Talking Faces

Jun-yong Noh, Ulrich Neumann
Computer Science Department
Integrated Media Systems Center
University of Southern California

ABSTRACT

We present a method to generate mouth animations coupled with moderate movements of the head. Our method uses a coarse 2D textured mesh representation of images. With hardware texture mapping, we obtain 30 frames/sec animations on modest PC platforms. The synthesized animation sequence is realistic and lively. Radial basis functions are used in the underlying animation engine. At each phoneme transition, the positions of the mesh nodes are determined by RBF coefficients associated with a set of sparse feature points. Blending textures from key frame images on the displaced mesh produces animation sequences of talking faces.

1. INTRODUCTION

Talking head animation research attracts attentions for its application potential. It can be applied to synthesize an intelligent desktop agent, a virtual friend, and an avatar either in a chat room, or in a low bit rate teleconferencing setting. The application of facial animation research to such systems requires easy creation of individual models, visually pleasing facial animation quality, and modest computing requirements. In achieving the above goals, facial animation synthesis often takes two approaches: 3D mesh based geometry deformations and 2D image manipulations.

In a typical 3D mesh approach, a person-specific model is prepared by deforming a generic mesh in a preprocessing step [8][15][18][24]. The generic mesh contains all the animation parameters necessary for the subsequent person specific animations. The model is animated by mesh node displacements based on motion rules specified by deformation engine such as vector muscles [25][26], spring muscles [15][19], free form deformations [13], volume morphing [10], or simple interpolation [18]. Texture mapping [10][15][18] is employed to improve realism with characteristics such as skin wrinkles that are hard to achieve by geometric deformations alone. Such 3D approaches hold promise for realistic interactions between users in future immersive teleconferencing systems.

In the constrained case of animating only the frontal face with small head rotations and translations, the 3D approach and the creation of the person specific model may be avoided. A 2D image-based approach bypasses the need for 3D face and head modeling. 2D based approaches are also attractive for mouth animation. The 3D mouth shapes created during speech are difficult to model and generalize. Tongue modeling and behavior also demand an elaborate mechanism [5][22]. In contrast, image-based approaches simplify these issues because the appearance and animation model can be readily captured in the form of images.

Ezzat et al. present a text to audiovisual speech synthesizer using dense pixel-based image warping and morphing between two key-frame viseme images [9]. Despite the fact that co-articulation effects are ignored, the synthesized talking animation appears very natural. Other approaches morph the images based on 2D meshes and associated feature points rather than relying on dense pixel warps. For instance, Gao et al. synthesize new mouth shapes by linear combinations of several base images [11]. Texture from the base images are blended together with weights determined by the tracked location of feature points on a triangular patch. Similarly, Koufakis et al. use three basis images captured from different views and synthesize slightly rotated views of the face by linear combination of these basis images [14].

In lieu of creating completely new images, existing images are combined to synthesize talking faces. Bregler et al. extract tri-phone images from video footages and constructs the database [3]. New utterances are synthesized by concatenations of the tri-phone segments that reproduce co-articulation effects. Cosatto et al. collect various image samples of a segmented face and parameterize them to synthesize a talking face [6]. By modeling different parts of the face from different sample segments, synthesized talking faces also exhibit emotions from eye and eyebrow movements and forehead wrinkles.

Our approach pursues the goal of creating talking faces with moderate head movements. Specifically, we opt to morph images using a coarse regular grid mesh placed on images. There are several rationales motivating this approach. Methods that exploit a collection of existing sample images [3][6] must search their database for the most appropriate segments to produce a needed animation. The success of the synthesized sequence heavily relies on the pre-constructed database. The lack of any ability to generate new images degrades the synthesized animation quality when newly encountered expressions or phrases are not already in the database.

Pixel image morphing operations [9] can consume large computing power. For example, the 60,000 pixels in a small 300 x 200 image region is a huge amount of data to process per frame. Consequently, animation is often restricted to only the mouth region. Incorporating wrinkles and expressions over the whole face forces the image size to be smaller to maintain real time performance.

Mesh based texture mapping techniques [11][14][17] are advantageous in that warping is computed for relatively few control points. Applying textures to 2D image meshes exploits the hardware texture mapping common in most graphics cards today. However, in many prior methods, mesh preparation for each individual is a tedious process. There are automated ways to determine optimal triangulations [12][21], but as pointed out in [14], automated triangulations do not capture the facial anatomy so manually prepared meshes are often preferred.

As mentioned earlier, we use a coarse regular mesh for both head movements and mouth animations. The mesh is constant for each individual. Realistic mouth animation is made possible by using radial basis functions (RBF) [20] in the underlying deformation engine. RBF allow sparsely scattered control points to control the denser grid point positions. Unlike other methods in which mesh resolutions are determined by the number and position of control points, neither of these issues constrains the shape or resolution of the textured mesh. With the RBF we gain the flexibility of locating the controlling feature points where we wish while keeping a simple regular textured mesh.

In fact, there have been attempts to apply RBF to create facial expressions [1]. Most approaches warped a single image to deform the face. However, the quality obtained from a single image deformation drops as more and more distortions are required. Also, single images lack information exposed during animation, e.g., mouth opening. Approaches without RBF [17] using only single images have similar pitfalls. We capture multiple images and use two as key frames to produce realistic animations.

The remainder of the paper proceeds as follows: We discuss the radial basis functions briefly in section 2. Section 3 describes implementation specifics for creating mouth animations using RBF. We discuss limitations to our approach and future works in section 4.

2. RADIAL BASIS FUNCTIONS

We compute the coefficients of the RBF that represent the best mapping between source and target feature points. The rest of the nodes in the facial mesh are transformed using these coefficients. The RBF approximation is of the form

$$F(x) = \sum_{i=1}^{N} c_i h(\| x - x_i \|) \qquad (1)$$

When Gaussian $h(r) = e^{-(\frac{r}{s})^2}$ is used, (1) becomes

$$\vec{x}^{target}{}_j = F(\vec{x}^{source}{}_j) = \sum_{i=1}^{N} c_i e^{-(\frac{\|\vec{x}^{source}{}_j - \vec{x}_i\|}{s_i})^2} \qquad (2)$$

where s is called a stiffness constant that regulates the local and global effects of the feature points, r is the Euclidean distance between a feature point and an input point, N is the number of feature points, and 1 = j =N. The dimension of x is 2 (i.e., the dimension of the feature point positions). Stiffness coefficient s is determined as suggested by Eck [7] for softer deformation where feature points are widely scattered and stronger where closely located.

$$s_i = \max_{j \neq i} \| \vec{x}^{source}{}_j - \vec{x}_i \| \qquad (3)$$

Substituting N feature points into equation (2) results in a linear system whose solution is of the form

$$\vec{c} = H^{-1} \vec{x}^{target} \qquad (4)$$

The solutions of equation (4) assume no spurious data, however, feature points may be noisy. We use the method suggested in [23] to replace the matrix H by $(H + lI)$ in equation (4)

$$\vec{c} = (H + lI)^{-1} \vec{x}^{target} \qquad (5)$$

where l is a "small" parameter (we set to 0.01), proportional to the noise. The system is easily solved by matrix inversion to obtain the coefficient set and the motion vectors of the nodes on the mesh are retrieved.

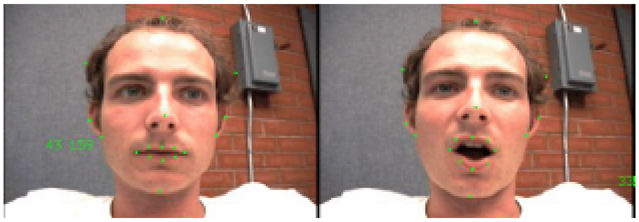

Figure 1. Selected features are shown on closed mouth (a) and AA (b) visemes. 8 points on the lips, 8 points on the boundary of the face, and 1 point on the nose control the RBF deformations.

3. APPLICATION

To demonstrate the capability of the radial basis functions for easy but powerful image synthesis, we illustrate some mouth animations with moderate head movements. Snapshots of the 26 viseme images are taken. The number of viseme images may differ depending on how the phonemes are classified into viseme groups. We divide the 39 phonemes into 26 visemes, similar to [3]: (1) CH, JH, SH, ZH (2) K, G, N, L (3) T, D, S, Z (4) P, B, M (5) F, V (6) TH, DH (7) W, R (8) HH (9) Y (10) NG (11) EH (12)

EY (13) ER (14) UH (15) AA (16) AO (17) AW (18) AY (19) UW (20) OW (21) OY (22) IY (23) IH (24) AE (25) AH and (26) CM for closed mouth.

We manually label feature points on each image, 8 on the boundary of the face, 1 on the nose, and 8 on the lips (figure 1). To ensure the convexity of the RBF interpolation, the four corners of the mesh are also used as feature points. If viseme images are extracted from the video sequences, feature points selections can be automated using feature tracking systems [16][27]. A 3 x 4 mesh grid for the face movements and a 14 x 13 mesh grid for the mouth movements are defined implicitly on the image (figure 2). One face mesh grid consists of 60 x 60 pixels and one mouth mesh grid consists of 6 x 6 pixels.

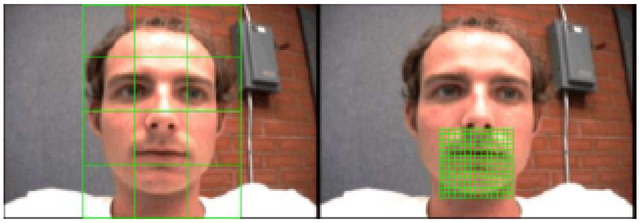

Figure 2. A coarse mesh grid is placed on the face (a) and a finer mesh is placed on the mouth region (b).

To produce an animation, a user types in arbitrary text that is decomposed into phonemes by the CMU dictionary [4] and, in turn, mapped to corresponding viseme images. At transitions from one phoneme to the other, the associated two visemes are used as key frames and RBF coefficients are computed for each direction. RBF coefficients enable us to construct a displacement map for the mesh nodes. Interpolating the displacement map and blending the key frame textures synthesize the in-between frames.

$$I^{new} = (1-c)I^1(c) + cI^2(1-c) \qquad (5)$$

I^{new} is the synthesized image, I^1, I^2 are the two key frames, and c is the interpolation coefficient. This approach is similar to [9], but we use a coarse mesh grid, take advantage of hardware texture mapping, need no pre-computation, and achieve 30 frames/sec on a modest PC.

Mouth animation on a stationary face is unnatural. We allow the movements of the face when we capture the snapshots of the visemes and use face images without normalizing the frontal position. This approach may have the drawback of coupling a specific viseme with a specific head position, but moderate movements of the head during speech are common. A synthesized talking person looks much more alive this way than in a stationary frontal pose. Clearly, stabilized pose could also have head motions added manually. Some examples are shown in figure 3 and 4. MPEG movies can be found at http://scf.usc.edu/~noh.

4. DISCUSSIONS AND FUTURE WORKS

We do not consider co-articulation effects in the sample animations. If this technique were applied to performance driven animation for low bit rate teleconferencing [10][14], however, feature tracking may provide clues to the natural evolution of mouth shapes on the subject. Alternatively, compact parameterization can also be used for co-articulation effects [6].

Integration of a text to speech module is beyond the current scope of our system. Consequently, the absence of the timing information for each phoneme transition is apparent in the sample animations. Nevertheless, subjective tests we performed on 10 people revealed that all could read the lip animations generated by simple text input.

We plan to synthesize wrinkles in the future research. It seems straightforward to apply the technique for mouth animation to generating wrinkles. A systematic method for generating small head motions is plausible. Ultimately, it appears feasible to synthesize natural facial animations based on a small set of feature points and base images [2].

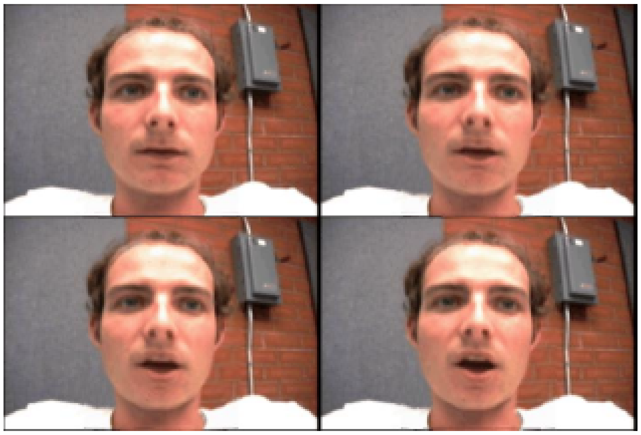

Figure 3. Sequence of the synthesized images is shown. Closed mouth and AA visemes shown in figure 1are used as key-frames.

5. CONCLUSION

We describe a simple method for generating mouth animations using snapshots of key viseme images. The synthesis of frames is based on RBF deformation and texture blending. These methods produce animations in real time. Low bit rate teleconferencing may be possible utilizing this approach.

6. ACKNOWLEGEMENTS

This work was supported by the Annenberg Center at USC. Additional Support and research facilities are provided by the NSF through its ERC funding of the Integrated Media Systems Center. We also thank Intel and IBM for support.

7. REFERENCES

[1] N. Arad, N. Dyn, D. Reisfeld, Y, Yeshurun, Image Warping by Radial Basis Functions: Application to Facial Expressions, CVGIP: Graphical Models and Image Processing, vol. 56, No. 2, March, 1994, pp. 161-172

[2] D. Beymer, A. Shashua, T. Poggio, Example Based Image Analysis and Synthesis, MIT Technical report from http://www.ai.mit.edu/publications/pubsDB/pubsDB/onlinehtml, 1993

[3] C. Bregler, M. Covell, M. Slaney, Video Rewrite: Driving Visual Speech with Audio, Siggraph proceedings, 1997, pp. 353-360

[4] CMUDictionary http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[5] M.M. Cohen, D.W. Massaro, Modeling Coarticulation in Synthetic Visual Speech, in Models and Techniques in Computer Animation, M. Magnenat-Thalmann and D. Thalmann (eds.), Tokyo, 1993, Springer Verlag

[6] E. Cosatto, H. P. Graf, Sample-Based Synthesis of Photo-Realistic Talking Heads, In proceedings of Computer Animation 1998, pp. 103-110

[7] M. Eck, Interpolation Methods for Reconstruction of 3D Surfaces from Sequences of Planar Slices, CAD und Computergraphik, Vol. 13, No. 5, Feb. 1991, 109 – 120

[8] R. Enciso, J. Li, D. Fidaleo, T-Y. Kim, J-Y. Noh, U. Neumann, Synthesis of 3D Faces, Digital and Computational Video (DCV 99)

[9] T. Ezzat, T. Poggio, Mike Talk: A Talking Facial Display Based on Morphing Visemes, In proceedings of Computer Animation 1998, pp. 96-102

[10] D. Fidaleo, J-Y. Noh, T. Kim, R. Enciso, U. Neumann, Classification and Volume Morphing for Performance-Driven Facial Animation, Digital and Computational Video (DCV99)

[11] L. Gao, Y. Mukaigawa, Y. Ohta, Synthesis of Facial Images with Lip Motion from Several Real Views, In proceedings of Automatic Face and Gesture Recognition, 1998, pp. 181-186

[12] L.Guibas, D. Knuth, M. Sharir, Randomized incremental construction of delaunay and voronoi diagrams, Algorithmica 7 (4) 1992, pp. 381-413

[13] P. Kalra, A. Mangili, N. M. Thalmann, D. Thalmann, Simulation of Facial Muscle Actions Based on Rational Free From Deformations, Eurographics 1992, vol. 11(3), pp. 59–69

[14] I. Koufakis, B.F. Buxton, Very low bit rate face video compression using linear combination of 2D face view and principal components analysis, Image and Vision Computing 17, 1999, pp. 1031-1051

[15] Y. C. Lee, D. Terzopoulos, K. Waters. Realistic face modeling for animation. Siggraph proceedings, 1995, pp. 55-62

[16] T. Maurer, C. von der Malsburg, Tracking and learning graphs of image sequence of faces, In Proceedings of International Conference on Artificial Neural Networks, Bochum, Germany, 1996

[17] W. Perng, Y. Wu, M. Ouhyoung, Image Talk: A Real Time Synthetic Talking Head Using One Single Image with Chinese Text-To-Speech Capability. IEEE 1998, pp. 140-148

[18] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin, Synthesizing Realistic Facial Expressions from Photographs, Siggraph proceedings, 1998, pp. 75-84

[19] S. Platt, N. Badler, Animating facial expression. Computer Graphics, 1981, vol. 15(3) pp. 245-252

[20] M. J. D. Powell, Radial basis functions for multivariate interpolation: a review. Algorithms for Approximation, Clarendon Press, Oxford, 1987

[21] J. Shewchuk, Triangle: Engineering a 2D quality mesh generator and delaunay triangulation, in Proceedings of the First Workshop of Applied Computational Geometry, ACM, 1996, pp. 124-133

[22] M. Stone, Toward a model of three-dimensional tongue movement. Journal of Phonetics, 1991, vol. 19, pp. 309-320

[23] A.N. Tikhonov, V.Y. Arsenin, Solution of Ill-Posed Problems and the regularization method. Soviet Math. Dokl., 4:1035-1038, 1963

[24] F. Ulgen, A step Toward universal facial animation via volume morphing, 6th IEEE International Workshop on Robot and Human communication, 1997, pp. 358-363

[25] K. Waters, J. Frisbie, A Coordinated Muscle Model for Speech Animation, Graphics Interface, 1995 pp. 163 – 170

[26] K. Waters. A muscle model for animating three-dimensional facial expression. In Maureen C. Stone, editor, Computer Graphics (Siggraph proceedings, 1987) vol. 21 pp. 17-24

[27] S. You, Y. Zhang, Z. Pen, G. Xu, "A Multi-Pose Face Recognition System", *Journal of Software*, 11, 1996.

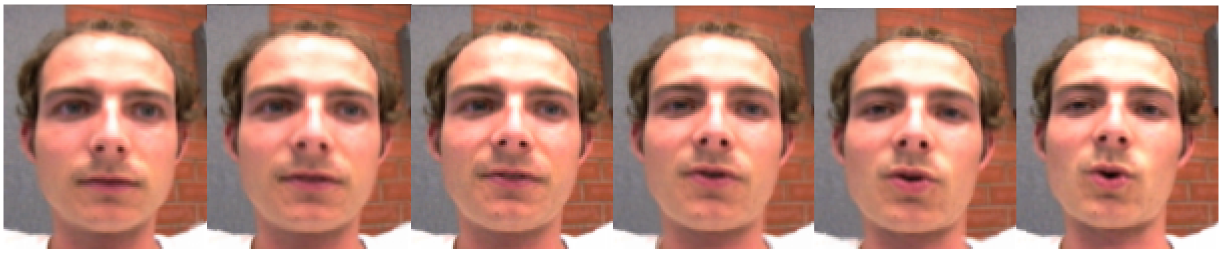

Figure 4. Synthesized transitions from closed mouth to F and to OW when pronouncing ‘phone’