Semantic 3D Motion Retargeting for Facial Animation

Cristóbal Curio* Martin Breidt* Mario Kleiner* Quoc C. Vuong* Martin A. Giese† Heinrich H. Bülthoff*
*Max Planck Institute for Biological Cybernetics, Tübingen, Germany
†Laboratory for Action Representation and Learning
Department of Cognitive Neurology
University Clinic Tübingen, Germany

Figure 1: Three examples of facial expressions retargeted from Motion Capture onto a morphable 3D face model.

Abstract

We present a system for realistic facial animation that decomposes facial motion capture data into semantically meaningful motion channels based on the Facial Action Coding System. A captured performance is retargeted onto a morphable 3D face model based on a semantic correspondence between motion capture and 3D scan data. The resulting facial animation reveals a high level of realism by combining the high spatial resolution of a 3D scanner with the high temporal accuracy of motion capture data that accounts for subtle facial movements with sparse measurements.

Such an animation system allows us to systematically investigate human perception of moving faces. It offers control over many aspects of the appearance of a dynamic face, while utilizing as much measured data as possible to avoid artistic biases. Using our animation system, we report results of an experiment that investigates the perceived naturalness of facial motion in a preference task. For expressions with small amounts of head motion, we find a benefit for our part-based generative animation system over an example-based approach that deforms the whole face at once.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation; H.1.2 [Models And Principles]: User/Machine Systems—Human information processing

Keywords: facial motion, facial animation, performance capture, human perception, motion retargeting

*{firstname.lastname}@tuebingen.mpg.de
†martin.giese@uni-tuebingen.de

APGV 2006, Boston, Massachusetts, July 28–29, 2006.
© 2006 ACM 1-59593-429-4/06/0007 $5.00

1 Introduction

Psychology in general and Psychophysics in particular have successfully researched the human perception of faces using synthetic images in the past decade (e.g. [Troje and Bülthoff 1996]). With recent advances in computing power, data acquisition and Computer Graphics, it has become feasible to produce stimuli of high physical realism for this purpose. At the same time, the Computer Graphics community is showing increased interest in systematically understanding human perception for achieving the desired result in the observer without spending unnecessary computational and artistic effort.

The human face is a challenging object for both fields of research. In psychology, there is demand for realistic, but controllable face stimuli. On the other hand, a good understanding of the cognitive processes of face perception in humans would clearly help Computer Graphics researchers and artists in the difficult task of synthesizing realistic virtual humans. From this, a variety of industrial applications could benefit: Computer games, human-computer interfaces, teleconferencing, medical rehabilitation systems, computer-based training and consulting as well as the film industry.

In order to achieve the level of realism required for psychophysical experiments, we use real-world data extensively. Shape and color information of the face and its deformation states are measured in a 3D scanner and then converted into a morphable 3D face model. Additionally, motion information for a sparse set of facial markers is acquired using an optical Motion Capture system. The morph activation time courses for the 3D model are computed from motion capture data by decomposing the marker trajectories into semantically meaningful motion elements based on the Facial Action Coding System (FACS) [Ekman and Friesen 1978], which defines a set of basic facial motions called Action Units (AUs). These AUs approximately correspond to natural muscle activations, providing an intuitive and accurate system for annotating facial motion. Using FACS as a basis has two additional advantages. Its semantics allow for easy retargeting of the motion onto any face model that uses the same semantic structure. In contrast to approaches that use statistical concepts such as Principle Component Analysis, Action Units can be verbally described. Thus, matching facial expressions can be

generated by actors or artists. Furthermore, Action Units describe local effects in the face which is beneficial for a generative model of facial motion. It has been argued that Action Unit activations might fail to describe the facial state accurately since they might reflect the combined activations of multiple muscles and do not take temporal information into account [Essa and Pentland 1994]. So far we have neglected temporal information in our system and have worked on a frame-by-frame basis.

After reviewing related work in Section 1.1, we give an overview over our animation system in Section 2, explaining the data acquisition process, the construction of the animation model and the motion analysis. A comparative perceptual experiment of our animation setup is described in Section 3. Section 4 concludes this paper with a general summary and outlook.

1.1 Related Work

Realistic Facial Animation Starting with the pioneering work of [Parke 1972], rendering lifelike performers has been of great interest in Computer Graphics. With [Williams 1990] an approach for the construction of realistic human head model based on photographic texture mapping has been suggested. The results initiated the development of a range of systems that analyze the expressions of a human performer and transfer the animation onto models, e.g. [Guenter et al. 1998]. [Pighin et al. 2002] build a morphable model by fitting a generic face model to multiple photographs of one facial expression using manually placed feature points; that model is afterwards used to track video sequences of the same person. [Eisert and Girod 1998] present an animation system designed for teleconferencing based on landmark tracking in videos. This process is aided by a deformable 3D mesh model that has been obtained from a 3D laser scan.

As an alternative to motion tracking, [Joshi et al. 2003] present a learning approach for the estimation of generative 3D morph models based on a combination of a reference 3D head model and a RBF deformation approach driven by sparse Motion Capture data projected onto the reference model. Similar to that idea, a muscle-based head animation system has been demonstrated by [Sifakis et al. 2005] that uses motion capture data in a non-linear optimization process to estimate facial muscle activation parameters.

For better generalization across identities, [Blanz and Vetter 1999] used a statistical approach for which a large corpus of colored 3D scanned head models in neutral pose were obtained and automatically put into correspondence. This 3D morphable model was augmented with scans of facial expressions and was subsequently used to track and alter facial motion in video sequences [Blanz et al. 2003]. [Vlasic et al. 2005] presented a video-driven motion retargeting approach for animation that controls animation factors such as the identity, type of expression and visemes of animated head models. These 3D head models are determined by a multi-linear model estimation based on tensor algebra and are directly estimated from high-temporal resolution 3D scans that are in dense correspondence.

[Kalberer and Gool 2002] learn a morphable model directly from high temporal resolution 3D scans. Their approach is based on unsupervised learning. Initially, the data acquisition process uses marker tracking. Later, the model is used for analyzing new marker-less scan data and applied for dense performance capture of mouth regions. [Chuang and Bregler 2005] introduced an animation blendshape technique for facial expression generation that is composed of different regionally selected submodels which offer an intuitive approach to control the animation. As head motion seems to be a key contribution for convincing facial expressions the authors learn a dependency for it from speech. Since facial Action Units are spatially localized across the face they have been used in Computer Vision to estimate facial states from video footage [Lien et al. 1998; Tian et al. 2000] with specialized classifiers. Instead of the determination of discrete class labels, we are interested in the suitability of FACS for the analysis and synthesis of continuous facial motion, exploiting the availability of Motion Capture and 3D colored scans of the same Action Units as a basis space.

3D Facial animation in perception research For the development of animation and retargeting techniques the assessment of their quality and realism is an important research problem. [Geiger et al. 2003] have evaluated the quality of 2D speech-driven face animation [Ezzat et al. 2002] by devising a "Turing test" that compares animations with real videos in a forced choice paradigm. Another study [Wallraven et al. 2005] has investigated different factors that influence quality of face animation. Factors such as texture, shape quality and simple animation techniques have been compared with performance measures in facial expression categorization tasks. [Kleiner et al. 2004] has developed a multi-camera 3D texture manipulation system to study the perception of facial expressions. This system allows to investigate the importance of facial regions in conversational expressions [Cunningham et al. 2004]. Facial animation technology is also used for investigating the influence of facial motion on identification performance [Hill and Johnston 2001; Knappmeyer et al. 2001; Knappmeyer et al. 2003].

2 Facial Animation System

The quality of facial animation depends both on the adequate representation of facial deformations and correct timing. Following the work of [Breidt et al. 2003], we developed a facial animation system using a morphable face model generated from 3D scans and animated by morph weights automatically computed from motion capture data.

FACS was originally designed for the analysis of natural facial motion by human observers. Nevertheless, it was used successfully for animation purposes (e.g. the character *Gollum*, created by Jason Schleifer and Bay Raitt for the recent movie trilogy *Lord Of The Rings*), although there is no systematic proof that it can produce the full range of facial motion. For our work, we make the assumption that all expressions in the normal range of conversational and emotional expressions are encodable by AUs and can be synthesized by their linear combination. (Note that for Gollum, a large number of corrective shapes were created to enhance the result of the different shape combinations).

The basic idea is to use the verbal descriptions of Action Units to instruct actors to perform basic facial actions. We record those actions in two modalities: in a 3D scanner and an optical Motion Capture system. Due to the semantic match between the two recording sessions, we are able to transfer new Motion Capture information from one modality to the other (see Figure 3). We decided to use the marker-based tracking of a commercial system for optimal speed and accuracy, even in areas of the face that show little contrast. From this we obtained very accurate data for the rigid motion of the head and the non-rigid motion of the face. No information on eye motion was captured. With eyes being such an important factor in the perception of faces, we decided not to display any eyes at all to avoid the presentation of incorrect and distracting information.

2.1 Recording and processing data

3D Scanning Shape and color data for the animatable face model was captured with a customized version of a commercial *ABW* scanner (Figure 2, top row). The scanner consists of two LCD lineprojectors and three CCD video cameras, covering an entire face from ear to ear. Additionally, the system was equipped with one digital *Canon EOS 10D* SLR camera, connected to a *ProFoto* flash system. After an initial calibration process, the scanner can compute 3D information using a Coded Light approach, producing a measurement for each pixel of each CCD camera. With a theoretical maximum resolution of 900,000 3D points and 6.2 million color samples, a typical face scan had about 400,000 points and took 2 seconds to record.

The face regions in the scans were manually masked out and small holes in the surface caused by poor reflection of the stripe pattern (e.g. eyes, eyebrows) were filled by linear interpolation of the surrounding values. Since one scan consists of four independent 3D data sets, a cylindrical resampling was performed for each face to produce one connected surface. Using the calibration information, the color information from the SLR camera was projected onto the surface to produce a detailed texture map. This procedure was used to scan one face actor who was able to perform 46 basic facial actions.

Motion Capture Facial motion data was acquired from a second actor with a commercial *Vicon 612* Motion Capture system with six cameras running at 120Hz arranged in a semi- circle at a distance of roughly 1.5m from the face (Figure 2, lower left). 69 reflective markers were attached to the motion performer's face, three more on a rigid head tracking target (Figure 2, lower right). The markers with a diameter of 2mm were no longer noticed by the performer after a few seconds and did not alter facial motion. From the second actor, we recorded facial actions using the same instructions as for the first actor in the 3D scanning step. He was able to perform a subset of 25 basic facial actions.

After capturing the facial actions, the reconstructed markers were labeled, occasional triangulation errors manually removed and gaps in any marker trajectory filled by cubic spline interpolation. Finally, the peak frame with maximum amplitude of the facial action was identified for each expression for later use in the analysis step described in Section 2.3.

2.2 Building the 3D face model

In order to build a morphable 3D face model from the scanned data, each scan was aligned to the scan of the neutral face using an Iterative Closest Point (ICP) algorithm with manual exclusion of strongly deformed regions to remove any existing rigid head motion that occurred in between the individual scans.

Next, the scan of the neutral face was taken into the 3D software *headus CySlice* where a low-resolution surface network with 136 control points was constructed on the scanned surface, omitting the areas of eyeballs and the interior of the mouth. For all of the remaining expressions, this network was copied onto the scanned surface and manually adjusted to correct for facial deformations. Exported in polygonal format, the network was converted into a triangle mesh with 3980 vertices that closely followed the scanned surface.

Finally, the individual facial action shapes were loaded into the animation package *Autodesk 3ds Max* and used within one single morph object. Increasing the morph weight for any action shape added increasing amounts of that action shape to the overall morph result. Animating all morph weights results in a facial animation based on the linear combination of the basic action shapes.

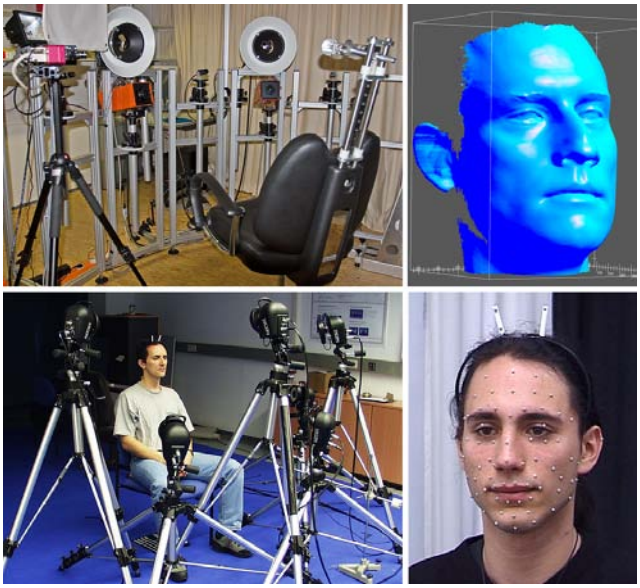

Figure 2: ABW Scanner (top left) and face scan example (top right); Motion Capture system (lower left) and reflective marker setup with additional head tracking target (lower right).

2.3 Motion analysis

In order to compute the time course of morph weights, the rigid head motion was separated from the Motion Capture data by aligning each frame of the motion sequence to the neutral start frame using an SVD alignment method [Arun et al. 1987].

For the decomposition of natural facial motion, a morph basis is required that can express all facial deformations as a linear combination of the basis elements. Analogous to the construction of the 3D shape model, we build such a basis from the peak frames of each of the basic facial expressions of the second actor, effectively creating a second morphable model for 72 marker positions that spans the same basis as used for the dense 3D shape model. This morphable model of Motion Capture markers is then used in an optimization process that estimates the contributions of the individual basis elements to a compound facial expression for each time step. Our system now finds the optimal linear combination with minimal least-squares error to the recorded marker positions. The Euclidean distance between a marker in the compound expression and the same marker in the linear combination denotes the error in the optimization problem. This error is minimized by quadratic programming enforcing positivity constraints on the morph weights as in [Choe and Ko 2001]. This decomposition of complex face configurations assumes linearity of the shape basis and does not yet exploit temporal dependencies. It produces a set of vectors of morph weights over all time steps, which is applied to the morphable 3D face model. See Figure 3 for a schematic overview of the motion retargeting process.

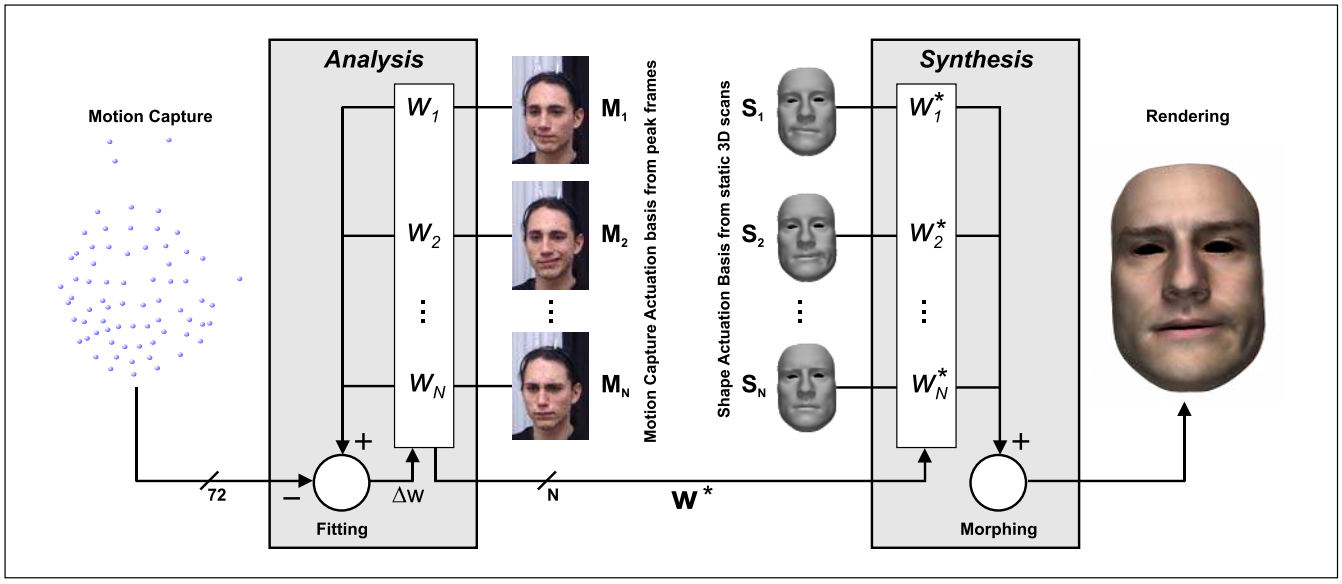

Figure 3: Schematic of our facial animation system: Using a morphable model M of motion markers (center left) that semantically matches the morphable 3D shape model S based on scans (center right), motion capture data of a complex expression (left) can be decomposed into N morph weights $\mathbf{w}^*$ and retargeted onto the 3D model (right).

2.4 Face synthesis

From the intersection of the sets of scanned basic facial actions and captured motion elements of the two respective actors, we selected a common basis of 17 basic facial actions. After the motion analysis, the morph activation values were directly used for morphing, exploiting the semantic correspondence between the basis elements. Rendering of the facial animations was done using the default rendering system of 3ds Max. It should be noted that our face model can easily be displayed and animated in real-time using current graphics hardware. See Figure 1 for some snapshots from animated expressions (from left to right: *Confusion*, *Thinking* and *Pleasant Surprise*). Demo movies are available for download at `http://www.tuebingen.mpg.de/~mbreidt/apgv06`.

3 Experiment

In order to test the quality of our animation system, we conducted an experiment on the perceived naturalness of facial motion. The main question we wanted to investigate was whether our system's ability to independently analyze and animate individual parts of a face results in an increased perceptual quality. For this, we asked subjects to directly compare two animation techniques. For technique A, we used the generative animation system as described above. For technique B we used a *global* version of the system using only two basis elements, effectively animating the entire face at once

3.1 Stimuli

We motion-captured 12 complex universal and conversational expressions (see table 1) giving verbal instructions for a *Method Acting* procedure to the same motion performer employed to construct the motion capture morph basis.

Agreement Continue	Agreement Yes	Confusion
Disagreement	Disgust	Staged Fear
Fear	Happy	Sad
Surprise	Thinking Problem	Thinking Remember

Table 1: List of complex expressions used in the experiment.

The morphable 3D model for technique A was built as described in Section 2.2. The facial animations for condition A were produced by estimating 17 morph weights for each of the recorded complex facial motions and applying them to the 3D model.

Technique A was also used to construct the two morph shapes of technique B (see Figure 4 top right): After identifying the peak frames of the 12 complex expressions, each of the peak frames were decomposed into morph weights for the morphable 3D model of technique A to produce one peak shape of model B: A static copy of the resulting face mesh was taken and stored as the compound peak shape for this expression. Since eye blinks occur often in natural facial motion and lack of them would immediately produce a very noticeable difference between technique A and B, the morph shape responsible for closing the eyes was additionally included in technique B. Effectively, we produced a simple morphable model with only two morph shapes for each complex expression. This was done analogously for the morphable marker model (Figure 4 top left).

Using the same optimization process as in technique A, activation values for the simple morph model of technique B were estimated and used as animated morph weights for the two morph channels (Figure 4 bottom). Figure 5 illustrates the resulting morph weight time courses of the two analysis techniques. The neutral shape and the eye blink shape were identical for both techniques.

The original rigid head motion present in the complex expressions was directly transferred to the morphable 3D face model for both techniques. Each complex motion started from and ended with a neutral facial expression. Differences in head orientation at the start

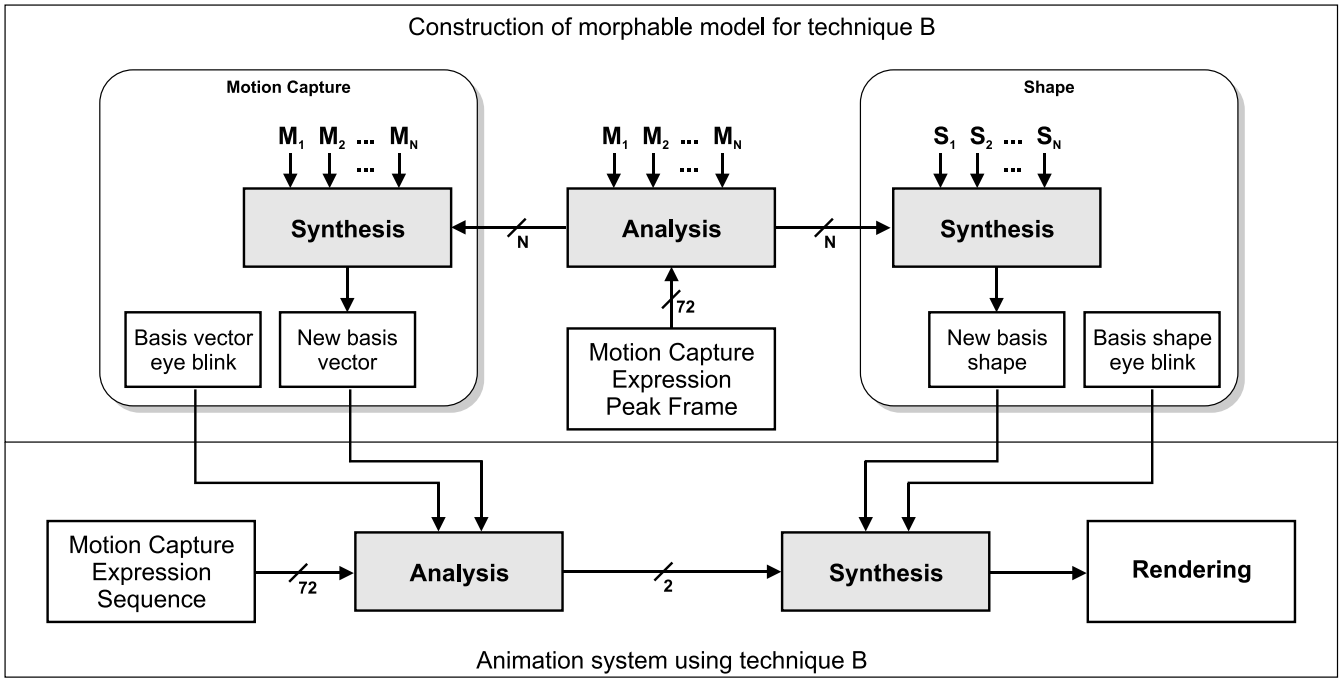

Figure 4: Construction of morphable model for technique B (top) and subsequent analysis of motion capture data (bottom).

of the motion were corrected to ensure identical start conditions for all expressions.

For a smoother appearance, we applied a single subdivision step to the face mesh before rendering the facial animation using the 3ds Max rendering system. For both techniques, a static color texture map taken from the neutral face scan was used in combination with a directional lighting setup producing relatively flat and symmetric lighting from above with soft shadows. To improve the rendering quality, a simple ambient term based on the angle of the surface normal to the observer was added to the Blinn shader of the face model. The hard edges of the face model were blended into the background using a transparency map.

From these animations, 24 Quicktime movies using H.264 compression were produced using identical perspective and lighting conditions, half of them using technique A, the other half using technique B. Each movie had a resolution of 480 x 640 pixels, with the face initially covering 50% of the image frame and never leaving it during the motion. For rendering, the temporal resolution of our animations was reduced to 30Hz. To compensate for this, we applied motion blur during rendering.

As there were varying amounts of rigid and non-rigid head motion in individual expressions, we computed the total length of the head trajectory and normalized it by the duration of the expression, effectively calculating the average speed of the rigid head motion in 3D space for each expression. Relative differences between the two non-rigid deformations were measured by comparing the displacements of all vertices of each face without rigid head motion against the neutral start of the expression. For each time step and each technique, the differences were computed and normalized by the vertex count. Then the average mesh difference between technique A and B over the duration of the expression was computed as a measure of effective 3D shape difference between the two techniques.

3.2 Experimental Design

Initial tests indicated that the two techniques produced animations that looked very similar for naive observers. Therefore, we decided to present both conditions simultaneous to allow for closer assessment of the motion details.

Our stimulus material comprised 12 pairs of animations, corresponding to the 12 different complex facial expressions recorded with Motion Capture. Each pair consisted of one movie clip showing the expression created by animation technique A and one clip showing the *same* facial expression created by technique B. A total of 19 voluntary subjects (10 males and 9 females, age between 24 and 35 years) participated in a preference study: They were seated in front of a computer monitor at a viewing distance of approximately 80 cm in a room with controlled lighting. The pairs of facial animation clips where shown to them simultaneously and side by side, using the *Psychophysics* toolbox for Matlab on Mac OS X [Brainard 1997] for movie presentation and response collection. The clips with a combined resolution of 960 x 640 pixels were played back at a frame rate of 30 frames per second on a display with a resolution of 1280 x 1024 and a refresh rate of 90 Hz. Special care was taken to prevent any jerkiness in the playback that could have disturbed the perceived facial motion. The two faces subtended approximately 17 x 10 degrees visual angle.

For each presented pair, participants had to select which of the two movies displayed a facial motion that looked *more natural* to them. It was stressed in the instructions to only focus on the facial movement, not on size, color or proportion of the faces. Animations were repeated until the participant responded with a key press indicating which movie clip was preferred. After the response the next pair of facial expressions was shown. One session consisted of six blocks: In each block, all of the 12 pairs of facial expressions were shown to the subject in a randomized order, resulting in a total of 72 trials. The presentation of the two animation techniques was counter-balanced for presentation on the left or the right side of the

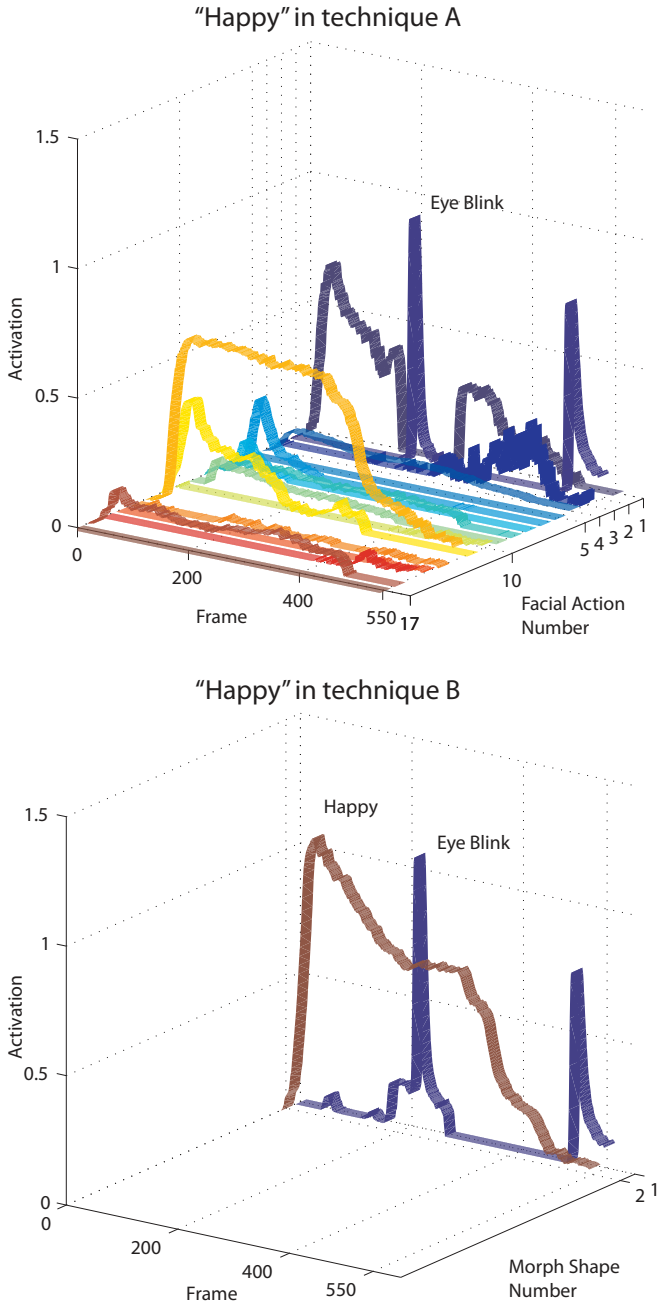

Figure 5: Morph weights for expression 'Happy' computed by technique A using 17 basis elements (top) and technique B using the two basis elements 'eye blink' and 'happy' (bottom). Note the very similar activation curves for eye blinking (dark blue ribbon #1, see color plates)

screen for each expression, in order to rule out a possible response bias for one side. At the end of a session, participants were asked for a self-assessment of their experience with 3D computer games, computer graphics, facial animation and 3D animated movies. The interviewer rated their response on a five point scale where a value of one meant "no experience at all" and a value of five corresponded to "being an expert".

3.3 Results

From the 72 trials of each subject, we computed an average preference score for technique A over technique B, separately for each expression. Figure 6 shows the mean preference and standard error of the mean.

For all statistical tests, we used a critical p-value of 0.05. A one-tailed t- test of the preference scores revealed that for half of the presented expressions the participants significantly preferred animations with technique A, as indicated by the light colored bars in Figure 6. Additionally, a repeated-measures ANOVA revealed a significant influence of the expression type ($F(11, 198)=2.4$). We also correlated participantsṕreference for technique A with computational measures of our animations and participantsśubjective ratings of their experience level. The correlation between duration of the animation and preference was significant ($r=0.782$, $n=12$).

Also, a significant negative correlation was found between rigid head speed and preference ($r= -0.646$, $n=12$). The correlation between mesh difference and preference for technique A was not significant ($r=0.254$). From the subjective expertise ratings of our participants, we found significant correlations between male gender and computer games experience level ($r=0.601$, $n=19$), male gender and computer graphics experience ($r=0.651$) and interestingly a significant negative correlation between experience with animated movies and preference ($r=-0.566$). To check for a learning effect across blocks for each expression, we ran a repeated-measures ANOVA with block as factor that showed no significant linear trend in the preference scores ($F(1,11)=3.328$).

3.4 Discussion

The correlation between average speed of the head and chance-level preference values for each expression shows that large amounts of rigid head motion made it difficult for the participants to judge subtle differences in the non-rigid motion. Taking this into account, our results indicate that there is an overall advantage for our proposed animation system based on basic facial motion elements. This is particularly interesting since all but two participants spontaneously reported after the experiment that the task was very hard to do and some were not sure there was a difference at all between the different conditions.

The computed average difference of 3D mesh displacement between the two techniques is not correlated to any of our results. This shows that such a computational measure is not necessarily useful for estimating the perceptual impact of different techniques and cannot yet replace psychophysical experiments.

The negative correlation of experience with animated movies and preference scores for technique A could be explained by the fact that the noise of the Motion Capture data and the optimization process is more visible in the part-based animation technique A than in the global animation of technique B. Animation experts might be particularly sensitive to this and therefore prefer the smoother holistic animation, whereas less experienced participants did not notice the increased noise level in particular and preferred technique A for its more natural overall motion.

It should be noted that it can prove difficult for an untrained person to reliably produce the full range of AUs (e.g. not all people can raise their left and right eyebrow independently). Two independent, certified FACS experts are needed to verify the correctness of the recorded actions, which we plan to do for our data.

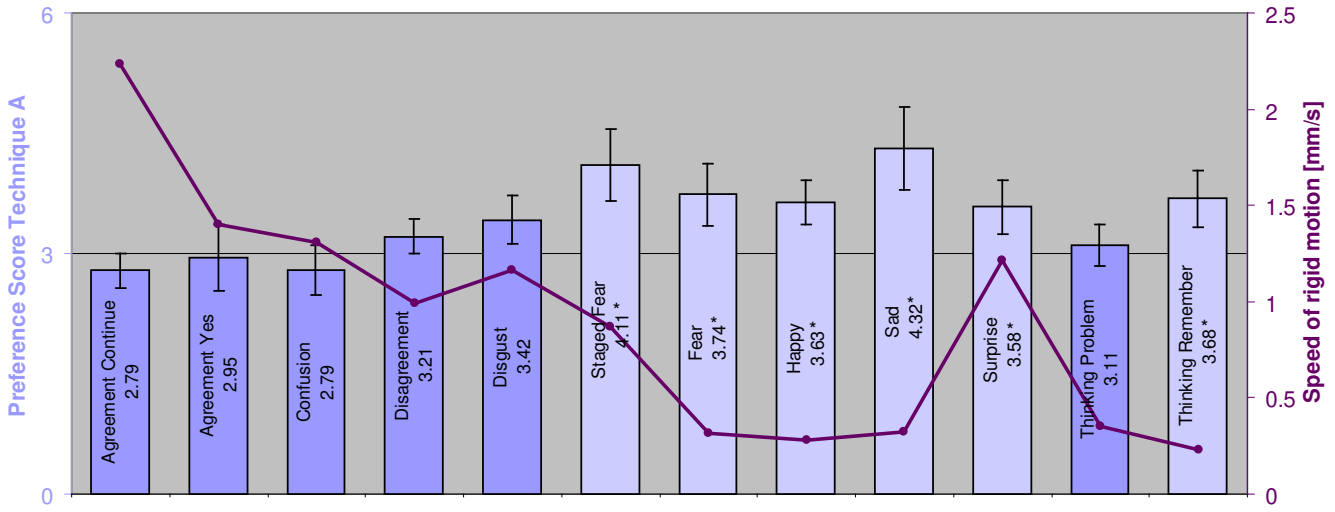

Figure 6: Preference score for technique A (bar diagram, significant values in light color, see color plates) and average head speed (line).

4 Summary

We have presented an animation system for retargeting motion-captured facial performances onto a morphable 3D face model using a semantic correspondence between the two modalities. The face model was created as a linear combination of accurate 3D scans of basic facial action units as defined by FACS. A semantically identical linear model was created from Motion Capture peak frames and used by an optimization process that fitted the model optimally into the observed marker positions recorded from complex expressions. The obtained weights of the linear combination were directly used as morph weights in the rendering system. In addition, our system could automate the very time-consuming manual FACS annotation procedure commonly used by certified experts for analyzing facial motion in Psychology.

Outlook We plan to build a richer morphable model by recording more Action Unit performers. With this augmented, manually constructed basis we intend to develop a model similar to the multilinear model of [Vlasic et al. 2005] enabling us to transfer Action Unit deformations to other identities. We expect that new Action Unit basis sets can be learned or adapted to new subjects not being able to perform Action Units. We want to evaluate our animation system with the help of certified FACS experts. Also, we plan to do motion recordings of longer sequences in natural situations. In parallel to this line of research, we are investigating real-time capabilities of the system for closed-loop conversational perception studies of facial expressions. This will include optimizing marker positions with the help of machine learning methodology. To add more realism to our animation system an eye motion model may be learned from facial expression state space models.

Acknowledgements The authors would like to thank Jens Vielhaben as face actor, Matthias Ernst as motion performer, Bernd Eberhardt for providing the Vicon system, Douglas W. Cunningham for his help with recording the data and Martin Kampe and Björn Günter for post-processing it.

This publication has been partially supported by the 6th Framework Programme of the European Commission contract number: BACS FP6-IST-027140, Action line: Cognitive Systems, and the EU IST project COMIC, IST-2002-32311.

References

ARUN, K. S., HUANG, T. S., AND BLOSTEIN, S. D. 1987. Least-squares fitting of two 3-d point sets. *IEEE Trans Pattern Anal Machine Intell 9*, 698–700.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 187–194.

BLANZ, V., BASSO, C., POGGIO, T., AND VETTER, T. 2003. Reanimating faces in images and video. *Comput. Graph. Forum 22*, 3, 641–650.

BRAINARD, D. H. 1997. The psychophysics toolbox. *Spatial Vision 10*, 433–436.

BREIDT, M., WALLRAVEN, C., CUNNINGHAM, D. W., AND BÜLTHOFF, H. H. 2003. Facial animation based on 3d scans and motion capture. In *SIGGRAPH '03 Sketches & Applications*, ACM Press, New York, NY, USA, N. Campbell, Ed.

CHOE, B., AND KO, H.-S. 2001. Analysis and synthesis of facial expressions with hand-generated muscle actuation basis. In *Proceedings of Computer Animation*, 12–19.

CHUANG, E., AND BREGLER, C. 2005. Mood swings: expressive speech animation. *ACM Trans. Graph. 24*, 2, 331–347.

CUNNINGHAM, D. W., KLEINER, M., BÜLTHOFF, H. H., AND WALLRAVEN, C. 2004. The components of conversational facial expressions. In *APGV '04: Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, ACM Press, New York, NY, USA, 143–150.

EISERT, P., AND GIROD, B. 1998. Analyzing facial expressions for virtual conferencing. *IEEE Comput. Graph. Appl. 18*, 5, 70–78.

EKMAN, P., AND FRIESEN, W. V. 1978. *Facial Action Coding System: A Technique for the Measurements of Facial Movement*. Palo Alto, CA, USA.

ESSA, I., AND PENTLAND, A. 1994. A vision system for observing and extracting facial action parameters. In *Computer Vision and Pattern Recognition Conference*, 76–83.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable videorealistic speech animation. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, ACM Press, New York, NY, USA, 388–398.

GEIGER, G., EZZAT, T., AND POGGIO, T. 2003. CBCL paper #224/ AI memo #2003-003. Tech. rep., Massachusetts Institute of Technology, Cambridge, MA, USA.

GUENTER, B., GRIMM, C., WOOD, D., MALVAR, H., AND PIGHIN, F. 1998. Making faces. In *Proceedings of SIGGRAPH 98*, Addison Wesley, M. Cohen, Ed., Annual Conference Series, Addison Wesley, 55–66.

HILL, H., AND JOHNSTON, A. 2001. Categorizing sex and identity from the biological motion of faces. *Current Biology 11*, 11, 880–885.

JOSHI, P., TIEN, W. C., DESBRUN, M., AND PIGHIN, F. 2003. Learning controls for blend shape based realistic facial animation. In *SCA '03: Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 187–192.

KALBERER, G. A., AND GOOL, L. V. 2002. Realistic face animation for speech. *Journal of Visualization and Computer Animation 13*, 97–106.

KLEINER, M., SCHWANINGER, A., CUNNINGHAM, D. W., AND KNAPPMEYER, B. 2004. Using facial texture manipulation to study facial motion perception. In *APGV '04: Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, ACM Press, New York, NY, USA, 180.

KNAPPMEYER, B., THORNTON, I., AND BÜLTHOFF, H. H. 2001. Facial motion can determine facial identity. *J. Vis. 1*, 3 (12), 338a.

KNAPPMEYER, B., GIESE, M., AND BÜLTHOFF, H. H. 2003. Spatio-temporal caricature effects for facial motion. *J. Vis. 3*, 9 (10), 304.

LIEN, J.-J. J., KANADE, T., COHN, J., AND LI, C.-C. 1998. Automated facial expression recognition based on facs action units. In *Third IEEE International Conference on Automatic Face and Gesture Recognition*, 390–395.

PARKE, F. I. 1972. *Computer generated animation of faces*. PhD thesis, University of Utah, Salt Lake City.

PIGHIN, F., SZELISKI, R., AND SALESIN, D. H. 2002. Modeling and animating realistic faces from images. *International Journal of Computer Vision 50*, 2 (November), 143–169.

SIFAKIS, E., NEVEROV, I., AND FEDKIW, R. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. *ACM Trans. Graph. 24*, 3, 417–425.

TIAN, Y.-L., KANADE, T., AND COHN, J. 2000. Recognizing lower face action units for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, 484–490.

TROJE, N. F., AND BÜLTHOFF, H. H. 1996. Face recognition under varying poses: The role of texture and shape. *Vision Research 36*, 1761–1771.

VLASIC, D., BRAND, M., PFISTER, H., AND POPOVIĆ, J. 2005. Face transfer with multilinear models. *ACM Trans. Graph. 24*, 3, 426–433.

WALLRAVEN, C., BREIDT, M., CUNNINGHAM, D. W., AND BÜLTHOFF, H. H. 2005. Psychophysical evaluation of animated facial expressions. In *APGV '05: Proceedings of the 2nd symposium on Applied perception in graphics and visualization*, ACM Press, New York, NY, USA, 17–24.

WILLIAMS, L. 1990. Performance-driven facial animation. In *SIGGRAPH '90: Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, ACM Press, New York, NY, USA, 235–242.

Semantic 3D Motion Retargeting for Facial Animation

Cristóbal Curio Martin Breidt Mario Kleiner Quoc C. Vuong Martin A. Giese Heinrich H. Bülthoff

Figure 1: Three examples of facial expressions retargeted from Motion Capture onto a morphable 3D face model (from left to right: *Confusion, Thinking, Pleasant Surprise*).

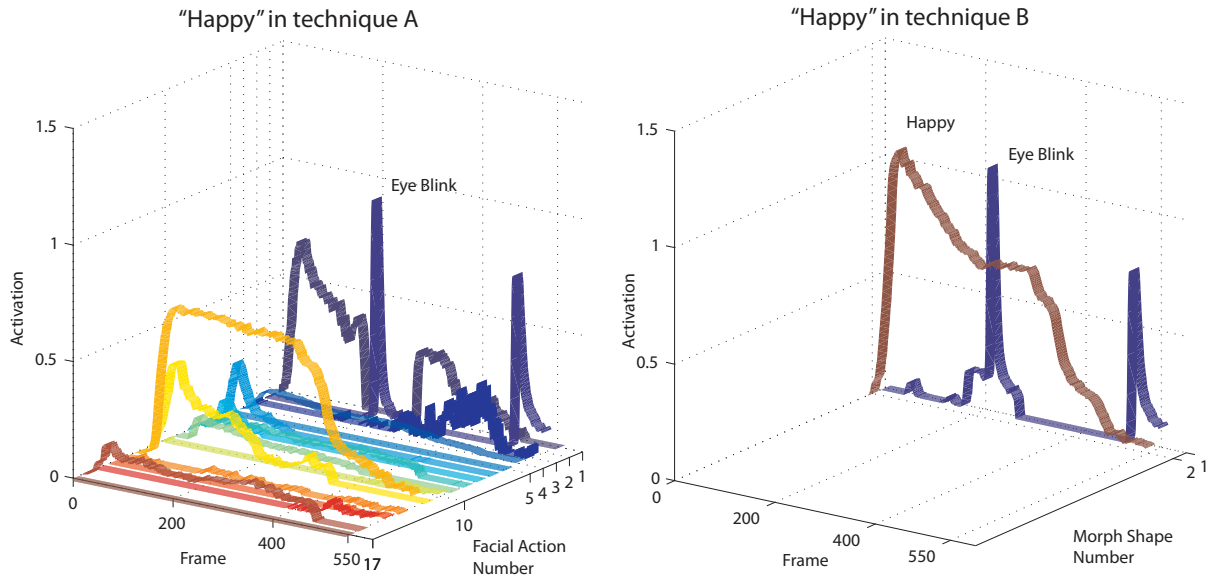

Figure 4: Morph weights for expression 'Happy' computed by technique A using 17 basis elements (left) and technique B using the two basis elements 'eyeblink' and 'happy' (right). Note the very similar activation curves for eye blinking (dark blue ribbon).

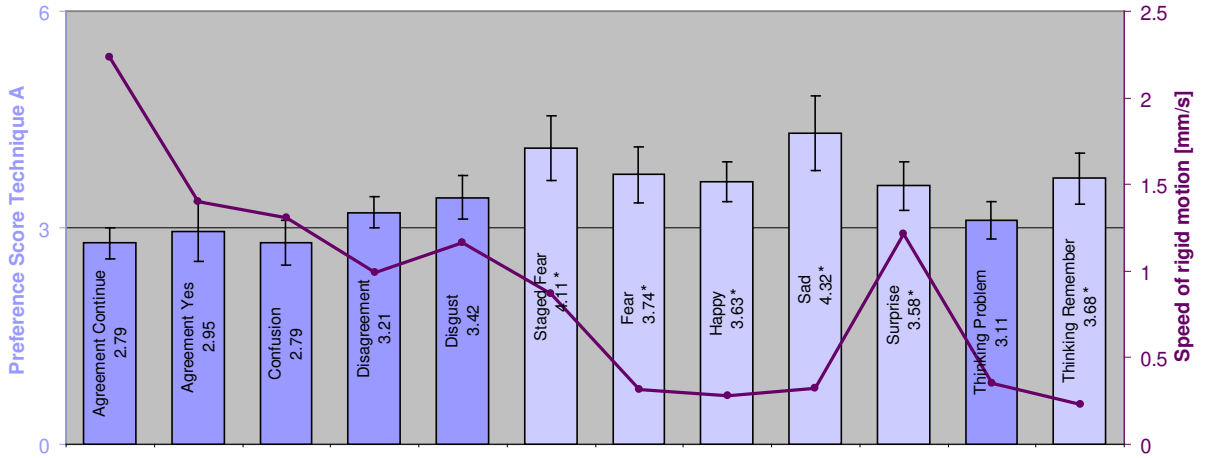

Figure 5: Preference score for technique A (bar diagram, significant values in light blue) and average head speed (purple line).