FACIAL ANIMATION WITH MOTION CAPTURE BASED ON SURFACE BLENDING

Lijia Zhu and Won-Sook Lee
School of Information Technology and Engineering, University of Ottawa
800 King Edward Ave., Ottawa, Ontario, Canada, K1N 6N5

Keywords: Surface Blending, Feature Point, Genetic Algorithms, Motion Capture, Facial Animation, Consistent Parameterization, MPEG-4, Laser Scanner.

Abstract: This paper proposes a methodology to reconstruct 3D facial expressions with motion capture data. Feature-point based facial animation provides easy control of expression usually by moving surface points using mathematical deformation. However it does not support the high quality of surface animation of the face where the feature points are not present. In this paper, we focus on animating a 3D facial model using only feature points, but keeping the high quality animation by using an expression databank obtained from surface scanning. Firstly, a facial expression databank is prepared by processing raw laser-scanned human face data with a consistent parameterization technique. Secondly, sparse motion capture data is obtained using an optical tracking system. Thirdly, guided by the captured MPEG-4 feature point motions, we find the corresponding surface information in the existing examples in the databank by linear combination of them. The optimized blending weights are obtained implicitly by Genetic Algorithms (GA). Finally, the surface blending result is retargeted into the performer's neutral facial mesh. Consequently, motions of the facial surface points are reconstructed.

1 INTRODUCTION

Facial animation can be driven by the facial motion data captured from a live actor's performance. Williams (Williams, 1990) tracks expressions of a live actor and then he maps 2D tracked data onto a scanned 3D face model. In his work, motion is captured using a single video camera and a mirror which generate multiple views required for vision reconstruction. Since Williams' pioneering work, performance-driven facial animation has been widely studied. Terzopoulos and Waters (Terzopoulos, 1993) estimate the dynamic facial muscle contractions from video sequences of expressive human faces. They use deformable contour models to track facial motions in video images. Then the tracked muscle contractions are used to animate the physically based model. Guenter *et al*. (Guenter, 1998) capture the geometry and texture information on a live actor's face from video streams and produce life-like facial animation based on the captured markers. Given video footage of a person's face, Pighin *et al*. (Pighin, 1999) present techniques to automatically recover the face position and the facial expression from each frame in the video sequence.

The aforementioned works utilize video sequences to drive the facial animation. They reconstruct 3D motion data based on 2D video images. However, it is difficult to capture detailed motions of the face from video sequences. Recently, with more advancing optical motion capture technology (e.g. *VICON*^{TM} system), higher resolution and more consistent facial motions can be captured. Some researchers utilize a *VICON*^{TM} system for facial animation. For example, Kshirsagar *et al.* (Kshirsagar, 2001) use a *VICON*^{TM} system to track feature point motions of a talking person. Then for producing realistic speech animation, they use Principal Component Analysis (PCA) to perform statistical analysis on the actual speech data.

Normally, the motion capture data consists of motions for the sparse feature points. While the feature-point based approach provides easier control for facial animation, it brings the issue of how to generate natural-looking results with merely controlling a few key points on the face.

To tackle this issue, it is better to have additional information on how the facial mesh should move (i.e.

Zhu L. and Lee W. (2007).
FACIAL ANIMATION WITH MOTION CAPTURE BASED ON SURFACE BLENDING.
In *Proceedings of the Second International Conference on Computer Graphics Theory and Applications - AS/IE*, pages 63-68
DOI: 10.5220/0002083700630068

how the facial surface points should move). Mapping the feature point motions into an available databank would be a suitable solution. When an expressive databank is available, it would be very beneficial to make use of that data and surface blending is probably the most suitable animation technique for this case. Some works combine performance-driven approach with the surface blending technique. Kouadio *et al*. (Kouadio, 1998) present an animation system that captures facial expressions from a performance actor. They animate the human face based upon a bank of 3D facial expressions. A linear combination of the key expressions in the bank is used to produce facial expressions. In their work, blending weights are obtained by minimizing *Euclidean* distance between corresponding feature points in the face model and live markers. Chai *et al*. (Chai, 2003) show that facial actions can be created from a pre-processed motion capture database. They develop a facial tracking system to extract animation control parameters from video. They reconstruct facial motions by blending the *K* closest examples in the database.

In this paper, we aim to control facial animation merely with the facial feature point motions. In order to produce realistic motions for facial surface points, we make use of a databank. Then guided by the captured feature point motions, we produce the surface blending result based on the examples in the databank. Here we propose an approach to obtain blending weights implicitly by Genetic Algorithms (GA). Then we retarget the surface blending result onto the neutral model of the performer.

The rest of the paper is organized as follows. Section 2 prepares a facial expression databank. A consistent parameterization technique for processing laser scanned data is also shown in Section 2. Section 3 uses a *VICON*^{TM} system to capture the facial motions performed by the human subject. Section 4 shows a feature-point guided surface blending technique via GA. Section 5 utilizes a facial motion retargeting technique. More experimental results are shown in Section 6. Finally we conclude our paper with Section 7.

2 PREPARE A FACIAL EXPRESSION DATABANK USING A LASER SCANNER

With advanced laser scanning technology, a laser scanner is able to capture the subject's face in very high resolution. In this section, we utilize a *Cyberware*^{TM} *3030* laser scanner to construct a facial expression databank. We scan a person when he performs various expressions and visemes statically. Totally, eleven expressions and fourteen visemes of this person are scanned. Examples of scanned data are depicted in Figure 1. The first row is the example expressions and the second row is the example visemes.

Figure 1: Raw laser scanned data.

However, the raw scan data is so dense, irregular and inconsistent that it can not be directly used for optimal model construction and animation. A common approach to tackle this is to pre-define a generic model with all the necessary structures. Then this generic model is adapted to the raw laser-scanned data. Many works have been done on adapting a generic model to raw laser-scanned data. (e.g. (Zhang, 2004), (Jeong, 2002), and (Kähler, 2002)). Here we adopt the similar idea to parameterize the raw-laser scanned data consistently.

Firstly we prepare a generic model (Figure 2(b)) with predefined feature points. Then we detect the same feature points in the raw laser-scanned data in the semi-automatic way (Figure 2(c)). After that, we deform the generic model using a *Radial Basis Function* (RBF) deformation technique (Figure 2(d)). We then increase the triangle count by applying Loop's subdivision scheme twice (Figure 2(e)). Finally we apply a cylindrical projection technique to further adapt the subdivision result (Figure 2(e)) to the raw laser scanned data (Figure 2(a)). The details of the aformentioned techniuqes can be found in our previous works ((Zhu, May 2006) and (Lee, 2006)). As we can see, the resulting model shown as Figure 2(f) bears very close resemblance to the raw laser-scanned data.

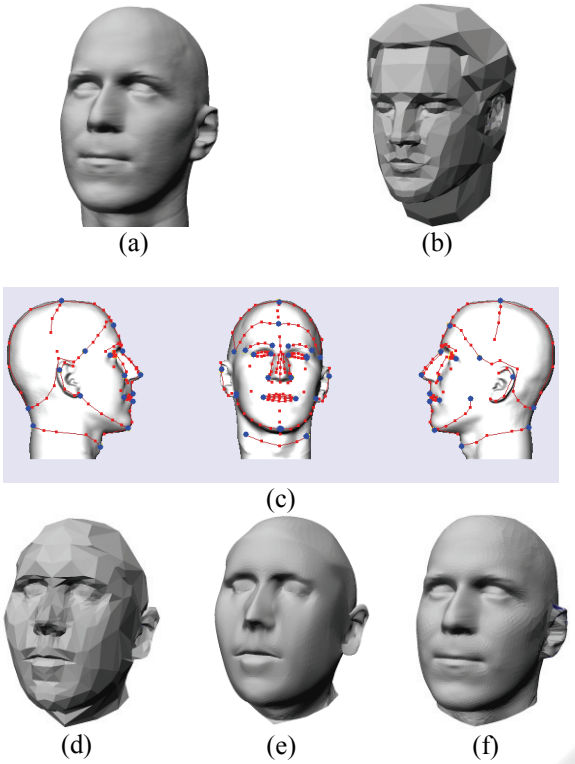

Figure 2: Consistent parameterization technique for processing the raw laser-scanned data (a) Original laser-scanned data (699,392 triangles); (b) Generic model (1,485 triangles); (c) Feature point detection for raw laser-scanned data; (d) Deformed generic model (1,485 triangles); (e) After applying Loop's subdivision twice (23,760 triangles); (f) After cylindrical projection (23,760 triangles).

Using this consistent parameterization technique, we can process all the laser scanned data in the consistent way. Figure 3 shows examples of the resulting consistent meshes. The first row is the example expressions and the second row is the example visemes.

Figure 3: Example consistent meshes of various expressions and visemes.

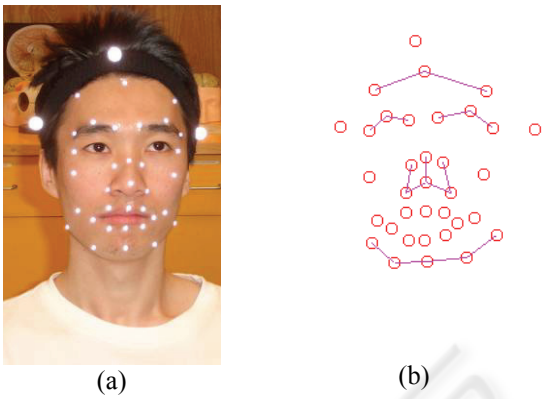

Figure 4: Facial motion capture using a *VICON*^{TM} system. (a) Sparse markers are attached to the subject's face; (b) 3D virtual markers reconstructed by the *VICON*^{TM} system.

3 FACIAL MOTION CAPTURE

We utilize an optical tracking system (*VICON*^{TM}) to capture facial motions performed by a human subject. As shown in Figure 4(a), a set of reflective markers is attached to the human subject's face. A set of 32 markers is used here which is the subset of the feature point set specified in the MPEG-4 standard. Not all MPEG-4 feature points defined in the standard are put reflective markers here since it is difficult to track the markers when they are too close to each other. We do not place markers on the eyelids since those markers may interfere with the subject when performing facial motions. Also we do not put markers in the inner lip region since the subject has trouble in performing the mouth actions. Note that three additional markers on the head mounted jig are used in our motion capture session (Figure 4(a)). Although they do not contribute to facial motions, they are important for removing global movements when analyzing facial motions. With reflective markers on the subject's face, the *VICON*^{TM} system is able to track those markers when the subject performs facial motions. A *VICON*^{TM} software suite is then used to process the 2D data captured from eight high-resolution cameras. The *VICON*^{TM} software suite is able to clean and reconstruct the 3D trajectory data from those 2D motion data. Figure 4 (b) shows the reconstructed 3D marker data of a frame.

The reconstructed 3D data needs to be normalized among frames. The following presented normalization technique is similar to the one proposed by Busso et al. (Busso, 2004). In the normalization step, the first neutral frame is used as

the reference frame. Firstly, the three head jig points are used for removing global head motions. In each frame, these three rigid points define a local coordinate system. Each frame is then rotated to align it with the reference frame. Then we use the nose tip point as the reference point to calculate the translation vector so that the nose tip in each frame aligns with that in the neutral frame. After the normalization step, the facial motion for each marker is calculated using the neutral frame as a reference. Once facial motions for the sparse markers are known, next question confronting us is how to reconstruct motions for facial surface points of the subject.

4 FEATURE POINT GUIDED SURFACE BLENDING VIA GENETIC ALGORITHMS

Given sparse marker motions, the goal of this section is to reconstruct motions for facial surface points of the subject. Mapping sparse marker motions into an available databank is a suitable solution. In Section 2, a facial expression databank is constructed. We make use of that databank here. Guided by captured feature point motions, motions for facial surface points are obtained by blending the examples in the databank using the surface blending technique.

4.1 Convert the Motion Capture Data into the Space of the Databank

Firstly, the sparse motion capture data has to be converted into the space of the databank. Chai et al. (Chai, 2003) scale the extracted animation parameters to ensure that motions from the tracked data have approximately the same magnitude as those in the database. They compute this scale by the ratio of the mouth width between the tracked data and the model in the database.

The following explains how we decide the scale factor. We calculate the major feature point distances in both neutral motion capture data (Figure 5(a)) and the neutral model of the databank (Figure 5 (b)). Then these distances are used to decide the final scale factor. The final scale factor S is given by:

$$S = \frac{1}{4} \times \frac{x_1}{x_1'} + \frac{1}{4} \times \frac{x_2}{x_2'} + \frac{1}{4} \times \frac{y_1}{y_1'} + \frac{1}{4} \times \frac{y_2}{y_2'} \quad (1)$$

The four distance scales contribute equally in deciding the final scale factor S. The averaging function decreases the amount of uncertainty. This scale factor S is then applied on the sparse marker motions. After that, the captured facial motions are in the same space as that of the databank.

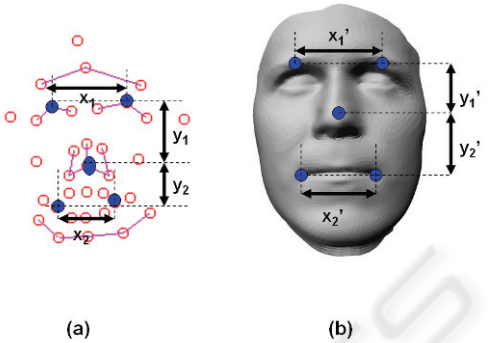

Figure 5: Illustration of calculating the scale factor to convert motion capture data into the space of the databank. (a) Neutral motion capture data; (b) Neutral model in the databank.

4.2 Our Genetic Algorithm Approach for Finding Blending Weights

Given the scaled sparse facial motions as input (i.e. specified feature point motions), we blend the examples in the databank. In this section, we find the blending weights implicitly via Genetic Algorithms (GA). More details of the following GA technique can be found in our previous work (Zhu, July 2006).

Our GA population is initialized with the previously constructed databank. In each generation, three example faces from the population pool are selected randomly. Then we use a fitness function to evaluate randomly selected three examples according to the input (i.e. captured feature point motions). The example with the worst fitness value in this tournament is not allowed to survive in the next generation. It is removed from the population and replaced with the offspring of two winners. For producing the offspring, we adopt a crossover operator to mix the two winners. After that, the offspring is used to replace the worst example in the previous tournament. The original population now comes to the next generation. We iterate such process again and again until N generations are processed. It is our GA termination condition. N =70 is chosen here experimentally.

We keep track of the best individual over all generations in the whole GA process. When GA process meets the termination condition, the best individual face so far is returned as the final surface blending result. From Figure 6, we can see the final surface blending result.

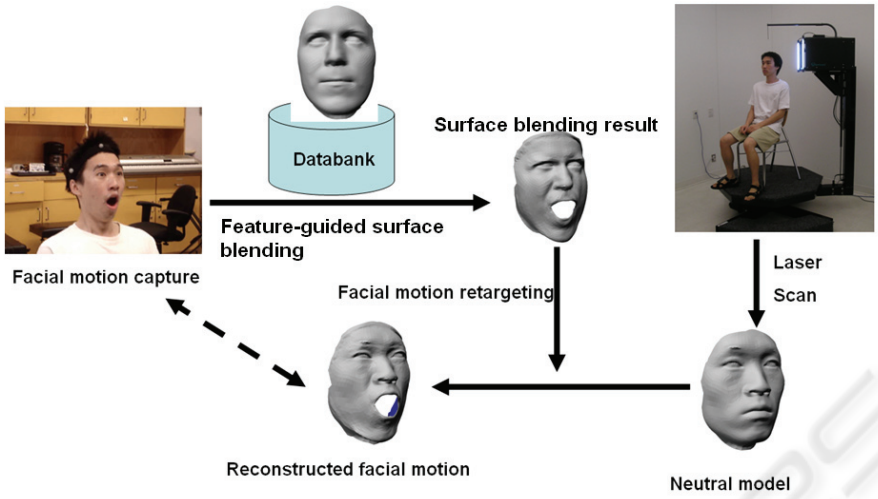

Figure 6: Illustration on how to reconstruct facial motion from sparse motion capture data.

5 FACIAL MOTION RETARGETING

Since the person in the databank is different from the performer in the motion capture session, we need to retarget facial motions to the performer's neutral face. Here we adopt the idea proposed by Noh and Neumann (Noh, 2001) where an expression cloning technique is used to clone motion vectors of the source face model onto the target model.

Before retargeting, we have to get the neutral shape of that performer's face. We utilize the *Cyberware*^{TM} *3030* laser-scanner to get the neutral shape for the performer (Figure 6). The raw laser-scanned data is processed using the consistent parameterization technique presented in Section 2. The resulting performer's neutral face model shares the same structure as the models in our databank. Different from the approach presented in (Noh, 2001) where the source and the target model have different structures, here is the trivial case where the source and the target share the same structure.

Figure 6 shows that the surface blending result is transferred to the neutral model of the performer using this facial motion retargeting technique. As shown in Figure 6, the final reconstructed facial motion is quite similar to the one in the real world.

6 RESULTS

In this section, more experimental results are shown. As described in Section 3, a *VICON*^{TM} motion capture system is used to capture sparse facial motions. Facial motions are captured when the subject is performing facial expressions and speaking. At the same time, a digital video camera is used to record facial motions of the subject.

Figure 7 shows reconstructed facial motions using the aforementioned methodology. The first, third and fifth rows are the snapshots recorded with a digital video camera. The rest rows show the corresponding reconstructed facial motions.

7 CONCLUSIONS

Our experimental results demonstrate that this methodology is effective for reconstructing realistic facial motions. By using a facial expression databank, we can achieve high quality facial animation using only feature points. The detailed methodology is: firstly, we utilize the consistent parameterization technique to process the raw laser-scanned data. Thus a facial expression databank can be constructed. Secondly, we capture sparse facial motions using an optical tracking system. Thirdly, guided by the captured feature point motions, we use a GA approach for surface blending the examples in the databank. Finally, the surface blending result is retargeted into the performer's neutral facial mesh.

However, our results depend on the similarity between the database of laser captured faces and the subject. Therefore individualized facial motion retargeting is our on-going research topic.

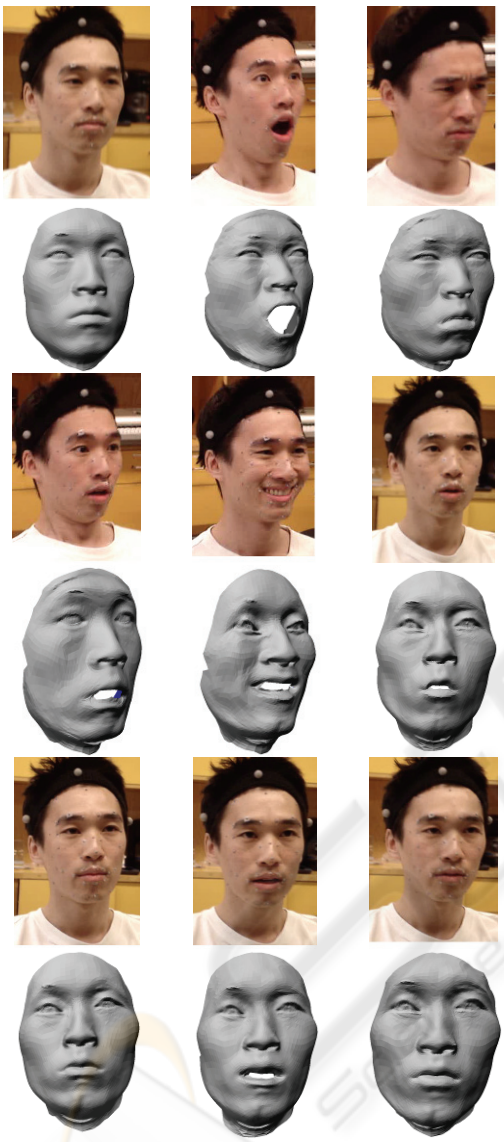

Figure 7: Results of reconstructed facial motions.

ACKNOWLEDGEMENTS

We would like to thank Ramesh Balasubramaniam in the Sensorimotor Neuroscience Laboratory of the University of Ottawa for allowing us to use his *VICON*^{TM} system. We are also grateful to Francois Malric and Andrew Soon for using their face data.

REFERENCES

Busso, C., Deng, Z., Lopes, Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S., 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. of ACM 6th International Conference on Multimodal Interfaces (ICMI 2004)*. p.205-211. State College, PA.

Chai, J., Xiao, J., Hodgins, J., 2003. Vision-based control of 3D facial animation. In *Proc. of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer animation*. San Diego, California.

Guenter, B., Grimm, C., Wood, D., Malvar, H., Pighin, F., 1998. Making faces. In *Proc. SIGGRAPH 1998.* p. 55—66.

Jeong, W.K., Kähler, K., Haber, J., Seidel, H.P., 2002. Automatic generation of subdivision surface head models from point cloud data. In *Proc. Graphics Interface*. p. 181-188.

Kähler, K., Haber, J., Yamauchi, H., Seidel, H.P., 2002. Head shop: generating animated head models with anatomical structure. In *Proc of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, San Antonio, Texas. p. 55 – 63.

Kouadio, C., Poulin, P., Lachapelle, P., 1998. Real-time facial animation based upon a bank of 3D facial expressions. In *Proc. of the Computer Animation*.

Kshirsagar, S., Molet, T., Magnenat-Thalmann, N., 2001. Principal components of expressive speech animation. In *Proc. of Computer Graphics International 2001*. p. 38-44.

Lee, W.-S., Soon, A., 2006. Facial shape and 3D skin. *Computer Animation and Virtual Worlds (SCI)*, *Vol. 17, Issue 3-4* (p. 501-512). July 2006, John Wiley.

Pighin, F., Szeliski, R., Salesin, D.H., 1999. Resynthesizing facial animation through 3D model-based tracking. In *Proc. of International Conference on Computer Vision*. p.143--150. Corfu, Greece.

Terzopoulos, D., Waters, K., 1993. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Pattern Analysis and Machine Intelligence.* 15(6):569—579.

Noh, J.Y., Neumann, U., 2001. Expression cloning. In *Proc. of SIGGRAPH 2001*. ACM Press. p. 277-288.

Williams, L., 1990. Performance-driven facial animation. In *Proc. of SIGGRAPH 1990.* p. 235--242.

Zhang, Y., Sim, T., Tan, C. L., 2004. Adaptation-based individualized face modeling for animation using displacement map. In *Proc. of Computer Graphics International 2004 (Crete, Greece)*. p. 518-521.

Zhu, L., Lee, W.-S., May 2006. Modeling and animating for the dense laser-scanned face in the low resolution level. In *Proc. of the 17th IASTED International Conference on Modelling & Simulation 2006.* Montreal, Quebec, Canada.

Zhu, L., Lee, W.-S., July 2006. Facial expression via genetic algorithms, In *Proc .of the 19th Annual Conference on Computer Animation and Social Agents (CASA) 2006.* Geneva, Switzerland.