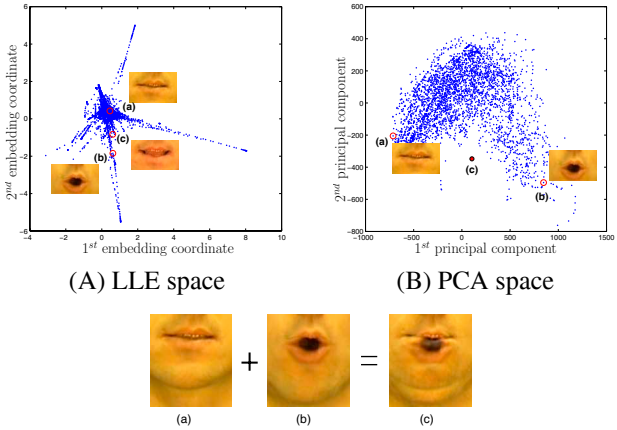

Fig. 4. Database is described in LLE (A) and PCA (B) space respectively. Representative mouth (a) and (b) are shown next to circled points. Image (C) shows averaging image in the PCA space.

increasing dimensionality. Compared to PCA, LLE requires a lower dimensionality to represent an image. Therefore, we can achieve high reconstruction quality in a low dimensional subspace by LLE, whereas PCA performs better, when the dimensionality of the subspace is over 80 for an images size of 8960 $pels$.

Our database contains about 20000 mouth images. The dimensionality of mouth images is reduced from 8960 to 12. Fig. 4 A and B shows the distribution of mouth images only by the first two coordinates in PCA and LLE (with $K = 6$) space. Note that while the linear projection by PCA has a somewhat uniform distribution about its mean, LLE has a distinctly spiny structure, with the mouth images of each spiny corresponding to mouth movements of sentences in the recorded video sequence. If the mouth images would describe a nearly linear manifold, these two methods would yield similar results; thus, the significant differences in these embeddings reveal the presence of nonlinear structures. Image C of Fig. 4 shows that mouth image c is reconstructed by PCA averaging the coefficients of the open mouth image a and the closed mouth image b. This mouth image shows two pairs of lips and chins, whereas the mouth image c derived from LLE space is a mouth image as shown in image A of Fig. 4. Hence, the mouth image set is neither a linear nor a convex data set.

4. EVALUATION AND EXPERIMENTAL RESULTS

This section discusses the impact of visual parameters by PCA or LLE on the facial animation system. First, we train the unit selection algorithm, which guarantees that the unit selection chooses the best mouth images using PCA or LLE so that the mouth movements correspond to the speech. The train-

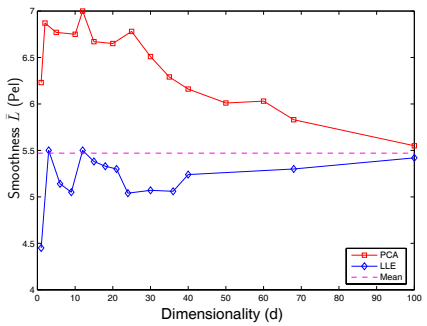

Fig. 5. Comparison of smoothness of facial animations generated by different dimensional pixel (PCA and LLE). The dashed indicates the mean smoothness $\bar{L}$ of real sequences.

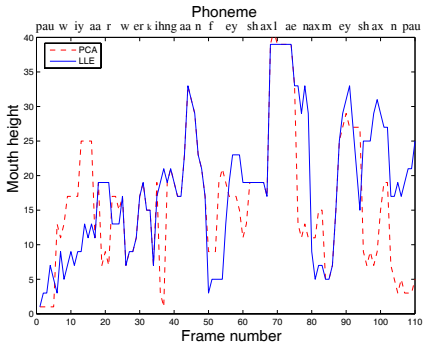

Fig. 6. Mouth movement in the synthesized sequence using PCA and LLE respectively.

ing process is presented in [3]. Then we choose PCA or LLE as different visual parameters to synthesize facial animation. More realistic facial animation can be achieved, if the selected mouth images change from one to the next smoothly. The smoothness criteria can be formulated as:

$$\bar{L} = \frac{1}{N-1} \sum_{i=1}^{N-1} \sqrt{(w_i - w_{i+1})^2 + (h_i - h_{i+1})^2} \qquad (6)$$

N is the number of synthesized sequence, w and h are width and height of mouth in pel, respectively.

Mean smoothness of the recorded video sequences is 5.47 pel and its variance is 1.20 pel^2. We assume that if the $\bar{L}$ is close to the mean smoothness value, the synthesized sequence is more photo-realistic. Fig. 5 shows smoothness criteria $\bar{L}$ for different dimensionality d for PCA and LLE. It is very clear that the smoothness criteria $\bar{L}$ by means of LLE are close to the mean value of the recorded videos, while the $\bar{L}$ by PCA are far from the mean value. But over 80 dimensionality the smoothness criteria $\bar{L}$ of both methods approach the mean value. Therefore, a more photo realistic facial animation is achieved by using LLE parameters in very low dimensional subspace according to the proposed objective criteria. Using only this low dimensional subspace reduces the computational load of the unit selection algorithm significantly. Fig.6 illustrates only one facial animation in which mouth images are labeled with 12 dimensional pixel feature parameters. It shows the mouth opening trajectory in the synthesized sequence using PCA and LLE separately. On the horizontal axis, the top presents the labels of the phoneme of the sentence:" We are working on facial animation. ", whilst the numbers at the bottom indicate the frame number. The curve produced using LLE is smoother. The optimal value $\bar{L}$ by LLE reaches 5.51 smaller than 6.99 by PCA.

Subjective test was carried out with two facial animation sequences using LLE and PCA, respectively. According to the results, the facial animation using LLE is distinctly smoother than the animation using PCA.

5. CONCLUSIONS

In this paper we have suggested that LLE is used in an image-based facial animation system instead of PCA for describing mouth images. PCA can describe the linear structure of the high dimensional data. Since the normalized mouth images form essentially nonlinear structures in a high dimensional space, PCA is not suitable for this special case. The LLE is a nonlinear dimensionality reduction method. It can reveal the nonlinear structure of the mouth images in a high dimensional space by exploiting the local linear reconstructions.

The objective criteria "smoothness of mouth motion" is proposed for evaluating the facial animation system. Experimental results show that the facial animation is more photo-realistic and computationally efficient by LLE parameters in low dimensionality.

6. 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, September, 2003.

[2] I. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, Issue 7/8, 1999.

[3] A. Weissenfeld, K. Liu, S. Klomp, and J. Ostermann, "Personalized unit selection for an image-based facial animation system," *7th International Workshop on Multimedia Signal Processing*, Shanghai, China, 2005.

[4] I. Jolliffe, "Principal Component Analysis," *Springer-Verlag*, New York, 1989.

[5] L. Saul and S. Roweis, "An introduction to local linearly embedding," http://www.cs.toronto.edu/ roweis/lle/publications.html.