

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/41781844>

# Evaluating the Perceptual Realism of Animated Facial Expressions

Article in ACM Transactions on Applied Perception · January 2008

DOI: 10.1145/1278760.1278764 · Source: OAI

CITATIONS

52

READS

283

4 authors:



**Christian Wallraven**

Korea University

224 PUBLICATIONS 2,743 CITATIONS

SEE PROFILE



**Martin Breidt**

Max Planck Institute for Biological Cybernetics

43 PUBLICATIONS 475 CITATIONS

SEE PROFILE



**Douglas William Cunningham**

Brandenburg University of Technology Cottbus - Senftenberg

127 PUBLICATIONS 1,511 CITATIONS

SEE PROFILE



**Heinrich H Bülthoff**

Max Planck Institute for Biological Cybernetics

1,024 PUBLICATIONS 22,504 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



The representation of complex, navigable space [View project](#)



Human-Robot-Interaction [View project](#)

# Evaluating the Perceptual Realism of Animated Facial Expressions

Christian Wallraven, Martin Breidt, Douglas W. Cunningham, Heinrich H. Bühlhoff  
Max Planck Institute for Biological Cybernetics, Tübingen, Germany

---

The human face is capable of producing an astonishing variety of expressions—expressions for which sometimes the smallest difference changes the perceived meaning considerably. Producing realistic-looking facial animations that are able to transmit this degree of complexity continues to be a challenging research topic in computer graphics. One important question that remains to be answered is: When are facial animations good enough? Here we present an integrated framework in which psychophysical experiments are used in a first step to systematically evaluate the *perceptual* quality of several different computer-generated animations with respect to real-world video sequences. The first experiment provides an evaluation of several animation techniques, exposing specific animation parameters that are important to achieve perceptual fidelity. In a second experiment we then use these benchmarked animation techniques in the context of perceptual research in order to systematically investigate the spatio-temporal characteristics of expressions. A third and final experiment uses the quality measures that were developed in the first two experiments to examine the perceptual impact of changing facial features to improve the animation techniques. Using such an integrated approach, we are able to provide important insights into facial expressions for both the perceptual and computer graphics community.

Categories and Subject Descriptors: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation; J.4 [Computer Application]: Social and Behavioural Sciences—Psychology

General Terms: Experimentation

Additional Key Words and Phrases: Evaluation of facial animations, 3D-scanning, avatar, recognition, psychophysics, perceptually adaptive graphics

---

One of the primary goals of computer graphics is to produce realistic images. In this, computer graphics has so far largely relied on the physical modeling of object properties—examples range from the rendering equation [Kajiya 1986] based on radiative heat transfer to facial animation based on physical simulation of muscles and skin tissue [Köhler et al. 2002]. Increasingly sophisticated algorithms together with an enormous increase in computing power have enabled researchers to produce amazing images of natural and artificial scenes. A question that has emerged in recent years is, however: “How do we know when to stop?”, or “When is realism ‘realistic enough’”? In this paper, we approach the question of realism from the viewpoint of human perception: What is needed in order to produce *perceptually realistic* images? By asking this question we have thus chosen the human

---

...

Permission to make digital/hard copy of all or part of this material without fee for personal or classroom use provided that the copies are not made or distributed for profit or commercial advantage, the ACM copyright/server notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© 20YY ACM 0000-0000/20YY/0000-0001 \$5.00

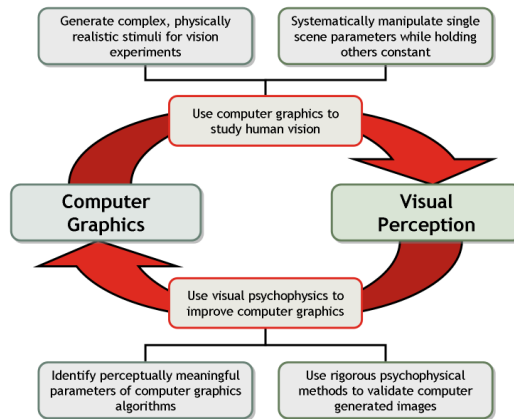


Fig. 1. Illustration of an integrated framework which establishes the connections between computer graphics and visual perception.

visual system rather than physical accuracy as our “gold standard”—after all, computer graphics produces images that humans look at and have to appreciate.

Our focus on the perceptual context of computer graphics requires a closer look at research on human perception. Although the human visual system is extremely complex, one can draw on a large body of research in the field of psychology, psychophysics, and neurophysiology. For example, in recent years, several researchers have applied knowledge from these fields to successfully design what have been dubbed “perceptually adaptive” computer graphics algorithms [O’Sullivan et al. 2004]—algorithms that take into account low-level properties of the human visual system such as its contrast or motion sensitivity (see, for example, [Myszkowski et al. 2001]). Furthermore, experimental methods from perceptual sciences allow one to both define and measure the perceptual realism of computer-generated imagery [Stokes et al. 2004]. In addition, these techniques can help to identify perceptually meaningful parameters of computer graphics algorithms, which in turn can lead to more efficient and effective rendering techniques.

This development, however, highlights only one aspect of perceptual realism (see Figure 1): Traditionally, visual perception has been studied using rather impoverished and abstract stimuli (such as simple lines, isolated objects, or line-drawings of faces). Until recently, the inability to produce visually complex scenes in a highly controllable and reproducible way has hampered the study of perception. The development of increasingly powerful graphics algorithms has enabled a systematic manipulation of complex scenes and has led to an increased use of computer-generated images and environments in perceptual research. Examples include perception of shape-from-shading [Langer and Bühlhoff 2000] and translucent materials [Fleming et al. 2004], or the systematic exploration of the perceptual space of human faces using morphable models [Leopold et al. 2001].

Taking these two developments together it seems that setting up close links between perception<sup>1</sup> research and computer graphics could provide benefits for both communities. In

<sup>1</sup>In this paper, “perception” is used in the broader context of both early and late stages of human visual processing.

this paper, we will illustrate such an integrative approach using topic that has been extensively studied in both perceptual research and computer graphics: *human facial expressions*.

Using psychophysical experiments, we will first address the question of *perceptual realism* by asking how well computer-animated facial expressions can be recognized and by showing how different animation techniques affect our qualitative judgment of these expressions. The experimental methodology has been applied previously to investigate the recognizability and believability of facial expressions in real video sequences [Cunningham et al. 2004], which enables us to directly compare our experimental results from various computer animations against the “ground-truth” of video sequences. In addition to benchmarking the recognizability and believability of animations, our results also provide some insights into which *animation parameters* affect the chosen perceptual measures. Computer animations offer complete, fine-grained control over both spatial and temporal stimulus aspects, including, for example, the timing of facial motion, the complexity of shape information, or the importance of facial components for recognition. With a benchmarked computer animation, it thus becomes possible to examine the perceptual impact of these stimulus aspects in a well-defined experimental setting. In a second experiment, we will demonstrate how the use of such computer-generated facial animations opens new possibilities in the systematic investigation of the perceptual processing of facial expressions. Finally, using the previous two experiments as baseline and motivation, we will report results from a third experiment in which we evaluated how the perceptual quality of one of the animation techniques could be improved.

The four main aims of this paper are then:

- to provide an experimental framework for evaluating facial expressions
- to present detailed evaluation results, using this framework, of several computer-generated, three-dimensional (3D) facial animation techniques
- to use the framework in a systematic investigation of important information channels in the visual processing of facial expressions
- to provide a demonstration of how the experimental results can be used to improve the perceptual quality of animation techniques

## 1. BACKGROUND AND STATE-OF-THE-ART

### 1.1 Perception of Facial Expressions

Facial motion—in particular in the form of facial expressions—has been studied extensively in the cognitive sciences over the last few decades (for two recent reviews, see [Adolphs 2002; Calder and Young 2005]). The work by [Ekman 1972], for example, suggests that there are seven universally recognized facial expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). Beyond these basic, emotionally-driven expressions, facial expressions also serve as a powerful non-verbal communication channel by modifying the meaning of what is being said, providing additional emphasis to spoken words, or controlling the flow of a conversation (see [Bull 2001]). Recent work [Cunningham et al. 2003] investigated the recognizability as well as the intensity and believability of

several conversational facial expressions using video sequences of actors. Subsequently, in [Cunningham et al. 2004] various facial regions were selectively frozen in these sequences using image manipulation techniques. This allowed them to determine which regions of the face support the recognition of an expression. Although this study represents an example of a selective investigation of information channels, the manipulation methods used are confined to the two dimensions of the image plane and therefore do not allow full control over all aspects of facial expressions. In particular, the chosen 2D image manipulation technique cannot be used to manipulate viewpoint, illumination, facial geometry or facial motion on a fine scale. In one of the few studies that have used 3D animated faces [Knappmeyer et al. 2003], the role of facial motion in identity recognition was investigated. The animation technique, however, allowed only limited facial deformations at a coarse spatio-temporal resolution. Having access to a flexible, yet at the same time visually “realistic” 3D animation thus would enable researchers to more *systematically* investigate the complex processes underlying face perception.

## 1.2 Facial Animation and Realism

Creating believable animated faces—either manually or automatically—remains one of the most challenging tasks in computer graphics. Here, we want to highlight a few approaches that have been taken to produce *three-dimensional* face animations (for a well-known example of highly realistic two-dimensional animation see [Ezzat et al. 2002]). The earliest attempts at three-dimensional animation were based on abstract muscle parameterizations [Parke 1982; Magnenat-Thalmann et al. 1988], as well as physical modeling of facial muscles using spring-mass-systems [Terzopoulos and Waters 1990]. These were later extended to include skin deformations [Köhler et al. 2002]. With the advent of laser scanners, it became possible to capture accurate facial geometry for animation (see [Lee et al. 1995]). In an extension of the “morphable model” by [Blanz and Vetter 1999], laser scans of both the neutral expression and several visemes were put into correspondence and then directly driven from video input [Blanz et al. 2003] to produce talking faces. In addition to laser scanners, motion capture systems have recently begun to be used for facial animation as they allow recording of facial motion for a set of markers attached to the face at a high temporal resolution. This data set can, for example, be used to deform the facial geometry of a laser scan (see [Williams 1990; Guenter et al. 1998] for examples of such performance-driven animation). Direct deformation of the geometry, however, leads to the problem of defining the influence regions of each of the markers. In one of the recent approaches to tackle this problem [Breidt et al. 2003], a hybrid system was demonstrated in which simple detectors translated the motion capture into morph channel activations, which were used to drive a high-resolution morphable face with high-resolution timing.

Despite these efforts, the question of how to evaluate the quality of the resulting—seemingly highly realistic—animations remains open. One interesting approach to this issue is provided by [Cosker et al. 2004], in which the quality of their two-dimensional facial animation system was tested using a well-known *perceptual* illusion (the ‘McGurk-effect’) relating to the integration of the auditory and visual senses. Their evaluation thus relied on an indirect, perceptual measure rather than on an explicit judgment of realism. Another example of a perceptual investigation of avatars is the work by [Rizzo et al. 2001], in which recognition of 3D animated facial expressions was compared with recognition of video

sequences. On average, their avatar performed consistently worse than the corresponding video sequences; unfortunately, they used different actors for each facial expression, which makes generalizations from the experimental results difficult.

Having access to an indirect, perceptual measure is especially important in the field of human-machine communication, which requires, for example, avatars for virtual kiosks or sales tasks (see, for example, [Bickmore and Cassell 2005]). For these applications it is crucial that the facial information displayed by the avatar should not only be recognizable but also believable and sincere (see also [Cunningham et al. 2003; Cunningham et al. 2004]). Depending on the scenario, one might also wish to be able to manipulate the intensity with which the avatar performs an expression or a gesture. As humans are sensitive to a large range of such aspects, any evaluation of animations needs to include a large set of *perceptual evaluation criteria* in order to capture the complexity of what it means for a facial expression to be judged as “real”. It is within this context that we will address the question of realism and perception of facial expressions using psychophysical methods.

## 2. PSYCHOPHYSICAL EXPERIMENTS

In this paper, we want to provide an experimental framework in which perceptual characteristics of animated and real facial expressions can be addressed. For this, we employ experimental paradigms derived from psychophysics in which systematic variations along selected input dimensions are mapped onto participants’ responses. This mapping then allows us to investigate the perceptual importance and meaning of these input dimensions.

In order to approach our first goal of examining the perceptual realism of computer animations, the first experiment will evaluate several different animation techniques using various perceptual measures. These measures allow us to characterize the perception of facial expressions at different levels: At the most basic level, expressions of computer animated faces have to be recognizable (see [Cunningham et al. 2003; Cunningham et al. 2004; Rizzo et al. 2001]). We have therefore chosen recognition as one of the central perceptual tasks in the following experiments. Furthermore, computer animated expressions should be recognized as quickly as human expressions are. Any human-machine interaction will be strongly and negatively affected if it takes a person too long to decipher the computer animated expressions. At the next higher level, the intensity of the animated expression should be comparable to the intensity of the corresponding human expression. Likewise, the expressions have to look sincere—who would want to interact with an avatar which seemed to be constantly lying? Finally, it might be important for an animated expression to exhibit the same degree of typicality as the corresponding human facial expression. Here, typicality is used in the sense of frequency—how often different variants of a facial expression are encountered in the real world. By using these different measures in psychophysical experiments, it becomes possible to form a more complete picture of the perceptual processing of facial expressions. This allows us not only to evaluate the perceptual realism of computer animated faces in relation to video recordings of human expressions (see Experiment 1, section 2.2); it also enables us to exploit the flexibility of computer animations to design highly controlled experiments to investigate human perception of expressions (see Experiment 2, section 2.3).

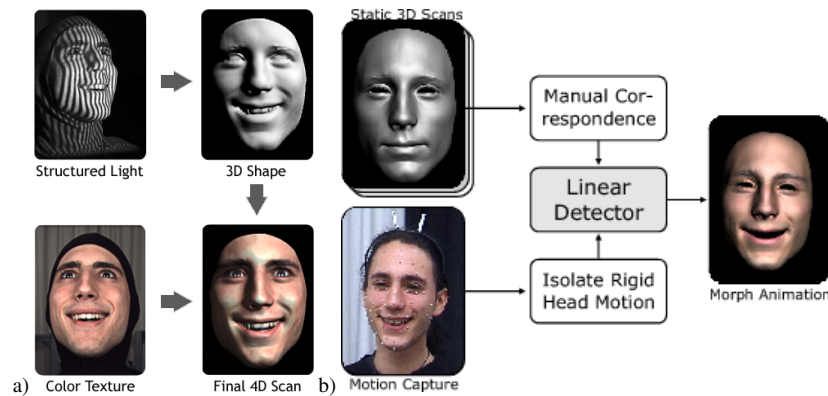


Fig. 2. Processing pipeline for a) the 4D-Scan, b) the avatar.

## 2.1 Recording and Animating Facial Expressions

In this section, we will describe the recording setups that were used to record several facial expressions as well as the system that we used to create computer animated faces. From these animated faces, we then created video sequences that were shown to participants in two psychophysical experiments.

**2.1.1 The 4D-Scanner.** In order to record human facial expressions in 3D we used one of the first commercially available 4D-Scanners (see also Figure 2a). The scanner was developed by ABW GmbH and uses a structured light projection. A 3D measurement in this setup is defined by four vertical stripe patterns, which are projected in rapid succession onto the face. The projected patterns are viewed by two high-speed digital video cameras arranged at  $22^\circ$  on either side of the projector. Image processing is then used to find the location of the edges of the stripes in each image. The four stripe patterns are needed to determine a unique correspondence for each line and to increase the accuracy of the scan. In a calibrated setup in which the spatial layout of projector and cameras is known, the deformation of each stripe pattern in the image yields information about the geometry of the object. The technical challenge for our application consisted of developing a setup that can project and record at least four stripe patterns @ 40 frames per second. In addition to the high-speed scanning the setup also contains an additional, synchronized digital color video camera to record the texture of the face. In order to ensure a more homogeneous lighting for this texture image, we also flash two stroboscopes at this time. The resulting sequences consist of 40 textured 3D shapes per second and thus capture *both shape and texture changes over time*—the only drawback of the current setup is its limited recording volume which does not yet allow for much rigid head motion without loss of coverage.

**2.1.2 The Avatar.** The avatar that was used in this paper is based on the design by [Breidt et al. 2003], in which the high *spatial* resolution of state-of-the-art 3D-scanners was combined with the high *temporal* resolution of motion capture (see Figure 2b). Facial geometry is first captured using a structured light scanner, which produces a detailed 3D-scan of each peak expression. This version of the scanner relies on the same recording technology as the dynamic scanner but provides a higher spatial resolution at the expense of temporal res-

olution (a scan takes 2s). In addition to the 3D-scan, a high-resolution texture is recorded with a digital SLR camera. The individual scans are cleaned and put into correspondence with each other in order to create a basis set of morphable scans. This correspondence is based on a control mesh whose 3D fiducial points are aligned to each scan. The second step in creating the avatar is to capture motion data with an optical motion capture system. In our setup we apply 72 markers to the face of the previously scanned person. The person is taken through the same expression elicitation protocol as was used in the 3D-scanner (see section 2.1.3 below; they are also asked to try to perform the expressions in the same way as they had in the 3D scanner). From the motion capture data we first isolate the non-rigid facial movement. In a next step, we specify simple linear detectors for facial action elements by using distances between markers to drive the morph channels. This yields an animation based on the amplitude and timing of marker motion in the motion capture data. By not using the motion capture to directly deform the facial geometry we are able to greatly increase the robustness of the resulting animation while retaining the high temporal resolution of the data. The morph animations of the final avatar<sup>2</sup> thus allow full control over important information channels used in facial expressions (such as internal motion of the face, rigid head motion, or fidelity of shape and texture) making it ideally suited for psychophysical experiments.

**2.1.3 Recording protocol.** For our experiments we recorded seven facial expressions from an amateur actor: confusion (as if the actor did not understand what was just said), disgust, fear, pleased/happy, pleasantly surprised, sad, and thinking (as if the actor was solving a math puzzle). These were elicited using a protocol based on method acting, in which a brief scenario was described in detail and the actor was asked to place himself in that situation and then react accordingly. In addition, the actor was asked to react without speaking. Note also, that this set of expressions includes five universal expressions and two expressions which usually occur in a conversational context. These particular expressions were chosen as the high-speed scanning setup currently allows only little rigid head motion without loss of coverage (see above). In addition, our choice allowed us to benchmark our results against previous experimental data that was available for *real video sequences* of the same expressions [Cunningham et al. 2004] for the *same actor*.

The recording protocol was repeated in the three different recording setups: motion capture, static 3D-scanner, as well as the 4D-Scanner. Both motion capture data and 4D-Scans of the resulting recordings were then edited so that the sequence began on the frame after the face began to move away from the neutral expression and ended after reaching the peak of the expression. Although the resulting seven video sequences varied in length across expressions (shortest 0.8s, longest 2.5s, average 1.5s), this variation was consistent across the different recording systems. The motion capture data and the static 3D-scans were then used to create animations of the avatar as described in section 2.2.1 below.

## 2.2 Experiment 1 - Evaluation of Animations

In the first psychophysical experiment, we were interested in evaluating several animation techniques with respect to their recognizability, intensity, sincerity, and typicality. With

<sup>2</sup>For a discussion of “eyes” and “teeth”, please see Experiment 3.



this experiment, we aim to establish a set of measures with which the task-dependent, perceptual quality of the animations can be determined. The different animation techniques are first treated in a “*black-box*” fashion in order to benchmark their performance with respect to real-world video sequences. In a second step, the results are discussed in more detail to examine which animation parameters might be responsible for the observed performance pattern.

**2.2.1 Methods.** In this experiment, we examined the perception of the seven recorded expressions using six different animation techniques. The first four animation techniques were derived from the full version of the avatar and were modeled after current approaches to facial animation, which can be found in the literature—it is important to stress that our main focus was not on *creating* perfectly realistic facial animations, but rather on investigating the usefulness of our evaluation framework with a variety of different animation techniques. Two additional animations were derived from the 4D-Scans, which resulted in the following six experimental conditions:

**Avatar:** The hybrid animation system as described in section 2.1.2. It combines natural timing from motion capture with high-resolution 3D-scans of neutral and peak expressions.

**AvaNoR:** In this condition, the rigid head motion, which was extracted from the motion capture data, was turned off in order to evaluate its impact on our perceptual measures. Such an animation technique could, for example, be derived from static 3D-scans together with hand-modelled timing information from video sequences.

**AvaLin:** This animation was produced by a linear morph between the 3D-scans of the neutral expression and each peak expression. Both rigid head motion as well as sequence length were the same as in the Avatar condition. This animation technique can be achieved with a combination of simple motion capture techniques to record the rigid head motion (“head tracking”) and two static 3D-scans.

**AvaClu:** This condition uses a static 3D-scan of the neutral expression, which is directly deformed (including rigid head motion) by the motion of facial markers tracked in motion capture—in animation terminology this is usually referred to as “cluster animation”. The challenge for this type of animation consists of defining the clusters, that is, finding a suitable mapping between motion of a facial marker and the surrounding 3D-geometry (the so-called “weight maps”).

**4DScan:** This condition consists of the four-dimensional captured data for the performed expressions (see section 2.1.1). Note, that this data includes changes in both geometry and texture.

**4DPeak:** Here, we use only the last (peak) frame of each expression from the 4D-Scan to evaluate the difference between perception of dynamic sequences and static images. In order to prevent timing differences between this and the previous condition, the peak frame was shown for the same duration as in the 4DScan condition.

The resulting video sequences were presented at 512x512 pixels on a 1024x768 monitor which participants viewed from a distance of approximately 0.5 meters (the face on the monitor subtended a visual angle of 11.4°). A single trial of the experiment consisted of the video sequence being shown repeatedly in the center of the screen. A 200 ms blank screen was inserted between repetitions of the video clip. When participants were ready to respond (they were instructed to answer “as soon as they knew what expression it was”

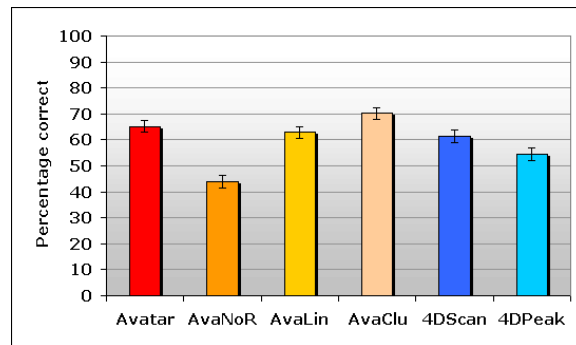


Fig. 3. Recognition results from Experiment 1 for different animation techniques (error bars represent standard error of the mean).

by pressing the space bar, which allowed us to measure their reaction time), the video sequence was removed from the screen, and the participants were to perform four tasks. The first task was to *identify* the expression by selecting the name of the expression from a list displayed on the side of the screen. The list of choices included all seven expressions as well as “none of the above” (an 8-alternative-non-forced-choice task, see [Cunningham et al. 2003] for a detailed discussion of this paradigm). The second task was to rate the *intensity* of the expressions on a scale from 1 (not intense) to 7 (very intense). In the third task, participants were to rate the *sincerity* of the expressions, with a rating of 1 indicating that the actor was clearly pretending and a value of 7 indicating that the actor really meant the underlying emotion. Finally, the participants were asked to rate the *typicality* of the expression. Specifically, participants were asked to indicate on a scale from 1 to 7 if what they just saw is something that people normally do. Participants were explicitly instructed to anchor all scales at a value of 4 (normal intensity, sincerity, and typicality) and to try and use the whole range of the scale during the experiment.

For this experiment, the crossing of 7 expressions, 6 animation techniques, and 5 repetitions of each sequence yielded a total of 210 trials.

**2.2.2 Results and Discussion.** The data from 12 participants were analyzed using standard “analysis of variance” (ANOVA) methods which yield the statistical significance of each experimental variable (expression, animation style) for the different measures (recognition, reaction time, intensity, sincerity, and typicality). In the following we will discuss the significant statistical effects for each measure.

**Recognition:** Figure 3 shows the overall recognition results broken down by animation style (ANOVA result for animation style:  $F(5,55)=3.270$ ,  $p<0.05$ ). It is immediately apparent that the recognition of expressions was far from perfect regardless of how the expressions were presented to participants (average recognition rate of 60.0%). This is consistent with earlier results by [Cunningham et al. 2004] in which recognition results for *video sequences* using the same expressions by the same actor were investigated and which will be discussed below. The worst condition was the avatar without rigid head motion (AvaNoR) with only 44.2% of the expressions correctly identified. In contrast to this, the full avatar including rigid head motion (Avatar) resulted in a recognition rate of 65.5%. The differ-

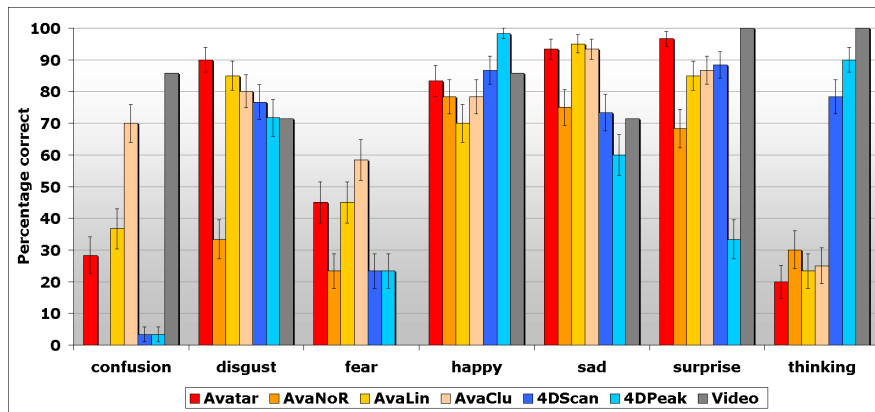


Fig. 4. Recognition results from Experiment 1 broken down by animation technique and individual expression (the bars labelled “Video” show data from [Cunningham et al. 2004]).

ence between these two condition highlights the importance of including rigid head motion in computer animated faces. Compared to the Avatar condition, the linear morph condition (AvaLin) resulted in a small but significant drop (T-test result:  $p < 0.05$ ) in performance to 62.9%. This demonstrates the perceptual difference of the timing of internal facial motion extracted from *real-world* motion capture data in contrast to simple *linear* timing. A similar effect of temporal information could also be observed in the decrease between the 4DScan (61.4%) and the static 4DPeak condition (54.3%).

In addition to recognition results for each expression, Figure 4 also shows the results from [Cunningham et al. 2004] (labeled “Video”; N.B. “fear” was not tested in their experiments). Using these values as a benchmark, we can see that the recognition rates of the Avatar were *as good or even better* for the tested expressions with the exceptions of “thinking” and “confusion”. The 4DScan condition also yielded very similar recognition rates to the video sequences except for “confusion”. This suggests that the 4D-scans capture the spatio-temporal characteristics of most expressions well enough to support a similar degree of recognizability. The differential results between the avatar and the 4D-Scanner conditions for the “thinking” expression seem to be due to the missing eye motion for the avatar (eye motion was present in the 4D-Scans). The consistent level of performance also shows that the actor was able to produce consistent expressions across the different recording setups (with the notable exception of “confusion”, which was recognized better in the real video sequences because the actor mouthed the word “What?”). Furthermore, the results also indicate that the reason for the surprisingly good recognition rates for the (simple) cluster animation are mainly due to the “confusion” and “fear” expressions. For the remaining expressions, the performance was lower or as good as for the Avatar condition. A closer look at the “confusion” and “fear” sequences in the AvaClu condition revealed jitter introduced by the noisy motion capture data which might have “helped” to identify these expressions. As mentioned above, the avatar contains a classification step that maps marker movement into activation of a peak expression results. This classification tends to smooth the motion capture data and thus might mask important high-frequency motion suggesting a further route for improvement of recognition performance for the avatar.

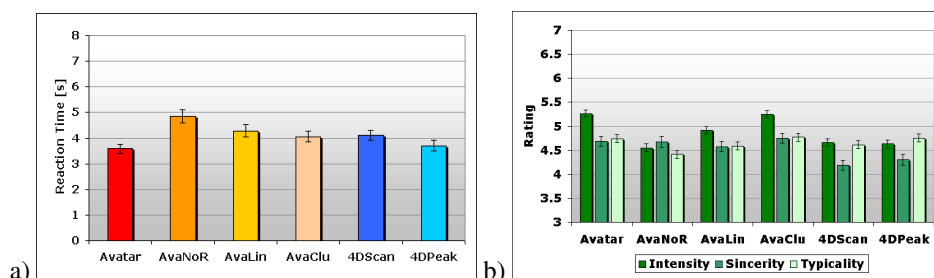


Fig. 5. Results from Experiment 1 for a) reaction times and b) intensity, sincerity and typicality

Finally, there was no straightforward correlation between sequence length and recognizability of the depicted expression showing that this factor could not account for the performance pattern found in this experiment.

**Reaction Time:** Reaction times reached stable values after two (out of five) repetitions of the sequences indicating that participants were able to perform consistently after a short time. In Figure 5a reaction times are shown for these last three repetitions broken down by animation style (ANOVA result:  $F(5,55)=2.953$ ,  $p<0.05$ ). In general, it took slightly less than 2 repetitions of the video sequence before participants answered. The overall range of reaction times corresponds closely to the results found in [Cunningham et al. 2004]. In both this and our study participants were not explicitly instructed to be prompt with their answer, which—together with the other tasks that needed to be performed on the video sequence after the recognition task—resulted in participants adopting a response strategy based on more than one viewing of the sequence. This strategy was also mentioned by several participants during the debriefing after the experiment.

Turning to the reaction time data in Figure 5a, participants responded fastest for the avatar and the static peak followed by the 4D-Scan, the cluster animation, the linear morph animation, and the AvaNoR condition. In addition to the worst recognition performance, the AvaNoR condition also had the highest reaction times. In contrast to our results on recognition accuracy, here the addition of either type of temporal information (rigid head motion and timing of internal motion) resulted in a *large* decrease in reaction times. This demonstrates that in order to judge the quality of an animation, it is important to take into account reaction times as well as recognition accuracy as they might result in different performance patterns. Finally, and not surprisingly, the static 4DPeak condition is recognized faster than the corresponding sequence (T-test result:  $p<0.05$ ) as it shows the peak frame from the start of the animation—it is interesting, however, that the avatar results in reaction times similar to the static condition.

**Intensity, Sincerity, and Typicality:** The final analysis is concerned with the three additional quality criteria that were used in the ratings of the participants (see Figure 5b). First of all, the *intensity* ratings for the AvaNoR, 4DPeak, and 4DScan conditions (ANOVA result for main effect of animation method:  $F(5,55)=9.580$ ,  $p<0.001$ ) were the same at an average value of 4.6 (all T-test results not significant, i.e.,  $p>>0.05$ ), which corresponds to slightly above normal intensity. Expressions using linear morph animations were judged

as slightly more intense, and both cluster animation and the avatar were judged to be most intense at an average value of 5.3 (all T-test results:  $p < 0.05$ ). The results thus show a clear correlation of intensity with presence of rigid head motion. Comparing the Avatar, AvaNoR and AvaLin conditions, one can see that the timing of the internal facial motion also affected intensity judgments. Both types of temporal information had an equally large impact on perceived intensity.

For the four avatar conditions, participants' judgments of *sincerity* followed the same pattern as for the intensity judgments (ANOVA result for main effect of animation method:  $F(5,55)=2.421$ ,  $p < 0.05$ ): Both the avatar and the cluster animation were rated as being most sincere followed by the linear morph and the avatar without rigid head motion (all T-test results  $p < 0.05$ ). Interestingly, both of the 4D-Scan conditions were seen as less sincere than the avatar conditions. A closer look at the data for individual expressions reveals that this was mainly due to the results in the "confusion" expression, which was rarely identified correctly either in the static or the dynamic 4D-Scan condition. For the remaining expressions, sincerity reached comparable levels to the Avatar condition; this suggests that in order to design an avatar that will be judged as sincere, it first has to be ensured that these expressions can be recognized reliably. Finally, the data for *typicality* (ANOVA result for main effect of animation method:  $F(5,55)=1.854$ ,  $p < 0.05$ ) shows that both the AvaNoR and AvaLin condition were judged as less typical than the other conditions (all T-test results  $p < 0.05$ ). This points again to the importance of rigid motion and timing of internal motion for optimal perceptual fidelity. There were no significant differences between the 4D-Scan and avatar conditions.

**Summary:** A comparison of the present recognition results with those obtained using real video sequences has shown that, for several expressions, the avatar already provides comparable recognition performance. Furthermore, we found that rigid head motion as well as natural internal motion play crucial roles across all investigated measures. Both types of temporal information ensure that the resulting facial expressions are recognized reliably and quickly as well as providing important cues for judging their intensity, sincerity, and typicality. Although the effects found for the ratings are small (which is to be expected from the 7-point Likert-scale we used), we observed a consistent pattern of changes in performance across all measures, which illustrates the robustness of the effects.

The results of Experiment 1 have important ramifications when considering the performance gain for computer-generated animations that can be achieved using relatively little additional bandwidth: adding only six floating point values per frame, which encode three translational and three rotational degrees of freedom, results in an increase of 43% in recognition accuracy and a decrease of 14% in reaction time, as well as increasing intensity by 8% and typicality by 5%. Adding another float value per frame for encoding the activation of the single morph channel results in an increase of 5% in recognition accuracy and a decrease of 19% in reaction time, as well as increasing intensity by 7% and typicality by 3%.

### 2.3 Experiment 2 - Spatio-temporal Characteristics of Expression Recognition

Using the perceptual evaluation of the avatar done in the previous experiment as a baseline, we can now systematically manipulate interesting aspects of facial expressions. Such a de-

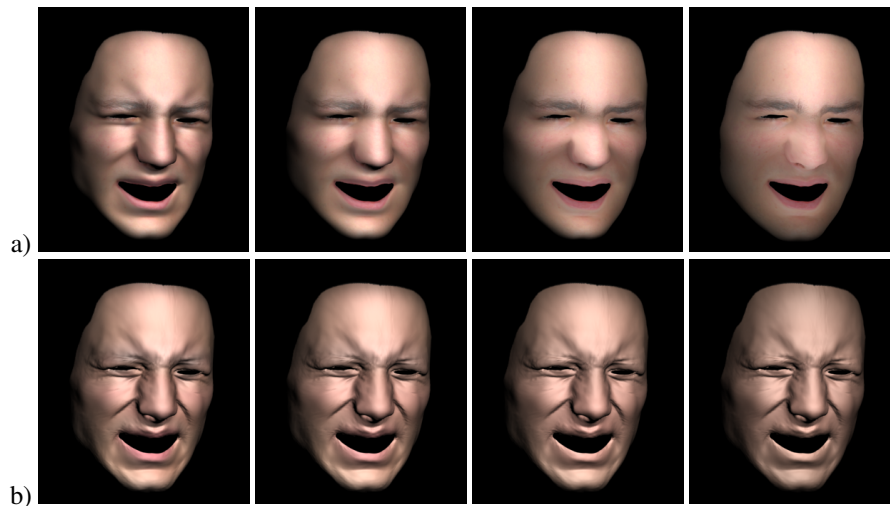


Fig. 6. Examples of the four a) shape and b) texture blur-levels for the “disgust” expression used in Experiment 2.

tailed, benchmarked 3D animation gives us full control over all relevant stimulus aspects, the manipulation of which would otherwise require extensive manual modeling. Specifically, in the following experiment we are interested in the information contained in the shape (the 3D geometry), texture (the coloring of each vertex) and motion (the temporal dimension) channels of the avatar. To date, there has been no detailed study of how these three information channels support the recognition of facial expressions. In addition to insights into spatio-temporal aspects of human visual processing of facial expressions, the investigation of these channels can also provide some additional suggestions on their importance in creating recognizable and believable avatars.

**2.3.1 Methods.** Both the setup and the experimental tasks remained the same as in Experiment 1. In this experiment, we investigated the three different information channels of shape, texture, and motion. Each channel was manipulated separately, using a Gaussian blur kernel which resulted in a gradual degradation of information across four distinct blur-levels. More specifically, the Gaussian kernel, whose width increased two-fold for each of the four blur-levels, was used to blur the texture of the avatar, the surface normals of each frame (effectively degrading shape information), or the overall motion (including both the activation of the morph channel as well as the rigid head motion). The end point of blurring was chosen so as to eliminate the information in the channel as much as possible (Figure 6 illustrates shape and texture blurring, note that texture blurring in Figure 6b results in a slight darkening of the overall texture). In order to be able to evaluate potential differential impacts on static versus dynamic sequences, the experiment used both blurred animations as well as blurred static (peak) frames. The experiment was split into three blocks, where each block manipulated a single information channel for both the static peak and the avatar conditions. The order of these blocks was randomized across participants in order to measure reaction times for each of the three blur-types independently.

The crossing of 7 expressions, 3 manipulation types (shape, texture, and motion), 2 motion types (dynamic versus peak only), and 5 blur levels yielded a total of 175 trials (NB: for motion blurring, the “peak only” condition was not run).

**2.3.2 Results and Discussion.** The resulting sequences were shown to another group of 12 participants who had not participated in the previous experiment. The results analyzed using the same statistical methods as in the previous experiment.

**Recognition:** Figure 7a shows the overall recognition results for blurring of motion, shape and texture for both the dynamic (MotionAva, ShapeAva, TexAva) as well as the static expressions (ShapePeak, TexPeak) (ANOVA result for main effect of blur-type:  $F(4,44)=8.289$ ,  $p<0.001$ ). Overall recognition rates were lower for the static peak frame than for the avatar (T-test result:  $p<0.01$ ), which provides further confirmation of the motion advantage found in Experiment 1. A closer look at shape and texture blurring in the static condition reveals that shape blurring severely affected recognition across blur-levels (T-test result:  $p<0.05$ ), whereas texture blurring did not (T-test result:  $p=0.19$ , not significant). Turning to the dynamic sequences, we see that recognition for the ShapeAva and TexAva conditions is the same as for the original avatar (Exp1Ava) (T-test results:  $p=0.29$ ,  $p=0.44$ , not significant). It seems that the additional temporal information is enough to cancel any effect of shape blurring on recognition of the animated expressions. This importance of temporal information for robust recognition is demonstrated in the MotionAva condition, in which blurring yielded significantly decreased recognition rates.

**Reaction Time:** The results in Figure 7b show that shape and texture blurring had a much larger overall effect on static expressions than on dynamic sequences. Reaction times for shape and texture blurring in the ShapeAva and TexAva conditions also were slower than in Experiment 1 (the plot shows recognition times for all 5 (Exp1Ava5) as well as for the last 3 repetitions (Exp1Ava3) of Experiment 1; T-test results:  $p<0.05$ ;  $p<0.001$ ). Degradation of motion information resulted in significantly slower reaction times dropping to the level of degraded static frames.

**Intensity, Sincerity, and Typicality:** Finally, Figures 7c and d show the experimental results for intensity and sincerity as a function of blur-level and blur-type. We found significant statistical effects of blur-level on intensity (ANOVA result for main effect of blur-level:  $F(4,44)=7.719$ ,  $p<0.001$ ) for both the shape as well as the motion channel. Similarly, blurring shape affected the sincerity of the resulting expression (ANOVA result for main effect of blur-level:  $F(4,44)=2.915$ ,  $p<0.05$ ). In both of these effects, the perceived intensity and sincerity decreased with increasing blur-level. These results are especially interesting as they demonstrate that although recognizable expressions could be encoded using fewer bytes for data transmission (see discussion of recognition results), doing so would degrade the perceptual intensity and sincerity of these expressions.

**Summary:** Experiment 2 has shown how degrading different information channels affects the perceptual measures. Interestingly, blurring the shape and texture channels had *no effect* on recognition rates as long as the motion channel was present, although it did increase reaction time in both conditions. As soon as the motion channel was missing—as in the static peak frame condition—shape blurring strongly affected recognition performance. In the context of perception research, this result shows how perceptual processing is able to

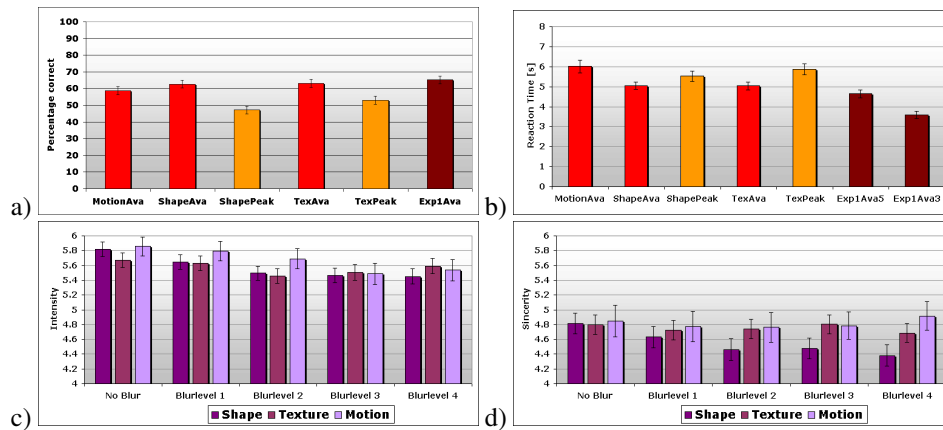


Fig. 7. Effects of motion, shape and texture blur in Experiment 2 for a) recognition, b) reaction times, c) intensity, and d) sincerity.

compensate for degraded information by shifting the analysis to a different information channel. As could be seen in the reaction time results, however, this shift comes at the cost of increased processing time.

Finally, we found clear effects of both shape and motion blurring on the perceived intensity as well as of shape blurring on the perceived sincerity of facial expressions. This underlines the importance of investigating facial expressions using a variety of different tasks or measures in order to capture a more complete picture of our ability to judge and evaluate faces.

## 2.4 Experiment 3 - The effect of eyes

One of the biggest drawbacks, in terms of realism and perceptual plausibility, of the avatar as it was used in the two previous experiments is its lack of eyes and teeth. It is well-known, for example, that humans are very sensitive to eye motion [Langton et al. 2000]. Indeed, recognition of some expressions - such as thinking - seems to rely critically on the presence of a characteristic eye motion ([Cunningham et al. 2004]). The results of the previous two experiments provided us with a baseline from which we could explicitly examine the effect of eyes and eye motion on the behavioral measures used in this study. Moreover, given the previously demonstrated ability of the human visual system to use motion information to partially compensate for missing or degraded static information, we suspected that the addition of eyes and eye motion would have differential effects on static and dynamic versions of facial expressions.

**2.4.1 Methods.** Both the setup and the experimental tasks remained the same as in Experiment 1. In this experiment, we investigated the addition of eyes and teeth on the different experimental measures both for static as well as for dynamic presentation of facial expressions.

Whereas adding teeth to the avatar is a fairly trivial task, modeling eye motion (be it



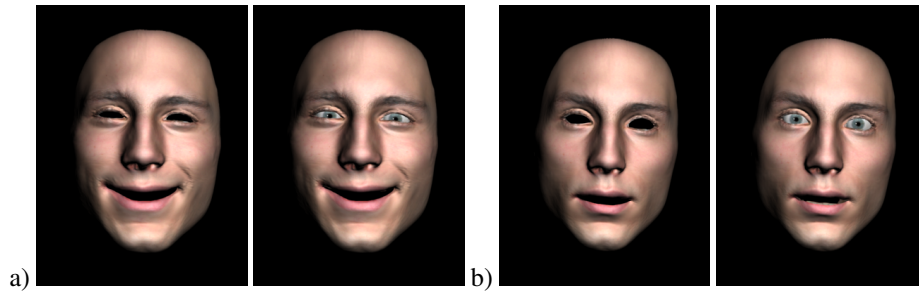


Fig. 8. Examples of the a) “happy” and b) “fear” expressions with and without eyes

“bottom-up” using recorded data or even fully model-driven [Itti et al. 2004]), is highly complex and any mistakes will most likely have an adverse effect on recognition performance. Unfortunately, the quality of recordings on which the avatar was based did not allow for either the automatic extraction of eye motion through tracking algorithms or for a manual reconstruction of the eye motion of our actor. Therefore, we followed a generic procedure for adding eyes to the avatar: First, two textured, spherical eyes were inserted into the eye sockets of the avatar. Second, eye motion was added to the animations by constraining the eyes to always look at the camera. Although the resulting animations using this simple scheme “looked natural” (see Figure 8 for two examples), we are fully aware of the limitations of our approach. Nevertheless, by comparing the animations with and without eyes across the whole spectrum of our measures, we can determine the perceptual quality of this approach. Furthermore, this approach represents the simplest possible way of introducing eyes, and thus provides a sort of baseline. In order to be able to evaluate a differential treatment of static versus dynamic information, the experiment used both types of animations in addition to static (peak) frames of the animated sequences, which were shown for the duration of the corresponding expression.

The crossing of 5 repetitions, 7 expressions, 2 motion types (dynamic versus peak only), and 2 eye conditions (present versus absent) yielded 140 trials. The latter two conditions will be jointly denoted by the terms Avatar, AvaPeak, Eyes, and EyesPeak.

**2.4.2 Results and Discussion.** The 140 trials were shown to a third group of 12 participants (none of whom had participated in the previous experiments). The results were analyzed using the same statistical methods as in the previous experiments.

**Recognition:** Figure 9 shows the recognition results collapsed across repetitions and expressions (ANOVA result for main effect of animation type:  $F(3,33)=30.647$ ,  $p<0.001$ ). This highlights the influence of the temporal context (“Static” versus “Dynamic”) as well as the effect of adding eyes and teeth (“Avatar” versus “Eyes”). Consistent with the results of the previous two experiments, the addition of dynamic information yielded a large increase (a gain of almost 20%, T-test result:  $p<0.001$ ) in recognition performance. The addition of eyes and teeth to the avatar also yielded a significant, albeit smaller, effect (a gain of 5%, T-test result:  $p<0.05$ ). The reason for this seems to lie in an interaction between motion context and the presence of eyes, as can be seen in Figure 10 (which breaks recognition performance down by animation type and expression; ANOVA result for inter-

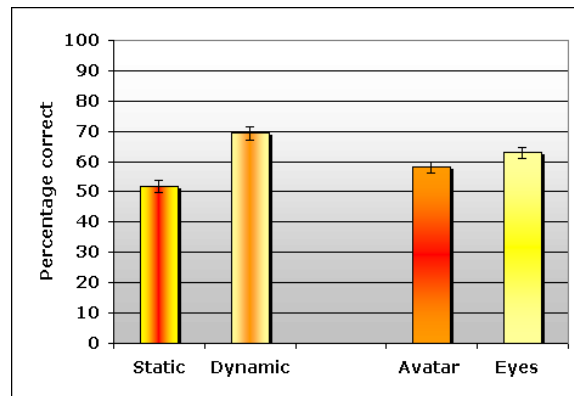


Fig. 9. Recognition results of Experiment 3 as a function of temporal context (left) and addition of eyes and teeth (right).

action between expression and animation type:  $F(18,198)=8.382$ ,  $p<0.001$ ). The addition of eyes almost always results in a significant performance gain for the dynamic sequences. In the static sequences, however, the addition of eyes sometimes actually hurt performance. This performance decrement is most notable for “thinking” and “happy”. Interestingly, the addition of eyes also hurt the *dynamic* version of “thinking”. Given that the addition of eye motion has a large, positive influence for real video sequences of this actor’s “thinking” expression (see [Cunningham et al. 2004]), the decrease in performance found here would seem to represent a limitation of our approach of integrating eyes into the final animations.

The “thinking” expression also produces an unusual pattern of results for motion context: It seems that the motion of the face for this expression actually impairs recognition performance. This stands in stark contrast to the other expressions, where the addition of motion results in a significant increase in performance. A closer look at the data shows that participants mistook the moving sequences most often for “confusion”, an error that was also found in the original study by [Cunningham et al. 2004]. Nevertheless, it is clear that both eye motion and overall facial motion need to be changed for the “thinking” expression in order to reach performance levels that are comparable to the original video recordings (see Figure 10).

**Reaction Time:** Reaction time data for this experiment showed a non-significant *trend* towards a faster recognition time for dynamic sequences. A post-hoc analysis of the data which excluded very long response times of two participants (whose average response times were three times higher than for other participants) showed a significant advantage for the dynamic sequences over the static sequences (ANOVA result for main effect of temporal context:  $F(1,11)=3.603$ ,  $p<0.05$ ). This, combined with the apparent absence of an effect of eyes and teeth on response times, confirms the critical role of motion in processing of facial expressions that was found in the previous two experiments.

**Intensity, Sincerity, and Typicality:** Figure 11 shows the ratings that were obtained for intensity, sincerity, and typicality in this experiment. Consistent with the previous experiments, the addition of temporal information produced a significant increase in *intensity*

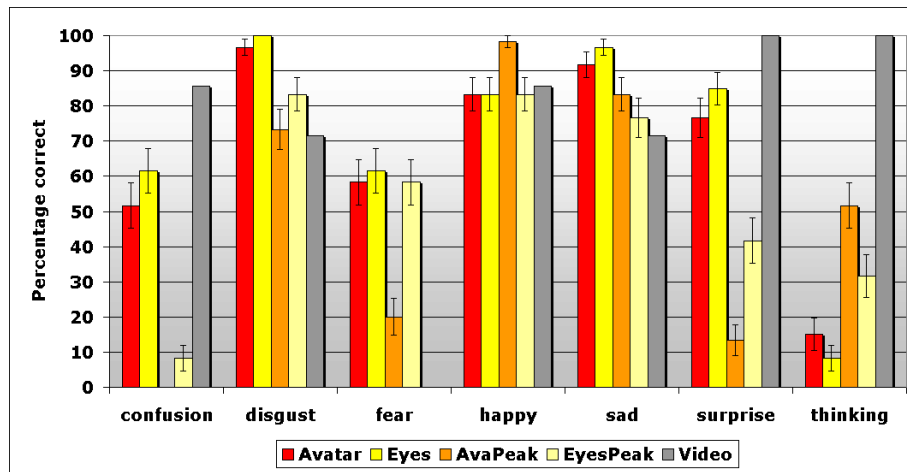


Fig. 10. Recognition results of Experiment 3 broken down by animation type and individual expressions (the bars labelled “Video” show data from [Cunningham et al. 2004]).

ratings (ANOVA result for main effect of temporal context:  $F(1,11)=11.851$ ,  $p<0.01$ ). This further underlines the importance of motion in the perception of facial expressions. Interestingly, the addition of eyes and teeth for the avatar also influenced the perceived intensity of expressions—this effect was particularly pronounced for “surprise”, “fear”, and “happy” (ANOVA result for main effect of eyes:  $F(1,11)=22.019$ ,  $p<0.001$ ).

The ratings for *sincerity* also showed a small, but significant, increase with presence of eyes in the animations (ANOVA result for interaction between eyes and expression:  $F(6,66)=3.967$ ,  $p<0.01$ ). Again, this increase was most pronounced for the “surprise” and “fear” expressions. Overall, we found no difference for static versus dynamic conditions in the sincerity ratings (ANOVA result for main effect of temporal context:  $F(1,11)=0.748$ ,  $p=0.406$ ).

Interestingly, both intensity and sincerity showed no difference in ratings for the “thinking” expression when adding eyes—a result, which is in accordance with the inferior recognition performance for this expression. This shows that the higher-level perceptual measures provide additional information related to the perceptual quality of animations.

Finally, the ratings for *typicality* (see Figure 11) were not affected by presence of eyes and teeth (ANOVA result for main effect of eyes:  $F(1,11)=0.602$ ,  $p=0.454$ ). Typicality ratings were, however, marginally affected by the addition of motion, indicating that static renderings of facial expressions with our avatar might result in less typical expressions (ANOVA result for main effect of temporal context:  $F(1,11)=3.509$ ,  $p=0.08$ ).

**Summary:** In this experiment, we used the perceptual measures developed in the previous two experiments to evaluate the impact on perceptual realism of the addition of eyes and teeth to our avatar. We found that their addition significantly increased recognition accuracy, intensity ratings, and sincerity ratings, despite the fact that the method of determining eye motion was relatively simple. Given the large body of evidence attesting to the importance of eye motion (see [Langton et al. 2000; Garau et al. 2001]), the perceptual implications of this result for the understanding of human perception are perhaps less far

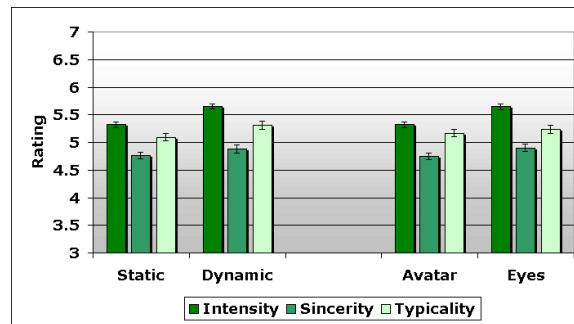


Fig. 11. Ratings of Experiment 3 for intensity, sincerity, and typicality as a function of temporal context (left) and addition of eyes and teeth (right).

reaching than the implications for the design of the avatar and for the framework in general. Using a battery of perceptual measures enabled us to paint a much clearer picture of the perceptual quality of the animation as well as to highlight specific areas that need improvement. This is especially well demonstrated with the “thinking” expression of our avatar, for which both the eye motion as well as the facial motion need to be improved in order to achieve real-world performance. Finally, this experiment again confirmed the importance of facial motion across a variety of perceptual measures (recognition accuracy, intensity, and typicality). This is an interesting result for both perceptual as well as computer animation researchers as it opens up an orthogonal dimension of analysis to which one might need to pay more attention in order to fully characterize the perceptual dimensions of facial expressions.

### 3. SUMMARY

Our proposed framework based on psychophysical methodology makes it possible to investigate dimensions that have not been accessible by standard methodologies in computer graphics or computer animation (see Figure 12 for a brief summary of the important results of this paper). Whereas the ultimate goal of computer graphics usually has been to achieve maximum physical realism, perceptual measures seem better suited to capture success or failure of animated facial expressions. We have shown how investigating these factors can help to identify the relative perceptual contribution of selected information channels of facial expressions. In particular, our first experiment elucidated the roles of rigid head motion, natural timing of internal facial motion, and the frequency content of internal facial motion. These results not only provide insights into how one could create animated facial expressions that are quickly and reliably recognized, they also help to ensure the perceptual fidelity of the animations in terms of intensity, sincerity, and typicality. With the help of our experimental methodology, it becomes possible to evaluate different animation techniques as well as compare their performance with real-world data [Cunningham et al. 2003; Cunningham et al. 2004]. Our work has provided an initial set of results, which can be used as a benchmark in future studies involving more sophisticated animation techniques.

Experiment 2 provided some initial insights into the separate roles of motion, shape, and texture in the perception of facial expressions. Interestingly, we did not find any effects of

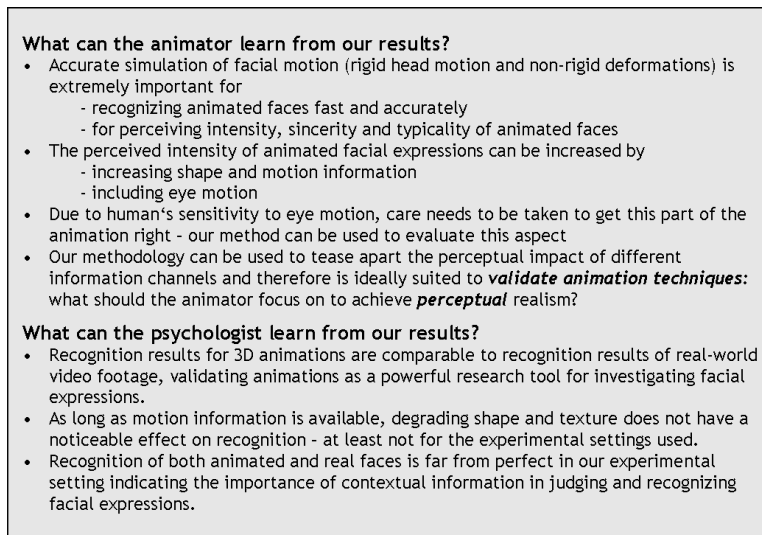


Fig. 12. Overview of the most important insights which can be gained from our research.

texture blurring on the performance of participants, which indicates that texture was not an important information channel (at least in our experimental setup). One reason for this might be our particular choice of material properties, which in combination with a Phong lighting model could have resulted in a perceptual bias towards shape information. An interesting follow-up experiment would be therefore to more closely investigate the impact of lighting and material properties on our measures. Most importantly, however, our recognition results have shown how the motion channel can compensate for degraded information in the shape channel. As seen in the static condition, if this motion information is missing, blurring shape strongly affects recognition. This result is in accordance with Experiment 1, where we found a strong impact of temporal information on recognition performance. This provides converging evidence for the perceptual importance of the temporal structure of the visual input in processing of facial expressions. In addition, in Experiment 2 we found clear effects of shape blurring on both intensity and sincerity judgments. This demonstrates that, even though reduced shape information does not affect recognition, it does influence the perceived quality of the expression.

Finally, Experiment 3 demonstrated how our battery of perceptual measures can be used to evaluate an improved version of our avatar. More specifically, we have shown that adding eyes (and teeth) results not only in an increase in recognition performance, but also affects the perceived intensity and sincerity of the facial expressions. Given the simple technique that was used to generate eye motion, this result is encouraging. Nevertheless, the full set of results has pointed to a number of areas in which performance needs to be improved in order for the avatar to be “perceptually equivalent” to its real counterpart.

Taken together, our experimental results shed light onto how humans use different information channels to evaluate various aspects of facial expressions—results that are not only of interest for perception research but also for computer animation (including the use of avatars in virtual reality and computer-supported collaborative work). In future work, we

plan to extend the capabilities of the avatar. Most importantly, we need to add more expressions and different actors in order to be able to fully generalize our results to perception of facial expressions (for a discussion of performance differences of actors, see [Cunningham et al. 2004]). In addition, we will conduct experiments directly comparing the video sequences and animations across all perceptual measures used here. This will allow us to integrate not only recognition accuracy but also higher-level measures such as intensity, sincerity, and typicality to establish a perceptual baseline for the video sequences. At the same time we will also record more expressions with the 4D-Scanner, as this allows us to capture ground-truth shape deformations as well as texture information against which we can evaluate the avatar. In addition, a large corpus of 4D-Scans would eventually allow us to create high-resolution, spatio-temporal models of facial expressions—a highly desirable goal both for computer animation and perception research.

### Acknowledgments

We would like to thank Kai Wolf for providing excellent technical support for our scanning hardware as well as Winfried Ilg and Martin Kampe for their support in processing the motion capture data. Finally, we are grateful for the input of several reviewers on portions of these data, which were published at APGV 2005.

### REFERENCES

- ADOLPHS, R. 2002. Recognizing emotions from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews* 1, 1, 21–61.
- BICKMORE, T. AND CASSELL, J. 2005. *Advances in Natural, Multimodal Dialogue Systems*. New York: Kluwer Academic, Chapter Social Dialogue with Embodied Conversational Agents.
- BLANZ, V., BASSO, C., POGGIO, T., AND VETTER, T. 2003. Reanimating faces in images and video. In *EUROGRAPHICS '03 Proceedings*. 641–650.
- BLANZ, V. AND VETTER, T. 1999. A Morphable Model for the Synthesis of 3D Faces. In *SIGGRAPH'99 Proceedings*. 187–194.
- BREIDT, M., WALLRAVEN, C., CUNNINGHAM, D. W., AND BÜLTHOFF, H. H. 2003. Facial animation based on 3d scans and motion capture. *SIGGRAPH '03 Sketches & Applications*.
- BULL, P. 2001. State of the art: Nonverbal communication. *The Psychologist* 14, 644–647.
- CALDER, A. J. AND YOUNG, A. W. 2005. Understanding the recognition of facial identity and facial expression. *Nat Rev Neurosci* 6, 8 (Aug), 641–651.
- COSKER, D., PADDOCK, S., MARSHALL, D., ROSIN, P. L., AND RUSHTON, S. 2004. Towards perceptually realistic talking heads: models, methods and McGurk. In *APGV 2004 - Symposium on Applied Perception in Graphics and Visualization*. 151–157.
- CUNNINGHAM, D. W., BREIDT, M., KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. H. 2003. How believable are real faces: Towards a perceptual basis for conversational animation. In *Computer Animation and Social Agents 2003*. 23–39.
- CUNNINGHAM, D. W., KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. H. 2004. The components of conversational facial expressions. In *APGV 2004 - Symposium on Applied Perception in Graphics and Visualization*. ACM Press, 143–149.
- EKMAN, P. 1972. *Universal and cultural differences in facial expressions of emotion*. University of Nebraska Press, 207–283.
- EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable videorealistic speech animation. In *ACM Transactions on Graphics*. ACM Press, 388–398.
- FLEMING, R. W., JENSEN, H. W., AND BÜLTHOFF, H. H. 2004. Perceiving translucent materials. In *APGV '04: Proceedings of the 1st Symposium on Applied perception in graphics and visualization*. ACM Press, New York, NY, USA, 127–134.

- GARAU, M., SLATER, M., BEE, S., AND SASSE, M. A. 2001. The impact of eye gaze on communication using humanoid avatars. In *Proceedings of SIGCHI '01*. ACM Press, New York, NY, USA, 309–316.
- GUENTER, B., GRIMM, C., WOOD, D., MALVAR, H., AND PIGHIN, F. 1998. Making faces. In *SIGGRAPH '98 Proceedings*. ACM Press, New York, NY, USA, 55–66.
- ITTI, L., DHAVALA, N., AND PIGHIN, F. 2004. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*. 64–78.
- KAJIYA, J. T. 1986. The rendering equation. In *SIGGRAPH '86 Proceedings*. 143–150.
- KNAPPMAYER, B., THORNTON, I., AND BÜLTHOFF, H. 2003. The use of facial motion and facial form during the processing of identity. *Vision Research* 43, 2.
- KÖHLER, K., HABER, J., YAMAUCHI, H., AND SEIDEL, H. 2002. Head shop: generating animated head models with anatomical structure. In *SCA '02: Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM Press, 55–63.
- LANGER, M. AND BÜLTHOFF, H. 2000. Measuring visual shape using computer graphics psychophysics. In *Rendering Techniques 2000. Proceedings of the Eurographics Workshop Rendering Techniques*. 1–9.
- LANGTON, S., WATT, R., AND BRUCE, V. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends Cogn Sci* 4, 2 (Feb), 50–59.
- LEE, Y., TERZOPOULOS, D., AND WALTERS, K. 1995. Realistic modeling for facial animation. In *SIGGRAPH '95 Proceedings*. 55–62.
- LEOPOLD, D. A., O'TOOLE, A. J., VETTER, T., AND BLANZ, V. 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat Neurosci* 4, 1, 89–94.
- MAGNENAT-THALMANN, N., PRIMEAU, N., AND THALMANN, D. 1988. Abstract muscle actions procedures for human face animation. *Visual Computer* 3, 5, 290–297.
- MYSZKOWSKI, K., TAWARA, T., AKAMINE, H., AND SEIDEL, H.-P. 2001. Perception-guided global illumination solution for animation rendering. In *SIGGRAPH '01 Proceedings*. ACM Press, New York, NY, USA, 221–230.
- O'SULLIVAN, C., HOWLETT, S., McDONNELL, R., MORVAN, Y., AND O'CONOR, K. 2004. Perceptually adaptive graphics. In *Eurographics '04, State-of-the-art-Report 6*.
- PARKE, F. 1982. Parametrized models for facial animation. *IEEE Computer Graphics and Applications* 2, 9, 61–68.
- RIZZO, A. A., NEUMANN, U., ENCISO, R., FIDALEO, D., AND NOH, J. Y. 2001. Performance-driven facial animation: Basic research on human judgments of emotional state in facial avatars. *CyberPsychology and Behavior* 4, 471–487.
- STOKES, W. A., FERWERDA, J. A., WALTER, B., AND GREENBERG, D. P. 2004. Perceptual illumination components: a new approach to efficient, high quality global illumination rendering. *ACM Trans. Graph.* 23, 3, 742–749.
- TERZOPOULUS, D. AND WATERS, K. 1990. Physically-based facial modeling, analysis and animation. *J. of Visualization and Computer Animation* 1, 73–80.
- WILLIAMS, L. 1990. Performance-driven facial animation. In *SIGGRAPH '90 Proceedings*. ACM Press, New York, NY, USA, 235–242.

December, 6th, 2005.