Combining 3D Scans and Motion Capture for Realistic Facial Animation

Martin Breidt, Christian Wallraven, Douglas W. Cunningham, Heinrich H. Buelthoff[†]

Max Planck Institute for Biological Cybernetics, Tuebingen, Germany

Abstract

We present ongoing work on the development of new methods for highly realistic facial animation. One of the main contributions is the use of real-world, high-precision data for both the timing of the animation and the deformation of the face geometry. For animation, a set of morph shapes acquired through a 3D scanner is linearly morphed according to timing extracted from point tracking data recorded with an optical motion capture system.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Animation

1. Introduction

One of the applications of realistic facial animation outside the film industry is psychophysical research on human perception of faces (e.g. [7]). For psychophysical experiments, it is desirable to have both realistic and controllable facial motion stimuli. For such scientific purposes, it is important to use real-world data in order to be independent of an artist's interpretation. Through the combination of high-resolution 3D scans and 3D motion capture, we have developed a pipeline satisfying these two needs and provide a prototypical example here.

State-of-the art 3D scanning systems deliver very high spatial resolution (commercial systems with resolutions up to 50 microns exist), but usually are too slow for real-time recording. Motion capture (mocap) systems on the other hand have fairly high temporal resolution (up to 1000 Hz) for a small set of tracking points (around 50–200).

The idea presented here is to combine these two systems in order to get high resolution data in both domains that is closely based upon real-world properties. While this is similar to previous work (e.g. [1 2 3 5 6 8 9 10]) the innovation of our approach lies in the combination of precision 3D geometry, high resolution motion tracking and photo-realistic textures.

2. Overview

The described animation system consists of two parts: The generation of facial geometry (upper branch of figure 1) and the calculation of timing and amplitude of the animation (lower branch of figure 1). These two data sets will then be combined in a conventional 3D morphing process: a weighted linear combination of geometric 3D shapes will be calculated according to a set of shape weight parameters (morph channels) extracted from the motion capture recordings. The first part of this paper will describe the geometry process, the second part will deal with the processing of motion data.

3. Morph shapes

As we want to represent facial deformation as a linear combination of basic facial action units, the first step is the creation of these action units as morph shapes.

For this, the geometry of facial poses is captured using a structured light scanner manufactured by ABW GmbH. The scan takes about two seconds allowing to capture facial expression without having to hold the pose for a long time. This results in about three million vertices for a single face scan (see figure 2 for an unprocessed scan; missing data are mostly due to occlusion or limited image contrast of the structured light projection).

After manual cleanup (spike removal, cropping of unwanted geometry, filling of holes), individual scans are all

[†] e-mail: firstname.lastname@tuebingen.mpg.de

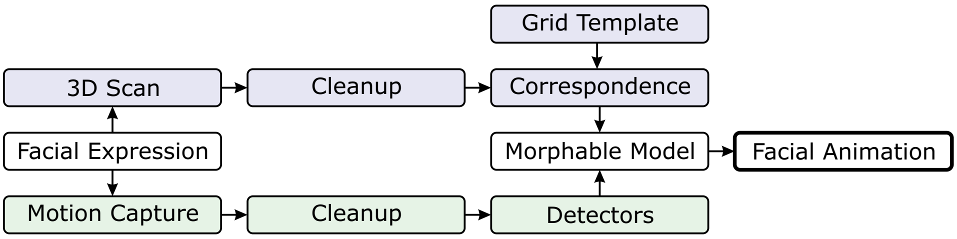

Figure 1: *The proposed facial animation pipeline.*

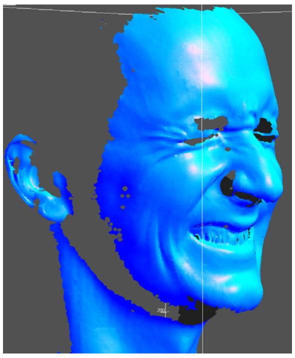

Figure 2: *Screenshot of an unprocessed 3D scan.*

put into correspondence in order to create a set of morphable polygon meshes. This is currently achieved by manually aligning a control grid of 238 points from a generic face template to each scan, taking about 3–4 hours per shape. The grid has been designed for animation purposes, providing more detail in the mouth and eye regions. Also, edge flow has been manually edited for optimal deformation by parallel or perpendicular alignment of polygonal edges to major deformation directions in the face. From this control grid, polygonal meshes of arbitrary resolution can be generated using subdivision schemes. The faces in figure 3 each consist of 11,656 vertices. This resolution is needed for real-time playback of the facial animation of an avatar in the IST research project COMIC[†].

Since the basic morph shapes will be used to produce all facial motion later on, the choice of this set is crucial for the deformation capabilities of the animatable face. As we want to capture a real face performing the individual actions for the basic morph shapes, we need to use a morph basis that can be both verbally described and practically performed by an actor. This also has the advantage that the basic morph shapes have a well-defined semantic description. It greatly facilitates any artistic manipulation of the facial animation and allows the manual creation of morph shapes for non-existing (and possibly even non-human) faces. Possible basis sets are currently being evaluated.

For this prototype, only a small set of facial expressions have been scanned but a large number of facial action units for a wide range of facial movements will be recorded in the future. We expect that 50–60 different poses will be required for most conversational expressions.

Since the scan system currently only captures black and white textures, we have fitted a separate photograph of the same neutral face to the neutral morph shape as a texture map for better visual fidelity. This texture map is deformed during morphing and partially suffers from stretching due to limited texture resolution and the frontal projection. In the final system, high resolution color pictures of the actual expression taken during the scanning process will be used to further increase the realism by providing details in texture that geometry cannot reproduce.

4. Motion Capture

Having generated a set of morphable face geometries, we capture motion data with an optical Vicon system using six cameras running at 120 Hz (figure 4). For this, 68 reflective markers are applied to the face of the same person that was previously scanned (figure 5). Five additional markers are used outside of the face for capturing the rigid head motion.

The facial performance for the animation sequence is then recorded. Since the mocap data are only used for timing and qualitative analysis, it is *not necessary* to have a precise geometric match between the scanned face and the mocap face. In principle, it is not even necessary to have the same person for scan and motion capture. Because we use the motion data at a higher abstraction level, we avoid the usual retargeting problems that arise when dealing with pure XYZ coordinate data.

Once the motion data are recorded, semi-automatic clean-up due to recording errors is required. Typical errors are missing data for certain markers over a short period of time, wrong labeling of markers, or short periods of high noise due to poor visibility of the markers in respect to the cameras.

[†] http://www.hcrc.ed.ac.uk/comic/

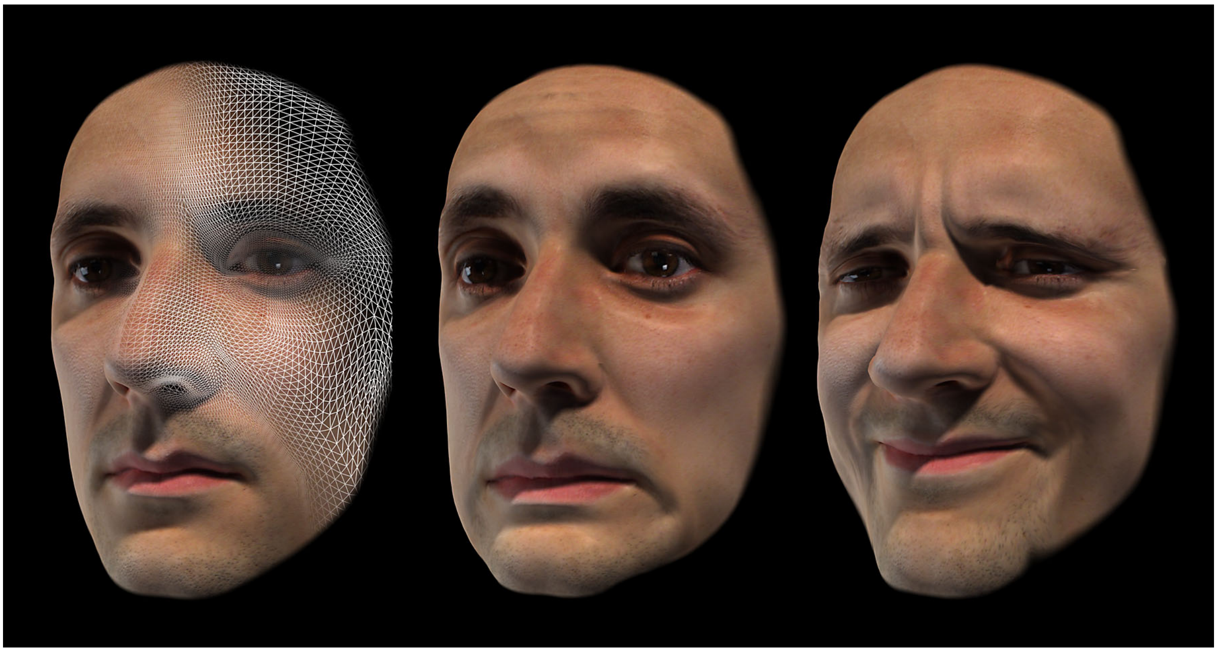

Figure 3: *Rendered 3D morph shapes.*

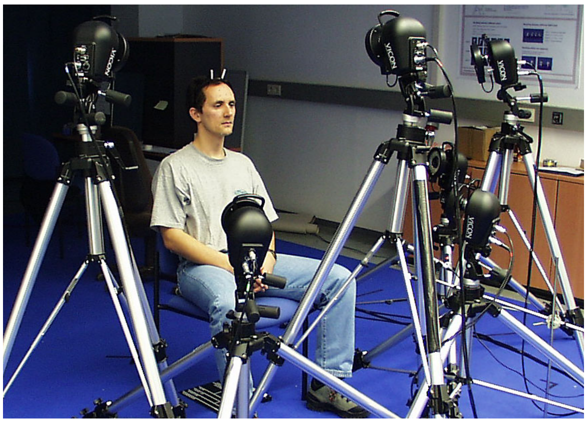

Figure 4: *Camera setup for motion capture.*

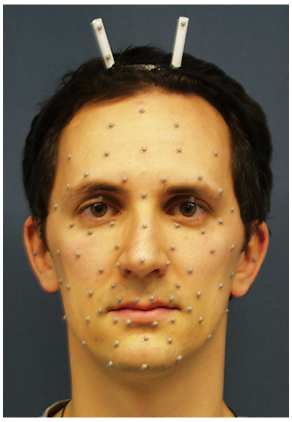

Figure 5: *Reflective marker placement for motion capture.*

For motion analysis, rigid head movement is temporarily removed in order to isolate the non-rigid facial movement. Changes in distances between markers are used to translate the XYZ data into morph animation channels. Each of these morph channels carries the contribution of a certain morph shape over time. So a morph channel contains the temporal information, but not the spatial distribution of facial features. Currently, simple linear detectors for facial action elements use euclidean distances between markers for generating the morph channels. This produces morph animation based on the amplitude and timing of marker motion in the mocap data. Figure 6 shows a plot of an extracted morph channel animation. The amplitude has been normalized to stay between 0 and 100 percent and describes the morph percentage of a 'scrunched' morph shape as shown on the right of figure 3.

In order to be able to extract a large number of morph channels, we are currently developing machine learning techniques for transcribing the mocap data into morph channel animation. Other research (e.g. [3] [8]) has used optimization techniques for recovering morph channel animation. This is not practical in our system, since the morph basis and the

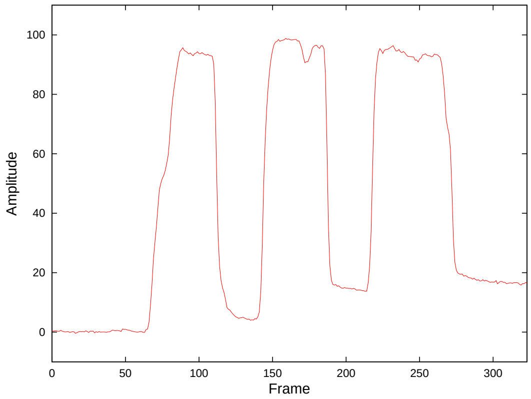

Figure 6: *Sample morph channel animation derived from tracking data.*

captured facial motion might not be identical or even from the same person, thus the different proportions of the two faces might interfere with a least-square optimization approach.

Since mocap data are not directly used for facial animation but for deriving morph channels instead (which contain essentially only timing, not a spatial description), it is easy to transfer the motion to other face models with the same set of morph shapes while retaining the temporal quality of the original recordings.

5. Conclusion

As our animation example[‡] shows, it is beneficial to use the best of both worlds: high definition scans for capturing the surface deformation and mocap data for amplitude, timing and coordination of the motion elements.

In order to make the described system practical for large-scale psychophysical research, we need to automate the process of correspondence calculation between facial expressions. A full set of individual muscle movements (for example based on Ekman's FACS[4]) will be scanned and put into correspondence.

The authors would like to thank Barbara Knappmeyer (Max Planck Institute for Biological Cybernetics), Winfried Ilg (Laboratory for Action Representation and Learning), the Department of Cognitive Neurology (University of Tübingen), and Kai Wolf (ABW GmbH) for their support in this project.

References

1. Choe, B., and Ko, H.-S. Analysis and synthesis of facial expressions with hand-generated muscle actuation basis. In *Proceedings of Computer Animation* (2001), IEEE, pp. 12–19.

2. Choe, B., Lee, H., and Ko, H.-S. Performance-driven muscle-based facial animation. *J. Visual. Comput. Animat. 12* (2001), 67–79.

3. Chuang, E., and Bregler, C. Performance driven facial animation using blendshape interpolation. Tech. Rep. CS-TR-2002-02, Stanford University, 2002.

4. Ekman, P., and Friesen, W. V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologist Press, Palo Alto, CA, 1978.

5. Guenter, B., Grimm, C., Wood, D., Malvar, H., and Pighin, F. Making faces. In *Proceedings of SIGGRAPH* (1998), ACM SIGGRAPH, ACM Press, NY, pp. 55–66.

6. Kalberer, G. A., and Gool, L. V. Realistic face animation for speech. *J. Visual. Comput. Animat. 13* (2002), 97–106.

7. Knappmeyer, B., Thornton, I. M., and Bülthoff, H. H. Interactions between facial form and facial motion during the processing of identity. Tech. Rep. 94, Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany, Nov 2002.

8. Kouadio, C., Poulin, P., and Lachapelle, P. Real-time facial animation based upon a bank of 3d facial expressions. In *Proc. Computer Animation* (1998).

9. Noh, J., Fidaleo, D., and Neumann, U. Gesture driven facial animation. Tech. Rep. 02-761, University of Southern California, 2002.

10. Pighin, F., Szeliski, R., and Salesin, D. H. Modelling and animating realistic faces from images. *J. of Computer Vision 50*, 2 (2002), 143–169.

[‡] http://www.kyb.mpg.de/~mbreidt/eg2003.html