Appeared in ***Proc. of IEEE Computer Animation and Social Agents (CASA) 2004***

Automatic Dynamic Expression Synthesis For Speech Animation

Zhigang Deng
Dept. of CS
zdeng@usc.edu

Murtaza Bulut
Dept. of EE
mbulut@usc.edu

Ulrich Neumann
Dept. of CS
uneumann@graphics.usc.edu

Shri Narayanan
Dept. of EE
shri@sipi.usc.edu

Integrated Media System Center
University of Southern California, Los Angeles
http://graphics.usc.edu

Abstract

Although a large amount of research has been done in speech animation, both 2D and 3D, one shortcoming of current speech animation methods is that they cannot generate dynamic expressions automatically. In this paper, an automatic technique for synthesizing novel dynamic expression for 3D speech animation is presented. After a Phoneme-Independent Expression Eigen-Space (*PIEES*) is extracted from expressive facial motion capture data by Principal Component Analysis (PCA), texture synthesis methods are used to synthesize novel dynamic expression from the PIEES. Control for expression intensity dynamics is also provided for animators via expression intensity curves.

Keywords: Facial Animation, Expressive Synthesis, Speech Animation

1. Introduction

Human expression is a challenging research topic not only in the graphics community, but also in other fields, including artificial intelligence and psychology. However, expression is critical to realistic facial animation, and automatically synthesizing expressive facial animation still remains an open problem. In this paper, a texture synthesis based approach for automatically synthesizing dynamic expressions for 3D speech animation is presented. This approach is compatible with various presented speech animation methods: expression is added on top of the animation synthesized by speech animation methods [1,2,3,4].

This automatic dynamic expression synthesis approach can be used for various applications. For example, in MPEG-4 facial animation, only textual descriptions are used to define high-level expression parameters [5], and no expression synthesis approach is specified. This approach can be used to automatically reconstruct expressive MPEG-4 facial animation upon receiving high-level expression parameters in FAPs (Facial Animation Parameters). This approach can also be used for web-based avatars [6] or low-bandwidth facial animation applications [7].

2. Related Work

An overview of various facial animation techniques can be found in [8]. The research topics closely connected to this work are speech animation and expressive facial animation. A brief review of recent work in these two fields and the motivations of our work are described next.

2.1 Speech Animation

Many techniques have been presented for synthesizing speech animation according to novel audio/text input. For example, phoneme-base methods [1,3,9] require animators to design key mouth shapes beforehand, and then hand-generated smooth functions or co-articulation rules are used to generate speech animation. Physics-based approaches [2] employ the laws of physics to drive the mouth movement. Bregler et al. [10] present "video rewrite" to synthesize novel speech animation by re-combining frames in existing video footage using triphone rules. Ezzat et al. [11]

present a multidimensional morphable model trained from phoneme-aligned video footage and successfully synthesize video-realistic speech animation given novel audio input. Instead of driving animation by phonemes, Kshirsagar and Thalmann [4] present a syllable-based approach to synthesize speech animation from a captured "visyllable motion" database. However, most of the above speech-animation approaches cannot automatically synthesize dynamic expressions.

2.2 Expressive Facial Animation

Cassell et al. [12] present a rule-based automatic system that generates expressions and speech for multiple conversation agents. In this work, FACS [13] is used to denote static facial expressions. Noh and Neumann [14] present an "expression cloning" technique to transfer existing expressive facial animation between different 3D face models. This technique and the further extended technique [15] are very useful to transfer expressive facial motion. However, they are not generative; they cannot be used for generating novel facial animation. Chuang et al. [16] learn a facial expression mapping/transformation from training footage using bilinear models, and then this learned mapping is used to transform novel video of neutral talking to expressive talking. In their work, the expressive face frames retain the same timing as the original neutral speech, which does not seem plausible in all cases. Cao et al. [17] present a motion editing technique that applies Independent Component Analysis (ICA) onto recorded expressive facial motion capture data and then perform more editing operations on these ICA components, interpreted as expression and speech components separately. This approach is used only for editing existing expressive facial motion, not for the purpose of synthesis. Zhang et al. [18] present a geometry-driven technique for synthesizing expression details for 2D faces. This method is used for static 2D expression synthesis, and the applicability of this method to animate images and 3D models has not been established.

Blanz et al. [19] present a novel technique to reanimate faces in images and video by learning an expression and mouth shape space from scanned 3D faces. This approach addresses both speech and expressions, but it only considers several static expression poses that are not enough to synthesize realistic dynamic expressive motion. Brand [20] learns an HMM-based facial control model by an entropy minimization algorithm from training audio/video data, and then novel facial motion (speech and expression) is synthesized by sampling from this generative model. In his work, audio features (a mixture of LPC and RASTA-PLP are used [20]) are implicitly mapped to some expressive facial configurations, and the implicit assumption is that there is a clear mapping between audio features and expressions. In the speech recognition literature [21,22,23], it is known that if only audio features are used as inputs, the average accuracy of the state-of-the-art emotion recognizers is about 70%. As such, expressive motion may not be learned correctly from only several audio features. Brand [20] also learns the audio-motion mapping as a monolithic system. Therefore, it is difficult for animators to intuitively control speech and expression separately. Kshirsagar et al. [6,24] present a PCA-based method for synthesizing expressive speech animation. In their work, static expression configurations are embedded in an expression and viseme space, constructed by PCA. Expressive speech animation is synthesized by weighted blending between expression configurations (corresponding to some points in the expression and viseme space) and speech motion. Besides the above work, significant effort has been done in expressive virtual characters in complex scenarios [25,26,27].

In this paper, an automatic approach for synthesizing *dynamic* expressions on 3D speech animation is presented. This approach is compatible with other speech animation techniques. On top of the speech animation synthesized by existing speech animation techniques, this approach synthesizes a novel dynamic expression sequence and adds it to the speech animation. First, an actress with motion capture markers on her face is captured, speaking a custom corpus with different expressions. By removing phoneme-dependent content from the expressive facial motion data, *Phoneme-Independent Expressive Motion Signals* (*PIEMS*) are extracted. Then these high dimensional motion signals are further used to construct a three dimensional *Phoneme-Independent Expression Eigen-Space* (*PIEES*)

using EM-PCA [28]. Based on the observation that an expression sequence is just a continuous curve in the three-dimensional PIEES, novel expressive motion is synthesized by texture-synthesis techniques [29,30].

This work shares similarities with [24,6], but the largest distinction of our work is that expressions are treated as a dynamic process, not as static poses as in [24,6]. In general, the expression dynamics include two aspects: (1) *Expressive Motion Dynamics* (EMD): even in an invariant level of anger, they seldom keep their eyebrows at the same height for the entire duration of the anger. As such, expressive motion is a dynamic process, not statically corresponding to some fixed facial configurations. (2) *Expression Intensity Dynamics* (EID): both the intensity of their expressions and the type of expression may vary over time, depending on many factors, including speech contexts.

Varying blending weights over time in [24,6] can simulate the EID, but the EMD are not modeled, because the same static expressive facial configurations are used. An automatic technique for synthesizing dynamic expressions, including EMD and EID, is presented in this paper. The EMD are embodied in the constructed PIEES as continuous curves. The optional expression-intensity control is used for simulating EID, similar to [24,6]. This approach is flexible, allowing the animators to use any desired speech animation technique.

3. System Overview

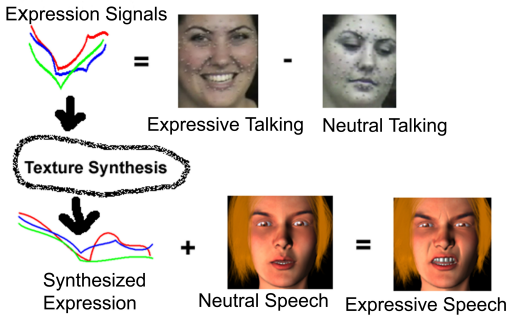

Figure 1: Illustration of the dynamic expression synthesis pipeline.

Figure 1 illustrates the pipeline of this dynamic expression synthesis approach. In the analysis stage (top row, Fig 1), PIEMS (Phoneme-Independent Expressive Motion Signals) are extracted by removing phoneme-dependent content from expressive speech motion data. These are further used to construct PIEES by PCA. In the synthesis stage (bottom row, Fig 1), texture synthesis is used to generate a novel dynamic expression sequence, and blend it with a neutral speech animation. The remainder of the paper is organized as follows: first, data acquisition, processing and analysis are described, then how to synthesize novel dynamic expressions and how to simulate expression intensity dynamics over time are described. The final part describes the results and conclusions.

4. Data Acquisition and Processing

A VICON motion capture system [31] with three cameras (left of Fig 2) is used to capture the expressive facial motion data with 120Hz sampling frequency. With 102 markers on her face (right of Fig 2), an actress is directed to speak a custom phoneme-balanced corpus four times. Each time the same corpus is spoken with different emotions. In this work, total four emotions (neutral, happy, angry, and sad) are considered. The actress is asked to speak the corpus with *full* specific emotion all the time. The markers' motion and aligned audio are captured by the system simultaneously.

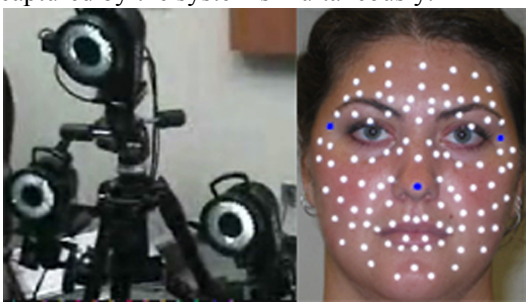

Figure 2: Illustration of motion data acquisition. The left shows the cameras of the motion capture system. The right shows the markers used in the data acquisition stage.

After the data are captured, the captured motion data are normalized. Then, the festival speech recognition system [32] is used to perform phoneme alignment on the captured audio. Since the same corpus is used for different captured expressions, the phoneme sequences, except for their timing, are the same. Based on this observation, a phoneme-

based time warping and resampling (super-sample/down-sample) is applied to the expressive capture data to make them align strictly with neutral data frame by frame. In this step, eyelid markers are ignored. Figures 3-4 illustrate this time-warping procedure for a short piece of angry data. Then, subtracting neutral motion from aligned expressive motion generates pure expressive motion signals. Since they are strictly phoneme-aligned, we assume that the above subtraction removes "phoneme-dependent" content from expressive speech motion capture data. As such, the extracted pure expressive motion signals are *Phoneme-Independent Expressive Motion Signals* (PIEMS).

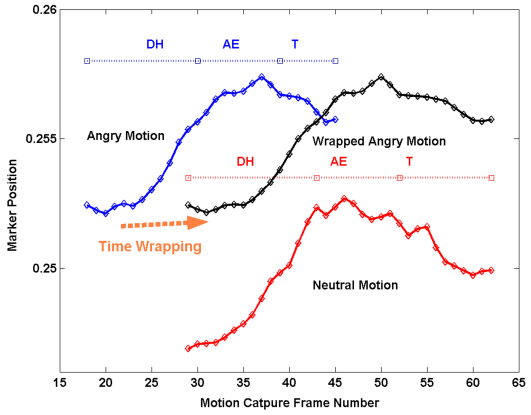

Figure 3: Illustration of phoneme-based time-warping for the Y position of a particular marker. Although the phoneme timings are different, the warped motion (black) is strictly frame aligned with neutral data (red).

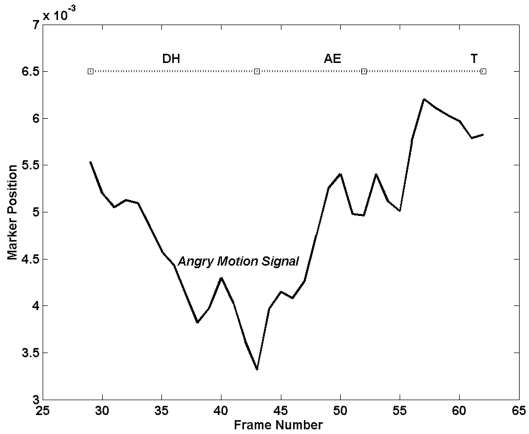

Figure 4: Extracted phoneme-independent angry motion signal from Fig 3.

The extracted PIEMS are high dimensional when the 3D motion of all markers are concatenated together. As such, all the PIEMS are put together and reduced to three dimensions, covering 86.5% of the variation. The EM-PCA algorithm [28] is used here, because EM-PCA can extract eigen-values and eigenvectors from large data sets more efficiently than regular PCA. In this way, we find three-dimensional PIEES where expression is a continuous 3D curve. Figures 5-6 illustrate the PIEES and the PIEMS. Note that the personality of the captured subject may be irreversibly reflected in the PIEES, and only four basic expressions are considered. Building person-independent expression eigen-space and modeling the universal expression space is beyond this work.

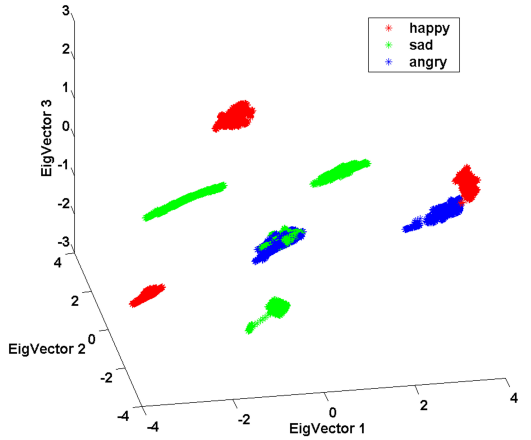

Figure 5: Plot of three expression signals on the PIEES. It shows that sad signals and angry signals intersect in some places.

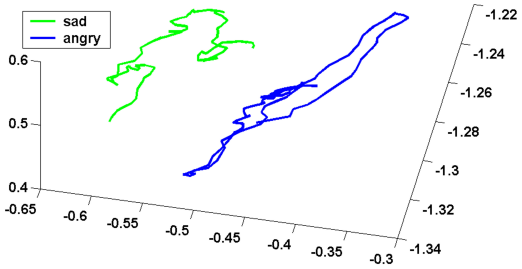

Figure 6: Plot of two expression sequences in the PIEES. It shows that expression is just a continuous curve in the PIEES.

5. Dynamic Expression Synthesis

From Fig 6, we observe that expression is just a continuous curve in the low-dimensional PIEES. *Texture Synthesis*, originally used in 2D image synthesis, is a natural choice for synthesizing novel expression sequences. Here

non-parametric sampling methods [29,30] are used. The patch-based sampling algorithm [30] is chosen, due to its time efficiency. Its basic idea is to grow one texture patch (fixed size) at a time, randomly chosen from qualified candidate patches in the input texture sample. In this work, each texture sample (analogous to a pixel in the 2D image texture case) consists of three elements: the three coefficients of the projection of a motion vector on the three-dimensional PIEES. Figure 7 sketches this synthesis procedure. The parameters of patch-based sampling [30] for this case: patch size = 30, the size of boundary zone = 5, and the tolerance extent = 0.03. For more details of this synthesis approach, see [30].

To ensure the synthesis quality, a high-resolution synthesis strategy is employed. As mentioned in data acquisition section, the original data are captured at a 120 Hz sampling rate. The above texture synthesis procedure is done on high-resolution (120 Hz) captured data and later down-sampled to ordinary video rate (30 samples/sec).

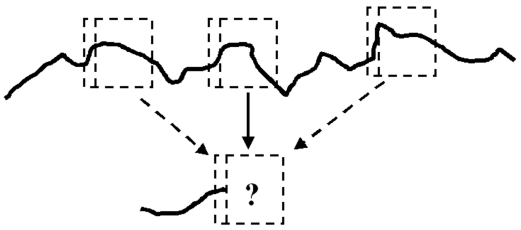

Figure 7: Illustration of the patch-based sampling procedure.

6. Intensity Dynamic Control

As mentioned in the data acquisition section, the expressive facial motion data used for extracting the PIEES are captured with full expressions. However, in real-world applications, humans usually vary their expression intensity over time. In this work, an optional expression-intensity curve scheme is provided to intuitively simulate the EID. Ideally, the EID (expression intensity curves in this work) should be automatically extracted from the given audio by emotion-recognition programs [21,22,23]. The optional expression-intensity control is a manual alternative to this program.

Basically, an expression intensity curve can be any continuous curve in time versus expression-intensity space, and its range is from 0 to 1. Zero represents "no specific expression" (neutral) and one represents "full specific expression". By interactively controlling expression-intensity curves, animators can conveniently control expression intensities over time.

7. Results and Evaluation

To test this approach, a female subject was recorded while she recited lines from a Shakespeare play (not included in the training corpus) with different expressions. Then, this audio is used to synthesize several expressive speech animations: after neutral speech animation is generated, expressions synthesized by this expression synthesis approach are added onto the speech animation. Figure 8 shows some frames of synthesized expressive speech animation. For more results, please see accompanying videos.

Figure 8: Some frames of synthesized expressive speech animation.

We compare synthesized expressive facial animation with the ground-truth video in Figure 9. Taken into consideration that the speech synthesis process is not yet perfect, the expression in the synthesized speech animation closely matches that of ground-truth video. The used 3D face model is a NURBS model, composed of 46 blend shapes. The method presented in [33] was used to map markers' motion to the weights of blend shapes, and eye motion was synthesized by the approach presented in [34]. Hair was modelled and rendered using the approach presented by Kim and Neumann [35].

8. Conclusions and Future Work

An automatic technique for synthesizing dynamic expressions for 3D speech animation is presented in this paper. Improving on previous expression synthesis work [6,24], this approach models both the EMD and EID. In this work, the EMD are embodied in the PIEES and the EID is provided by expression intensity curves. The advantages of this approach are that it works in conjunction with any speech animation method. This approach is simple and also benefits from the time efficiencies of texture synthesis techniques.

This approach can be used for reconstructing MPEG-4 expressive facial animation given high-level expression parameters and in low bandwidth facial animation applications [6,7]. A limitation of this approach is that the interaction between expression and speech is simplified. We assume there is a PIEES extracted by phoneme-based subtraction. Therefore, a large amount of expressive facial motion data are needed to construct the PIEES, because it is hard to anticipate in advance how much data are needed to avoid getting "stuck" during the synthesis process. However, after some animation is generated, it is easy to evaluate the variety of synthesized expressions and more data can be obtained if necessary.

We are aware that expressive eye motion and head motion are critical parts of expressive facial animation, since the eye is one of the strongest cues to the emotional state of a person. Future work on expressive eye motion and head motion can greatly enhance the realism of expressive facial animation synthesized by this approach. Although this approach is a step toward the goal, further research is needed to discover the principles of human expression creation and perception in order to rigorously solve the automatic expression synthesis problem.

Learning approaches, such as HMMs, will be explored to model expression patterns and interactions between speech and expressions. "Expression graphs" (similar to "motion graphs" [36]) or "expression textures" (similar to "video textures" [37]) will also be investigated in the future work.

Figure 9: Comparison of the synthesized expressive speech motion (left) with the ground-truth one (right).

Acknowledgements

This research has been funded by the Integrated Media System Center/USC, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152. Special Thanks go to J.P.Lewis for data capture, Hiroki Itokazu and Bret St. Clair for model preparation, Pamela Fox and Lisette Garcia for proof reading. We also appreciate many valuable comments from other colleagues in the CGIT Lab/USC.

References

[1] M.M. Cohen and D.W. Massaro, Modeling coarticulation in synthetic visual speech. M. Thalmann and D. Thalmann (editors), *Models and Techniques in Computer Animation*, 1993, 139-156.

[2] K. Waters and J. Frisble. A coordinated muscle model for speech animation. In *Proc. of Graphics Interface'95*,163-170.

[3] JP Lewis, Automated lip-sync: background and techniques. *Journal of Visualization and Computer Animation*, 1991, 118-122.

[4] S. Kshirsagar and N.M. Thalmann. Visyllable based speech animation, In *Proc. of EuroGraphics'03*, 631-640.

[5] J. Ostermann. Animation of synthetic faces in mpeg-4. In *Proc. of IEEE Computer Animation'98*.

[6] S. Kshirsagar, S. Garchery, G. Sannier, and N.M. Thalmann, Synthetic faces: analysis and applications. *International Journal of Imaging Systems and Technology*, 13(1), 2003, 65-73.

[7] I. S. Pandzic, Facial Animation framework for the web and mobile platform. In *Proc. of 7th Int'l conf. on 3D web technology*, 2002.

[8] F. I. Parke and K. Waters. *Computer Facial Animation*, A K Peters, Wellesley, 1996.

[9] C. Pelachaud. Communications and coarticulation in facial animation. *Ph.D. thesis*, 1991, U. of Penn.

[10] C. Bregler, M. Covell, and M. Slaney, video rewrite: driving visual speech with audio. In *Proc. of ACM SIGGRAPH'97*.

[11] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transaction on Graphics, 21(3)*. 2002, 353-360

[12] J.Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated Conversation: rule-based generation of facial expression, in *Proc. of ACM SIGGRAPH'94*, 413-420.

[13] P. Ekman and W. V. Friesen. Unmasking the face: a guide to recognizing emotions from facial cues. 1975.

[14] J.Y.Noh and U. Neumann. Expression Cloning, In *Proc. of ACM SIGGRAPH'01*, 277-288.

[15] H.Pyun, Y. Kim, W. Chae, H.W.Kang, and S.Y. Shin, An example-based approach for facial expression cloning, In *Proc. of SIGGRAPH/EG Symposium on Computer Animation (SCA) 2003*, 167-176.

[16] E.S. Chuang, H. Deshpande, and C. Bregler, Facial Expression space learning, In *Proc. of Pacific Graphics'02*, 68-76.

[17] Y. Cao, P. Faloutsos, and F. Pighin. Unsupervised learning for speech animation. In *Proc. of SIGGRAPH/EG Symposium on Computer Animation (SCA) 2003*.

[18] Q. Zhang, Z. Liu, B. Guo, and H. Shum. Geometry-driven photorealistic facial expression synthesis. In *Proc. of SIGGRAPH/EG Symposium on Computer Animation (SCA) 2003*.

[19] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video, In *Proc. of EuroGraphics'03*, 2003.

[20] M. Brand. Voice Puppetry, in *Proc. of ACM SIGGRAPH'99*, 21-28.

[21] R. Cowie, E. D. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellens, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Proc. Mag.*, 18(1), 2001, 32-80.

[22] V. Petrushin. Emotion in speech: recognition and application to call centers. *Artificial Neu. Net. In Engr.*, 1999,7-10

[23] C.M. Lee, S. Narayanan, and R. Pieraccini. Recognition of negative emotions from the speech signal. In *Proc. of Automatic Speech Recognition and Understanding*, 2003.

[24] S. Kshirsagar, T. Molet, and N.M. Thalmann. Principal Components of expressive speech animation. In *Proc. of Computer Graphics International*, 2001.

[25] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1): 1994, 1-46.

[26] C. Pelachaud and I. Poggio. Subtleties of facial expressions in embodied agents. *J. of Visual. & Com. Anim.*, 13(5), 2002.

[27] T. Rist, M. Schmitt, C. Pelachuad, and M. Bilvi. Towards a simulation of conversations with expressive embodies speakers and listeners. In *Proc. of IEEE Computer Animation and Social Agents 2003*, 5-10.

[28] S. Roweis. EM algorithm for pca and spca. *NIPS'97*, 1997, 137-148

[29] A. Efros and T.K. Leung. Texture synthesis by non-parametric sampling. In *ICCV'99*, 1033-1038.

[30] L. Liang, C. Liu, Y.Q. Xu, B. Guo, and H.Y. Shum. Real-time texture synthesis by patch-based sampling, *ACM Transaction on Graphics (TOG)*, 20(3), 2001.

[31] http://www.vicon.com

[32] http://www.cstr.ed.ac.uk/projects/festival/.

[33] P. Joshi, W.C. Tien, M. Desbrun, and F. Pighin. Learning controls for blend shape based realistic facial animation. In *Proc. of SIGGRAPH/EG Symposium on Computer Animation (SCA) 2003*

[34] Z. Deng, J.P. Lewis, and U. Neumann, Automated Eye Motion using Texture Synthesis, *IEEE Computer Graphics and Applications (CG&A)*, to appear.

[35] T.Y. Kim and U. Neumann, Interactive Multiresolution Hair Modeling and Editing. *ACM Transaction on Graphics*, 21(3), 2002.

[36] L. Kovar, M. Gleicher, and F. Pighin, Motion graphs. *ACM Transaction on Graphics*, 21(3), 2002.

[37] A. Schodl, R. Szeliski, D. H. Salesin, and I. Essa. Video Textures. In *Proc. of ACM SIGGRAPH'2000*, 489-498.

Appeared in ***Proc. of IEEE Computer Animation and Social Agents (CASA) 2004***