Minimized Database of Unit Selection in Visual Speech Synthesis without Loss of Naturalness

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

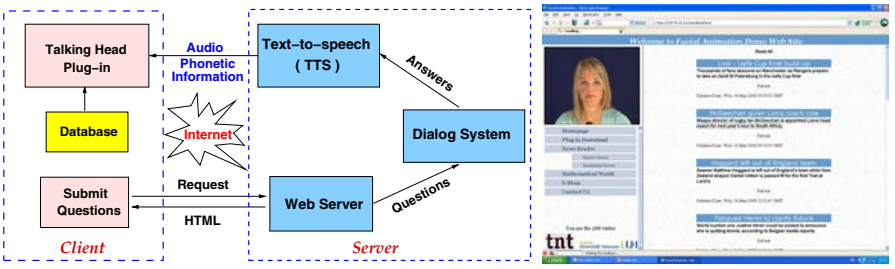
Abstract. Image-based modeling is very successful in the creation of realistic facial animations. Applications with dialog systems, such as e-Learning and customer information service, can integrate facial animations with synthesized speech in websites to improve human-machine communication. However, downloading a database with 11,594 mouth images (about 120MB in JPEG format) used by talking head needs about 15 minutes at 150 kBps. This paper presents a prototype framework of two-step database minimization. First, the key mouth images are identified by clustering algorithms and similar mouth images are discarded. Second, the clustered key mouth images are further compressed by JPEG. MST (Minimum Spanning Tree), RSST (Recursive Shortest Spanning Tree) and LBG-based clustering algorithms are developed and evaluated. Our experiments demonstrate that the number of mouth images is lowered by the LBG-based clustering algorithm and further compressed to 8MB by JPEG, which generates facial animations in CIF format without loss of naturalness and fulfill the need of talking head for Internet applications.

1 Introduction

Visual speech synthesis (talking head) is studied by researchers in computer graphics, image processing, speech processing, artificial intelligence, communication, psychology, etc. Different competing talking head systems have been presented in the first visual speech synthesis challenge LIPS 2008 [1]. The image-based talking head system [2] achieved the most natural animations in terms of audio-visual consistency [3]. An image-based talking head may be combined with dialog systems, such as desktop agents on personal computers, e-Learning and human-car-entertainment services [4], which will open many opportunities in modern human-machine communications.

A typical architecture of a talking head for an Web-based customer information service is shown in Fig. 1(a). A Web server will forward any questions from client to a dialog system, which sends the answer to a TTS (Text-To-Speech) synthesizer. The TTS converts the text of the answer to the corresponding spoken audio track as well as the phonetic information and their duration which is required by the talking head plug-in embedded in the website. The talking

X. Jiang and N. Petkov (Eds.): CAIP 2009, LNCS 5702, pp. 1212–1219, 2009.



(a) Architecture for customer information service. (b) Snapshot of a website

Fig. 1. Web-based applications with talking heads

head plug-in selects appropriate mouth images from the database to animate the talking head at the client. A snapshot of the Website is shown in Fig. 1(b).

For a talking head in PAL format, the database with about 12000 mouth images in JPEG format requires 120MB storage space and needs a long time to be downloaded from Internet. Therefore, the database should be minimized to realize the talking head for Internet applications. This paper proposes a prototype framework which can efficiently minimize the database and focuses on MST, RSST and LBG-based clustering algorithms.

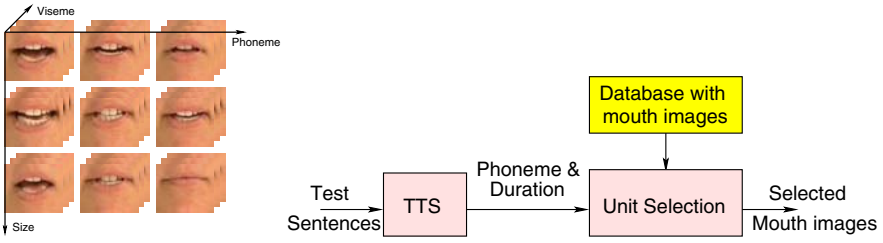
The rest of the paper is structured as follows. Section 2 describes the creation of the database. Section 3 presents the proposed framework of database minimization. Experimental results are shown in Section 4, and concluding remarks are drawn in the final section.

2 Database Creation

Our image-based talking head system includes two parts: analysis and synthesis. The audio-visual analysis part creates a database, which is available for the synthesis part to generate animations. The details of the analysis and synthesis were presented in [2] and this section describes the database creation briefly.

A subject is recorded while reading a predefined corpus including about 300 sentences. The motion parameters of recorded subject are estimated by a gradient-based approach [5], which is used to compensate head motion such that mouth images can be cropped from the normalized face sequences. A snapshot of the database with a large number of mouth images is shown in Fig. 2(a).

Since the dimension of normalized mouth image is very high and computation using the original mouth image is inefficient, the dimensionality should be reduced before clustering and compression. PCA (Principal Component Analysis) [6] is very efficient to reduce the dimension of the mouth images. In the PCA space, each base is an eigen mouth and each mouth image is the sum of weighted eigen mouths. The weight is the coordinate of the mouth image in the PCA space. Therefore, the problem of mouth image classification is simplified to cluster the data set in the low dimensional PCA space of the database.



(a) Database with mouth images (b) Determining selection frequency of mouth images

Fig. 2. Frequency determination of using mouth images

3 Database Minimization

This section proposes a framework to minimize the database. First, probability of using a mouth image is determined and only used mouth images are retained in the database. Second, the key mouth images are identified in the PCA space of the retained mouth images by clustering algorithms. For each cluster, one image is selected as a representative image from the cluster. All the representative images build a final database. Last, the final database is further compressed by JPEG. In order to cluster the database, three clustering algorithms are developed and an objective performance measurement of the clustering algorithms is defined.

3.1 Probability Determination of Using Mouth Images

In order to evaluate the relevance of each image in the database, the test corpus including 1457 sentences from different sources is comprised of:

- 400 titles of top news in different categories from BBC website;
- 100 sentences from the story "The Tale of Two Cities";
- 657 sentences from the corpus used for speech synthesis to cover all diphones in English;
- 300 sentences from AT&T research lab.

Using this test corpus, the relative frequency p_i of using the mouth image x_i is determined. By doing so, only the mouth images with $p_i > 0$ are retained and the mouth images with $p_i = 0$ are discarded.

Fig. 2(b) shows the process of the frequency determination, which is part of the synthesis system. Depending on the phonetic information from TTS, 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. 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

goal of smoothness is to reduce the visual distance at the transition of images in the final animation, favoring transitions between consecutively recorded images. The probability of using mouth images is derived from the selection frequency of mouth images.

3.2 Clustering Algorithms

The goal of the clustering algorithms is to identify key mouth images and discard similar mouth images from the database. For example, a lot of closed mouths are similar and only one is necessary for animations. Similarity depends on the distance d between two points in the PCA space. d is calculated as the weighted Euclidean distance between the two mouth images. Therefore, a threshold T is required to measure the similarity. $d(u, v) < T$ means that mouth image u and v are similar and should be classified in one cluster.

MST-Based Clustering

A spanning tree is an acyclic subgraph of a graph G, which contains all the vertices from G. The MST of a weighted graph is the minimum weight spanning tree of that graph. Prim's algorithm [7] is known to be a good algorithm to find an MST. The algorithm continuously increases the size of a tree starting with a single vertex until it spans all the vertices. The cost of constructing an MST is $O(mlogn)$, where m is the number of edges in the graph, n is the number of vertices. The classical MST based clustering algorithms begin with any point in the PCA space to construct an MST. From this tree, any edge with a weight $d(u, v) \geq T$ is removed from the tree. This leads to a set of disjoint subtrees $S_C = \{C_1, C_2, ...\}$. Each of the subtrees C_t is treated as a cluster, for which a representative point should be found. All representative points are collected to build a final database.

Each image in the database is a vertex of the graph. Prim's algorithm takes a long time and it becomes impossible to construct an MST for our database, since the database consists of a large number of mouth images, even though some non-used mouth images with $p_i = 0$ are discarded. In order to speed up MST construction for our database, we modify the algorithm by combining the two steps as follows. Once the weight of a new edge of the subtree is bigger than the threshold T, we stop constructing the subtree and the subtree is treated as a cluster. The remaining vertices of the database are treated the same to build clusters until all the vertices are part of the subtrees. The modified MST based clustering algorithm can build the same subtrees according to the threshold T as the classical MST based clustering does, but the computing time for our database is reduced to maximal 3 hours from several months or more.

To find the representative image, we assume that there are n_t points in the cluster C_t, the average node c_t and its standard deviation σ_t are computed considering the unit probability in the following way:

$$c_t = \sum_{j=1}^{n_t} p_j V_j \quad ; \qquad \sigma_t = \sqrt{\sum_{j=1}^{n_t} p_j (V_j - c_t)^2} \quad ,$$

where V_j is a vector with the PCA weights of mouth image j in C_t and p_j is the probability of using the mouth image j. If the condition $\sigma_t < T/6$ is fulfilled, the nearest point to c_t is selected as the representative point in the cluster C_t. Otherwise, cluster C_t will be approximated by two Gaussian mixture distributions or more, till the condition is fulfilled.

RSST-Based Clustering

RSST-based clustering algorithm presents a powerful solution to the problem of incorporating global information into clustering algorithm [9]. RSST begins with the shortest link in the graph and merges the two vertices joined by this link. A new vertex and link weights are recalculated in the region. The region represents a vertex or many vertices in the same partition. The process will be repeated until the number of regions are enough for the clustering.

In the case of mouth image clustering, we define the region as a mouth image in the PCA space at the initialization stage or many mouth images clustered in a partition. The link weights are calculated by the weighted Euclidean distance of two mouth images. The RSST-based clustering algorithm begins with finding the least link in the graph and merging the two mouth images adjoined by this link into a region. The average PCA weight of the region is calculated to represent the new vertex of the region and the link weights of the region are updated. This process will be done recursively until the desired regions are obtained. We define a threshold T that controls the building of RSST. If the next least link weight is bigger than T, we stop the construction of RSST. The mouth image, which is the nearest to the average PCA weight, is treated as the representative mouth image of the region.

LBG-Based Clustering

We assume that there is a training data set consisting of M vectors: $\tau = \{x_1, x_2, ..., x_M\}$ and the vectors are K-dimensional: $x_m = (x_{m,1}, x_{m,2}, ..., x_{m,K})$, $m = 1, 2, ..., M$. In our case, the vector represents the PCA weight of a mouth image. The LBG VQ design algorithm [8] is an iterative algorithm to find the partition $S = \{s_1, s_2, ..., s_P\}$ and their representative vectors $r_p = (r_{p,1}, r_{p,2}, ..., r_{p,K})$, $p = 1, 2, ..., P$, which are subject to $Q(x_m) = r_p$, $x_m \in s_p$, such that the global average distortion is minimized in the following way, considering the probability p_m of using the mouth image x_m:

$$D_{ave} = \frac{\sum_{m=1}^{M} p_m \cdot \left\| x_m - Q(x_m) \right\|^2}{K \cdot \sum_{m=1}^{M} p_m}.$$

However, the classic LBG does not consider the maximum distortion of the partition, which results in jerky animations, when the mouth images are selected from the partitions with a large distortion. To overcome this problem, we define a threshold R that controls the size of the partitions. The partitions are further split, if the maximal distortion of the partitions is larger than the threshold. The clustering algorithm is repeated until all the partitions fulfill the threshold

condition. Considering the probability of using a mouth image, the distortion $D(s_p)$ between any point and the representative point r_p in the partition s_p fulfills the following condition: $D(s_p) = \left\|x_m - r_p\right\|^2 < R, \forall\ x_m \in s_p$.

Objective Evaluation of Clustering Algorithms
To evaluate the clustering algorithms, PSNR is chosen as the objective measurement. The PSNR between the original database and the final database is defined as:

$$PSNR = 10 \cdot log_{10} \frac{255^2}{MSE} \quad ; \quad MSE = \frac{\sum_{m=1}^{M} P_{x_m} \cdot \left\|I_{x_m} - I_{Q(x_m)}\right\|^2}{w \cdot h \cdot \sum_{m=1}^{M} P_{x_m}}$$

where MSE is the mean square error, I_{x_m} is the pixel value vector of mouth image x_m, w and h are the width and height of the mouth image, $Q(x_m)$ is the representative image of x_m, P_{x_m} is the probability of using the mouth image x_m.

3.3 Compression of Final Database

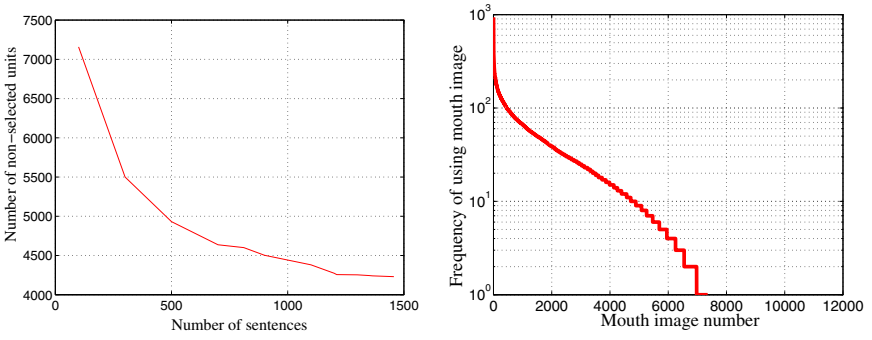
JPEG and H.264 are the most efficient coding methods for pictures and sequences. Due to the discontinuity of the mouth images in the final database, inter coding of images by H.264 is not useful. The efficiency of the intra coding of H.264 is similar to the JPEG efficiency. In practice, JPEG is the most popular and efficient coder for pictures and very suitable for our case. The size of compressed database in JPEG is proportional to the number of mouth images in the final database.

4 Experimental Results

In order to determine the probability of using mouth images, we have built an image-based talking head system [2] with 11594 mouth images in the database. The resolution of the talking head image is 720×576 and the cropped mouth image size is 176×208. This means the original database is about $640MB$ in YUV format and $120MB$ in JPEG format.

Fig. 3(a) shows the number of mouth images that are not selected by the unit selection given the number of test sentences. 4230 mouth images are never used by the test corpus. Almost 36.6% of the mouth images are not selected by the unit selection. 93.7% of the non-used mouth images are labeled with silence. Given 1457 test sentences, the logarithm of the selection frequency of the mouth images is shown in Fig. 3(b), where the mouth images are sorted in the order of selection frequency.

The performance of the clustering algorithms is measured by PSNR between the final database and the original database as shown in Fig. 4. The black curve represents the PSNR between the final database clustered by LBG and the original database. The red one corresponds to RSST-based clustering and the blue



(a) The number of non-used mouths (b) Logarithm of selection frequency

Fig. 3. Results of probability determination. (a) The number of non-used mouth images given the number of input sentences. (b) Logarithm of the selection frequency determined by unit selection from 1457 test sentences.

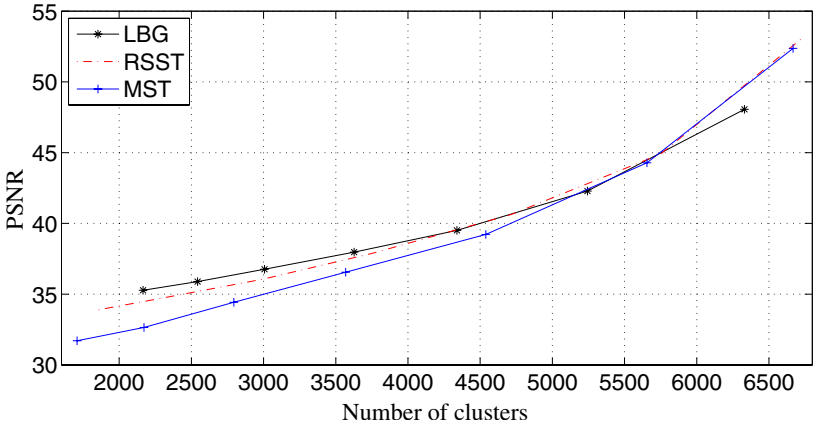


Fig. 4. PSNR Performance of clustering algorithms

one corresponds to MST-based clustering. LBG-based clustering performs better than the others, if the clustered database contains less than 5400 mouth images.

After clustering the database, the final database is compressed by JPEG. The size of the final database depends on the number of mouth images in the final database and the resolution of the talking head image. For example, 3000 mouth images need $29MB$ storage space in PAL format or $8MB$ in CIF format.

In order to evaluate the proposed clustering algorithms subjectively, we generated animations by using the original database and the final database clustered by MST, RSST and LBG with different size. These animations are shown to the viewers and they were asked to score the quality of the animations generated by using the final databases. The results of the informal subjective tests show

that the database with not less than 3000 mouth images clustered by LBG, can synthesize animations without loss of naturalness.

To see the animations according to Fig. 4, the reader is encouraged to visit `http://www.tnt.uni-hannover.de/project/facialanimation/demo/minidb/`

5 Conclusion

In this paper we have presented a prototype framework for minimizing the database of unit selection so that the real-time talking head for Internet applications is possible. The database reduction is carried out in two steps: First, the database with useful mouth images is clustered; Second, JPEG is used to compress the final database. MST, RSST and LBG-based clustering algorithms are proposed and evaluated.

Experimental results show that the proposed methods can reduce the database efficiently. LBG-based clustering algorithm performs better than others given a small size of database. According to the subjective tests, the animations can be generated by a small database with at least 3000 mouth images without loss of naturalness.

Furthermore, because the non-used mouth images are discarded from the database, the number of candidates are reduced in the Viterbi search, so that the unit selection performs faster and more efficiently.

References

1. Theobald, B., Fagel, S., Bailly, G., Elisei, F.: LIPS2008: Visual Speech Synthesis Challenge. In: Proc. Interspeech 2008, Brisbane, Australia, September 2008, pp. 2310–2313 (2008)
2. Liu, K., Ostermann, J.: Realistic Facial Animation System for Interactive Services. In: Proc. Interspeech 2008, Brisbane, Australia, September 2008, pp. 2330–2333 (2008)
3. LIPS 2008: Visual Speech Synthesis Challenge (2008), `http://www.lips2008.org/`
4. Liu, K., Ostermann, J.: Realistic Talking Head for Human-Car-Entertainment Services. In: Proc. IMA 2008 Informationssysteme fuer mobile Anwendungen, Braunschweig, Germany, September 2008, pp. 108–118 (2008)
5. Weissenfeld, A., Urfalioglu, O., Liu, K., Ostermann, J.: Robust Rigid Head Motion Estimation based on Differential Evolution. In: IEEE Proc. ICME 2006, Toronto, Canada, July 2006, pp. 225–228 (2006)
6. Jolliffe, I.: Principal Component Analysis. Springer, New York (1989)
7. Prim, R.C.: Shortest connection networks and some generalizations. Bell System Technical Journal 36, 1389–1401 (1957)
8. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. IEEE Trans. Commun. COM-28, 84–95 (1980)
9. Morris, O.J., Lee, M.J., Constantinides, A.G.: Graph theory for image analysis: an approach based on the shortest spanning tree. In: Proc. Inst. Electr. Eng., vol. 133, pp. 146–152 (1986)