Multi-viewpoint video capture for facial perception research

Mario Kleiner Christian Wallraven Martin Breidt
Douglas W. Cunningham Heinrich H. Bülthoff

Max Planck Institute for Biological Cybernetics
firstname.lastname@tuebingen.mpg.de
http://www.kyb.mpg.de/∼kleinerm

Abstract

In order to produce realistic-looking avatars, computer graphics has traditionally relied solely on physical realism. Research on cognitive aspects of face perception, however, can provide insights into how to produce believable and recognizable faces. In this paper, we describe a method for automatically manipulating video recordings of faces. The technique involves the use of a custom-built multi-viewpoint video capture system in combination with head motion tracking and a detailed 3D head shape model. We illustrate how the technique can be employed in studies on dynamic facial expression perception by summarizing the results of two psychophysical studies which provide suggestions for creating recognizable facial expressions.

Keywords: multi-viewpoint video recording, motion tracking, texture manipulation, computer graphics, psychophysics, facial perception

1 Introduction

Communication is one of the most important tasks that humans undertake. Since facial expressions can be a very powerful form of communication, it is only natural that they should be used in applied settings, such as Human-Machine interfaces, e.g., computer animated avatars. As the synthesis of proper conversational expressions is extremely challenging [1], a systematic description of the necessary and sufficient components of conversational expressions could prove very helpful in the synthesis of conversational agents. Given that temporal information seems to be of central importance to the perception and recognition of expressions [2], such a description should include an examination of the temporal aspects of expressions. However, providing an empirical basis for such descriptions via psychophysical investigation of facial expression perception is not an easy task. It is exceedingly difficult and time-consuming to systematically alter by hand (e.g., by use of standard image processing software like Photoshop) sub-regions of a face throughout entire video sequences in order to examine the role of different types of facial motion.

Our aim here is to exemplify how multi-viewpoint video capture can be combined with computer graphics methods like tracking, texture extraction and rendering to automatically manipulate video of real expressions in a controlled way, thereby making the aforementioned investigations feasible. Towards that end, we first present our implementation and software for the creation of stimuli for face perception studies, and then we summarize two studies that have been conducted using our framework.

2 Method

2.1 The MPI VideoLab

The MPI VideoLab is a custom built, digital video- and audio recording studio that enables high quality recordings of human actions from multiple viewpoints. The system was designed to meet the following requirements:

Time-synchronized recording: All cameras should capture time-synchronized video frames and corresponding audio signals, with microsecond accurate synchronization between corresponding frames maintained over unlimited recording periods, in order to be suitable for multi-view computer vision algorithms and controlled psychophysical investigations.

Image quality: To allow accurate color based tracking and the creation of high-quality visual stimuli, the video images need to be captured in color at a high image resolution under well controlled lighting conditions. The video footage should be free of interlacing, motion blur, video compression artifacts, and image noise.

Time resolution: To reconstruct the dynamics of human facial expressions or articulated movements without motion blur, exposure times below 5 ms and frame rates higher than 25 frames/s are necessary.

Flexibility: Due to the different requirements of recording facial motion vs. body motion, flexibility in camera arrangement is needed. Also, it should be easy to upgrade the system to a higher number of cameras and audio recording channels.

2.1.1 Related work

Several other researchers have implemented multi-viewpoint video capture systems. Most of these systems either use video cameras in combination with frame grabbers, e.g., [3], or cameras connected via a firewire link, thereby avoiding the need for a frame grabber, e.g., [4]. While these systems are well suited for their respective application domains, none of them meets the specific needs of our domain of application: Many systems cannot handle real-time writeout of the data to disk storage, therefore they are limited in recording time by available memory capacity to a couple of seconds. Other systems

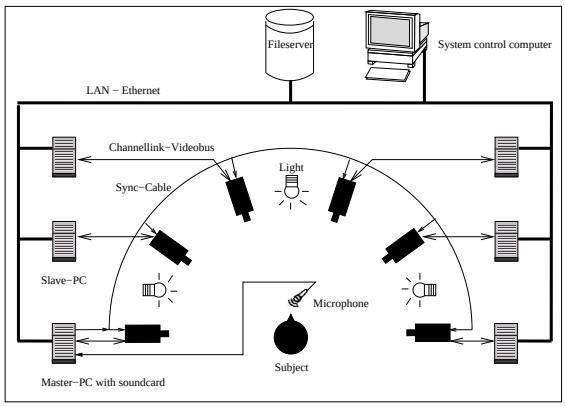

Figure 1: Schematic overview of the video setup, showing the connections between different recording nodes.

allow for unlimited recording times via harddisc recording, but are either limited in maximum recording framerate (as low as 15 fps) or in image quality due to lossy image compression to keep the data rate low.

2.1.2 System implementation

Our system is designed as an easily extendable distributed computer cluster of six video and audio recording nodes (see fig. 1), built from off-the-shelf computer components and open source software for maximum flexibility. Each node consists of a Pentium-III PC, equipped with a digital video camera, a frame grabber and - optionally - a sound card. The computers run a customized version of the GNU/Linux operating system, tuned for high disk write performance and low processing latencies. The frame grabbers are connected to each other via a dedicated cable for the transmission of a TTL trigger signal that allows for synchronisation of frame capture between the nodes with an accuracy of less than 1 μs. Using a standard LAN for exchange of control messages, our distributed control software presents the cluster as a unified system to the user and application programmer. Our cameras (Basler A302bc) are equipped with single chip, progressive scan CCD sensors and a Bayer color filter mask [5] for interpolation of 24 Bit true color RGB images from raw CCD intensity measurements. Their maximum image resolution is 782 x 582 pixels, and they allow a selection of exposure times between 10 μs and 1 s with frame rates between 3 and 60

full frames/s. Each camera is connected to a frame grabber (Silicon Software microEnable MXA36) via a ChannelLink interface for transmission of the digitized, raw 8 Bit per pixel sensor intensity measurements to the frame grabber, which in turn passes this data to its host computer. The raw data is written onto a pair of fast IDE hard disks, which are configured as a striped RAID-0 drive with a sustained write data rate of approximately 39 MB/s. The Bayer filtering step, to recreate a true color image from the stored sensor data, is performed on-the-fly by our software library, every time an application requests recorded video data. This "deferred filtering" allows us to keep the data rate for continuous video recording below 27 MB/s per camera node, while retaining full image resolution and frame rate without the use of lossy image compression. This way, we are able to perform uninterrupted recordings for several hours without loss of synchronization or dropping of frames.

For a more in-depth explanation of the system, see [6].

2.2 FaceFX

Our methodology for the automatic manipulation of facial video recordings employs a combination of 3D head motion tracking and computer graphics. The technique allows the systematic manipulation of video recordings of faces, more specifically rigid head motion and facial texture.

2.2.1 Method

Our method works as follows: First, we use a high resolution laser range scanner from Cyberware to capture the 3D geometry and facial texture of the actor's head while the actor displays a neutral facial expression. From the 3D scan, a detailed 3D morphable model of the actor's head is computed, using the methods described in [7]. The model consists of a 3D polygon mesh with approximately 150k triangles, and a texture map of the actor's facial texture with a resolution of 512 x 512 texels. The most crucial feature of the morphable model is its correspondence property: A fixed one-to-one mapping is established between specific semantic features of the face and corresponding vertex indices in the shape vector of a computed model as well as specific texture coordinates. This means, e.g., regardless of the shape of an actor's head, the vertex with index 30.000 and texture coordinates (256, 200) *always* denotes the tip of the nose. After computing the morphable model, the actor is filmed with a calibrated stereo camera pair of our video setup, while performing facial expressions. As can be seen in figure 3, the actor wears a tracking target with six green markers on his head. A standard stereo tracking algorithm detects the image positions of corresponding markers in each stereo image pair and performs stereo triangulation to recover the 3D positions of these markers. A geometric model of the tracking target is fitted to the 3D markers, thereby recovering 3D spatial position and orientation of the tracking target, and thus the rigid 3D head motion of the actor. The known camera parameters, head shape and rigid head pose for each recorded video frame allow registration of texels in the texture map of the actor's head model with corresponding pixels in each video frame by forward projection. This mapping allows the extraction of a texture map of the actor's facial texture for each video frame. For the creation of manipulated video footage, these extracted "texturemap movies" can now be automatically altered by standard image processing and video editing techniques in various ways like e.g., freezing, replacing or filtering parts of the texture, mixing texture parts from different sources or changing order and timing of "texture frames". All the manipulations to the facial regions need to be defined only once on a reference head and texture for a specific experiment and can then be applied *automatically* to different recordings of different actors, because all head models are in correspondence to each other. This saves a lot of manual setup work. The resulting manipulated texture for each video frame is then reapplied to an actor's head model and the model is rendered, either in isolation with possibly altered rigid head motion or head shape, or as an alpha-blended overlay to the original video footage, thereby manipulating the original facial texture of the video clip. The latter option preserves the original context like e.g., hair, neck, upper body and scene background. The demo videos at: *http://www.kyb.mpg.de/∼kleinerm/cap04.html*

show some examples of extracted texture movies, the quality of the tracking and correspondence as well as some typical manipulations used for conducting facial perception research.

3 Applications

3.1 Components of conversational facial expressions

In [1] we used our video manipulation technique (Sec. 2.2) to selectively "freeze" parts of an actor's or actress's face in video recordings in order to determine the necessary and sufficient facial motions for nine conversational expressions (agreement, disagreement, disgust, thinking, happy, sadness, pleasantly surprised, clueless and confused). The expressions were recorded from six different actors (3 males, 3 females, one of them was a professional actor), using an elicitation protocol based on method acting. From each of the 54 resulting video sequences, multiple versions were created: One unaltered version and five "freeze face" versions. In the frozen versions, different parts of the face were held still by replacing the corresponding facial texture of each "texture movie frame" with a still texture taken from a video frame were the actor displayed a neutral facial expression[1]. Our six versions were: Full face (original footage), rigid head motion only, head motion plus eye movements, head motion plus eye movements plus eyebrow movements, head plus eyes plus eyebrows plus mouth movement and finally rigid head motion plus mouth movement. The resulting 324 different video clips were all shown to 7 participants. Participants had to decide for each clip, which expression was shown and they had to rate on two separate seven point scales, how believable and how natural the expression looked to them. The results show that most of the tested expressions rely primarily on a single facial area to convey meaning, with different expressions using different facial areas: Agreement, disagreement and cluelessness seem to only need rigid head motion, while expressions of happiness and pleasant surprise seem to be mostly specified through mouth motion. Confusion seems to be mostly defined by eyebrow motion. Thinking relies heavily on eye motion. Only the expressions of sadness and disgust seem to require all types of motion. The results also show that the combination of rigid head, eye, eyebrow, and mouth motion is sufficient to produce versions of these expressions that are as easy to recognize as the original recordings and that our manipulation technique introduced few perceptible artifacts into the altered video sequences.

[1] *http://www.kyb.mpg.de/∼kleinerm/cap04.html#ff*

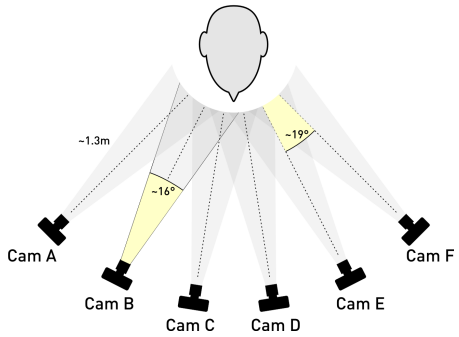

Figure 2: Camera layout used for acquisition of the Action Unit Database.

3.2 View dependency

3.2.1 Facial Action Unit Database

Using the MPI VideoLab, we recently started building a database of multi-viewpoint video sequences of facial actions. It contains a set of synchronously recorded facial movements from six different viewpoints, so far only from a single actor, but more recordings are planned.

Recordings were performed at 25 frames/sec with an exposure time of 2 ms to avoid motion blur. Four dimmable halogen studio lights with daylight color filters were used to produce uniform lighting at sufficient brightness. The camera layout is shown in figure 2.

During the recordings, the actor wore a black hat with a tracking target attached in order to be able to recover rigid head motion and thereby allow later application of our postprocessing method (Sec. 2.2). For the same reason, extrinsic and intrinsic camera parameters were measured using off-the-shelf calibration software.

Following the Facial Action Coding System (FACS) developed by Ekman/Friesen [8], the actor performed 46 actions units (AUs), each

three times, starting from and also returning to a neutral facial expression. Due to the inherent difficulty in correctly performing all individual AUs separately, not all AUs proposed by Ekman/Friesen could be recorded. Some actions were recorded both separately for left and right activation as well as in left-right combination. The recorded data were edited in order to trim unwanted pre- and post-roll footage so each sequence begins at the first indication of facial movement towards the specific AU. The peak expression of action unit 09 is shown in figure 3 (top) as an example of the action unit database.

3.2.2 View dependence of complex versus simple facial motions

In this study [9] we investigated the *viewpoint dependency* of complex facial expressions versus simple facial motions. The results not only shed light on the cognitive processes underlying the processing of complex and simple facial motion for expression recognition, but also suggest ways how one might incorporate these results into computer graphics and computer animation. For example, expression recognition might be highly viewpoint dependent making it difficult to recognize expressions from the side. As a direct consequence, modeling of expressions would then require only the frontal views to "look good", i.e., it would in principle be unnecessary to attempt detailed 3D modeling of expressions. If, however, recognition of expressions were view-invariant, then modeling would have to provide a faithful 3D rendering of facial expressions.

From the Facial Action Unit Database we extracted 14 action units, which included only internal motions of the face (no rigid head motion). Additionally, eight complex facial expressions were taken from the database (e.g., fig. 3). All sequences were recorded from four viewpoints spanning a total of 68°. Ten participants took part in the experiment, which consisted of a 22 alternative-forced choice task in which participants were instructed to view a looping video sequence and to indicate as quickly and accurately as possible which of the 14 action units or 8 facial expressions was depicted in the sequence (based on a table of names of both expressions and action units). Dependent variables

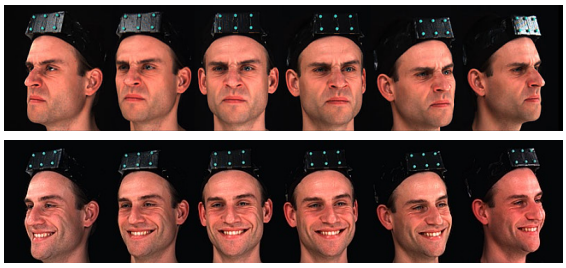

Figure 3: Six views of the peak expression from the video sequence of action unit au09 (top) and "happy" (bottom). From AU database at http://faces.kyb.mpg.de

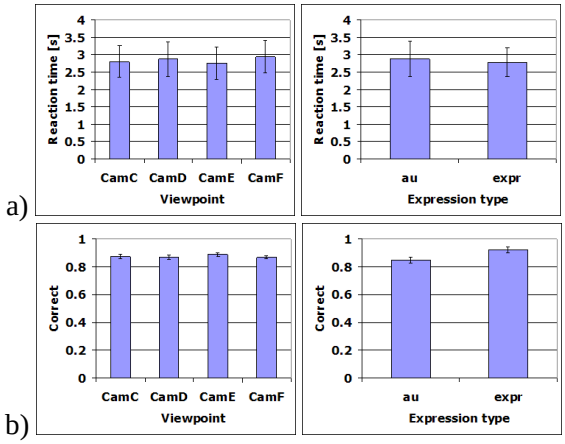

Figure 4: a) Reaction times collapsed across viewpoint and expression type; b) Correct responses collapsed across viewpoint and expression type.

in this experiment were reaction time and recognition accuracy. Statistical analysis was done using a multivariate ANOVA with factors "expression type" and "viewpoint" based on "reaction time" and "recognition accuracy". Participants had an average recognition accuracy of 88.6%, showing that the task was not too hard. Interestingly, an analysis of the confusion matrix showed that expressions were never confused with action units and vice versa, which demonstrates a clear semantic separation of simple from complex facial motions. Neither the analysis of reaction times nor of measured accuracy (see fig. 4) revealed *any effects* of either viewpoint or type of expression. One of the possible effects of view-dependent recognition could be that the recognition time varies with viewpoint: it might be more time-consuming to extract facial motion information from side views.

Recognition accuracy might also be affected by viewpoint: facial motion might be more ambiguous from the side than from the front, for example. The experimental results, however, showed no clear effects of viewpoint on either factors. It thus seems that humans are able to recognize facial motions in a largely *viewpoint invariant* manner (at least within the viewing range covered in this experiment), which supports the theoretical model of face recognition by [10]. In addition, our results suggest that in order to be recognized, computer generated facial expressions should "look good" from all viewpoints. The fact that we found no differential effects of action units and expressions sheds further light on processing strategies of expressions. First, untrained participants were able to recognize action units with a surprisingly high accuracy. Second, recognition performance of full expressions *cannot* be explained by simply adding the observed recognition performance of their constituent action units (for example, "sad" can be constructed by three simple action units). It thus seems that "the whole is more than the sum of its parts".

4 Conclusion

In this work, we illustrated by two examples, how multi-viewpoint video capture techniques can be employed together with computer graphics methods to provide a means to systematically examine the perception of facial expressions. Extending this approach in future work should not only provide fundamental insights into human perception, but also yield the basis for a systematic description of what needs to be animated in computer graphics avatars in order to produce realistic, recognizable facial expressions.

To facilitate our research on these and related projects, we plan to extend our multi-viewpoint database of facial action units (Sec. 3.2.1) and facial expressions with recordings of additional FACS performers. We also would like to validate the recorded action units in collaboration with certified FACS experts. The database will be made publicly available to the scientific community soon under the URL:
`http://faces.kyb.mpg.de`

Acknowledgements

This project was funded by the Max Planck Society. Thanks to Jens Hansen for performing the FACS Action Units.

References

[1] D.W. Cunningham, M. Kleiner, C. Wallraven, and H.H. Bülthoff. The components of conversational facial expressions. In *APGV 2004 - Symposium on Applied Perception in Graphics and Visualization*. ACM Press, 2004.

[2] J. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology*, 4, 1978.

[3] T. Kanade, H. Saito, and S. Vedula. The 3d room: Digitizing time-varying 3d events by synchronized multiple video streams. Technical Report 98-34, Robotics Institute, Carnegie Mellon University, Dec. 1998.

[4] C. Theobalt, M. Li, M. Magnor, and H. Seidel. A flexible and versatile studio for synchronized multi-view video recording. In *Proc. of Vision, Video and Graphics*, 2003.

[5] R. Ramanath, W. Snyder, G. Bilbro, and W. Sander. Demosaicking methods for bayer color arrays. *Journal of Electronic Imaging*, 11(3), 2002.

[6] M. Kleiner, C. Wallraven, and H.H. Bülthoff. The MPI VideoLab. Technical Report 123, MPI for Biological Cybernetics, May 2004. Available online at ftp://ftp.kyb.tuebingen.mpg.de/pub/mpi-memos/pdf/TR-123.pdf.

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In Alyn Rockwood, editor, *Proceedings of ACM SIGGRAPH 1999*, Computer Graphics Proceedings, Annual Conference Series. ACM, ACM Press / ACM SIGGRAPH, 1999.

[8] P. Ekman and W. Friesen. *Facial Action Coding System (FACS)*. Consulting Psychology Press, 1978.

[9] C. Wallraven, D.W. Cunningham, M. Breidt, and H.H. Bülthoff. View dependence of complex versus simple facial motions. In *APGV 2004 - Symposium on Applied Perception in Graphics and Visualization*. ACM Press, 2004.

[10] A. O'Toole, D. Roark, and H. Abdi. Recognizing moving faces: a psychological and neural synthesis. *Trends in Cognitive Sciences*, 6(6), 2003.