Psychophysical Investigation of Facial Expressions Using Computer Animated Faces

Rita T. Griesser*
Max Planck Institute
for Biological Cybernetics
Tübingen, Germany

Douglas W. Cunningham†
WSI for Computer Science
University of Tübingen
Germany

Christian Wallraven‡
Max Planck Institute
for Biological Cybernetics
Tübingen, Germany

Heinrich H. Bülthoff§
Max Planck Institute
for Biological Cybernetics
Tübingen, Germany

Abstract

The human face is capable of producing a large variety of facial expressions that supply important information for communication. As was shown in previous studies using unmanipulated video sequences, movements of single regions like mouth, eyes, and eyebrows as well as rigid head motion play a decisive role in the recognition of conversational facial expressions. Here, flexible but at the same time realistic computer animated faces were used to investigate the spatiotemporal coaction of facial movements systematically. For three psychophysical experiments, spatiotemporal properties were manipulated in a highly controlled manner. First, single regions (mouth, eyes, and eyebrows) of a computer animated face performing seven basic facial expressions were selected. These single regions, as well as combinations of these regions, were animated for each of the seven chosen facial expressions. Participants were then asked to recognize these animated expressions in the experiments. The findings show that the animated avatar in general is a useful tool for the investigation of facial expressions, although improvements have to be made to reach a higher recognition accuracy of certain expressions. Furthermore, the results shed light on the importance and interplay of individual facial regions for recognition. With this knowledge the perceptual quality of computer animations can be improved in order to reach a higher level of realism and effectiveness.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Realism—Animation; J.4 [Computer Application]: Social and Behavioural Sciences—Psychology; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Animations;

Keywords: perceptual graphics, psychophysics, facial animation, facial expressions, 3D-scanning, face recognition

1 Introduction

Exact modeling and animation of facial expressions turns out to be a very difficult task. Despite immense efforts in computer hardware and software development, including the improvement of sophisticated algorithms for computer graphics, today still no computational system exists that approximates the performance of humans. The main questions that arise are: *"What is needed to produce perceptually realistic images?"*, *"When are images real enough?"* and *"How can one map the space of expressions using the minimum possible amount of computing power and time?"* Considering facial expressions, one approach to meet these questions is to understand the details of perceptual and cognitive issues underlying human facial motion and recognition. Moreover, explicit information about animation parameters affecting the chosen perceptual measures is required. This knowledge can support efficient computing of image data in so far as time consuming rendering techniques could be reserved for significant facial regions. Finally, this in turn can lead to a higher level of realism and effectiveness (see Figure 1, [Wallraven et al. 2005]).

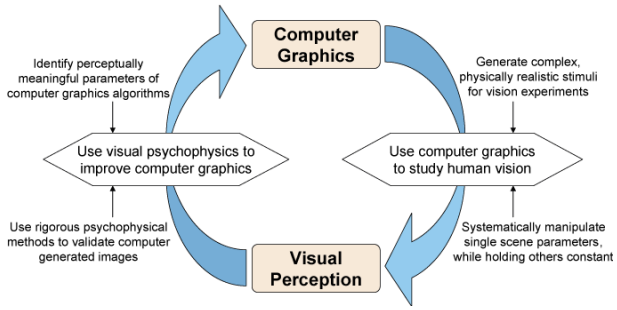

Figure 1: Illustration of the close link between computer graphics and visual perception research.

Facial expressions play a decisive role in communication. During a conversation, facial expressions can emphasize or modify the meaning of what is being said. When emphasizing a word in a sentence, the facial expression reflects this emphasis. For example, in the sentence: "Take the *blue* bowl, not the *red* one", the contrast in stressing the words blue and red is reflected by certain facial expressions. The relationship between vocal and facial emphasis is so strong that it can be nearly impossible to produce the proper vocal emphasis pattern without producing the accompanying facial motion. For example, a verbal statement of appreciation receives a totally different meaning when it is accompanied by a look of displeasure. Additionally, facial expressions may be useful in controlling the flow of a conversation, as the speaker can adapt his reaction according to the listener's facial expression: if the listener nods, the speaker knows that he or she is understood and is therefore encouraged to continue. Otherwise, if the listener looks confused, the speaker will probably explain his message again in a more detailed fashion.

Humans are able to recognize and perform a variety of different facial expressions. As shown in previous studies [Cunningham et al. 2005], movements of the three facial regions mouth, eyes, and eyebrows play a crucial role in the recognition of conversational facial expressions. For example, happiness seems to be mainly defined by motion of the mouth, confusion in contrast by motion of the eyebrows.

*e-mail: rita.griesser@tuebingen.mpg.de
†e-mail:douglas.cunningham@gris.uni-tuebingen.de
‡e-mail:christian.wallraven@tuebingen.mpg.de
§e-mail:heinrich.buelthoff@tuebingen.mpg.de

APGV 2007, Tübingen, Germany, July 26–27, 2007.
© 2007 ACM 978-1-59593-670-7/07/0007 $5.00

One reason why there is still no computational system that approximates the performance of humans is that humans are amazingly good in perceiving small differences in facial motion and meaning. This high cognitive performance is due to their use of comprehensive previous knowledge about faces and facial expressions. In every face, the shape and spatiotemporal alignment of parts like eyes, eyebrows, nose, and mouth follow certain natural rules. These rules determine which facial expressions actually occur in reality. Violating these rules leads to an unnatural facial movement, even if the differences are quite subtle. Yet, even if a physically accurate virtual human face imitates all spatial and temporal aspects of facial motion perfectly, it is still not guaranteed that this facial expression will be identified correctly. Moreover, there are more perceptual and cognitive parameters that need to be captured by an animation system. For example, sincerity seems to play a central role: if an expression is synthesized well, it may be correctly identified, but this still does not mean that it will be considered to be sincere. For example, who would purchase a life insurance policy from a virtual insurance agent who looks dishonest or insincere, independent of how realistic s/he looks? Likewise, who enjoys watching animated movies if the characters' facial expressions do not correspond to their intended emotional constitution? To avoid such violations of our expectations, explicit knowledge about natural facial movements and their importance is required. In addition to the media, other application areas would be in human-computer interaction, medical rehabilitation, multimedia applications as well as biometrics and the film industry.

In previous studies, video sequences of actors were used to investigate recognition, believability and intensity of conversational facial expressions [Cunningham et al. 2005]. In one set of studies, advanced computer graphics techniques were used to "freeze" several facial regions while leaving the others intact. These manipulated video sequences provided a reasonable approach for determining the important regions for facial recognition. These video-based techniques, however, are limited in their applicability. For example, examining the temporal synchronicity of individual facial regions during an expression is nearly impossible as the manipulation of speed in video sequences results in unbalanced movements. For this reason, a computer animated avatar was designed by [Breidt et al. 2003] combining 3D face scans and motion capture data. State-of-the-art 3D scanning systems provide very high spatial resolution while motion capture systems provide very high temporal resolution. With these two techniques, it is possible to create facial animations with a sufficient level of realism. By importing the animation sequences into the a commercial animation program, such as 3ds Max, one can selectively manipulate facial regions and the rigid head motion. Moreover, motion can be manipulated by nonlinear time functions, making the accurate investigation of temporal interplay of individual facial regions possible.

2 Background

2.1 Perception of Facial Expressions

The study of facial expressions goes back to the 19th century, when researchers wanted to learn more about the relationship between emotions, movements, and facial expressions. In 1872, C. Darwin suggested that facial expressions are innate and consist of habitual movements that depend on emotions and the state of the mind. He observed several expressions carefully and identified characteristic facial movements that appeared for the different expressions [Darwin 1872]. Through the years, a variety of issues have been investigated, for example, how emotions are produced and recognized, or whether there is a genetic coding of facial movements. For the investigations of facial motions, usually photos with persons performing expressions were shown to observers who had to identify the expressions. Bassili was one of the first to examine the role of *movement* of the facial surface [Bassili 1978]. He covered faces with black makeup and numerous white spots and recorded the performed facial expressions. He subsequently showed the recordings both with white spots and normally illuminated faces to observers. The results demonstrated that the white spots were sufficient to recognize the expressions, but that recognition accuracy was higher when the complete face was visible.
A number of systems have been developed to describe the movements of facial expressions. One of the most well-known representational systems is the *Facial Acting Coding System* (FACS) by Ekman and Friesen [Ekman and Friesen 1978]. It decomposes facial expressions into 46 small facial movement units, called "action units", which correspond to single or multiple facial muscle activations. The combination of action units produces expressions. Thus, this system provides an intuitive, semantic basis also for facial animation. So far, most studies investigated facial expressions that accompany basic emotions such as *happiness*, *sadness*, *fear*, *anger*, *surprise*, and *disgust*.

2.2 Computer Graphics in Facial Expression Research

To produce highly realistic facial animations, new face models and advanced animation techniques had to be developed. Parke developed the first parameterized face model in 1974, with the goal of producing facial animations quickly [Parke and Waters 1996]. Using photogrammetric techniques, he collected 3D data from real faces and created animations by interpolating between the facial expressions. From then on, the development of three-dimensional facial animation dominated in research, although two-dimensional animations were also refined with later applications in cartoon animation. Waters developed a muscle-based facial animation model in 1987 that allows one to create realistic basic facial expressions [Waters 1987]. In this approach, the simulation of skin deformation, which is not specific to the texture, is controllable by a limited number of parameters. Sifakis demonstrated a similar muscle-based animation system that uses motion capture data in a non-linear optimization process to estimate facial muscle activation parameters [Sifakis et al. 2005]. Blanz and Vetter developed an algorithm that fits a blendshape model onto a single image, resulting in an estimation of the geometry and texture of the person's face [Blanz and Vetter 1999]. In 2003, Joshi proposed an automatic, physically motivated segmentation that learns the controls and parameters directly from the set of blendshapes [Joshi et al. 2003]. Williams first introduced performance-driven animation in the late 80s, aquiring the expressions of real faces in two-dimensional space and applying them to computer generated faces [Williams 1987]. Recent animation systems derive facial movements in three-dimensional space by tracking markers attached to a person's face. Breidt et al. presented such a model in 2003, combining 3D scans and motion capture data for highly realistic facial animation [Breidt et al. 2003]. Martin et al. have presented a model of multimodal complex emotions involving gesture expressivity and blended facial expressions [Martin et al. 2006]. They defined a copy-synthesis approach to drive an embodied conversational agent from different sources of information.

2.2.1 Video Recordings and Computer Animated Faces

One important issue for the investigation of facial expressions is the determination of which individual facial regions carry information during a conversation. To partially answer this questions, [Cunningham et al. 2005] employed advanced computer graphics and computer vision techniques to track and freeze selected regions of a face in video recordings. The results show that facial expressions can be recognized from individual components and combina-

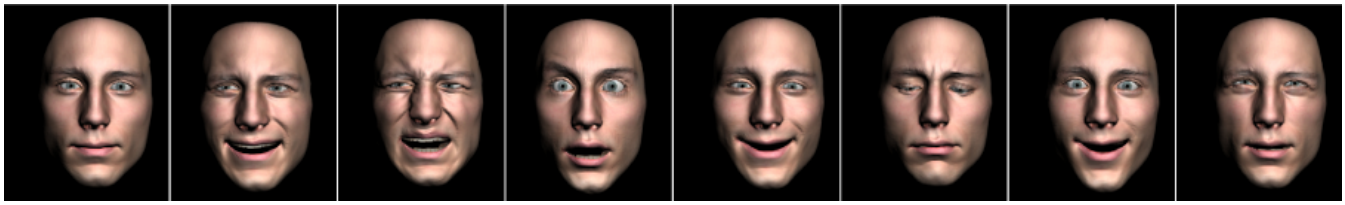

Figure 2: Computer animated avatar performing the seven basic peak expressions. From left to right: neutral, confusion (not understood), disgust, fear, happy, sad, pleasant surprise, thinking (about a problem).

tions of these components nearly as well as they can be from the original sequences. However, video recordings have the great disadvantage that, for example, the temporal interplay of individual facial regions cannot be examined properly. Therefore, a computer animated avatar (see Figure 2) was designed at the Max Planck Institute for Biological Cybernetics, providing the possibility of selective animating distinct facial regions with a high degree of control. Several experiments have already been conducted using this computer animated avatar, to examine its realism, perception and spatiotemporal characteristics [Wallraven et al. 2005].

3 Recording and Animation of Facial Expressions

For the creation of the computer animated avatar, the animation system shown in Figure 3 was applied. It consists of two branches: the upper branch describes the generation of facial geometry, the lower branch the calculation of amplitude and timing of facial motion. Since high resolution data concerning space, time, and texture are required for the examination of facial movements, a 3D scanning and a motion capture system were combined. After initial data acquisition, the data had to be cleaned and processed. Additionally, for better visual fidelity, texture maps of skin, eyes and teeth were recorded (using with a a digital SLR camera) high-resolution digital color camera and applied to the morph shape. Finally, using a 3D animation program facial animation can be created from these morph shapes and the motion capture data.

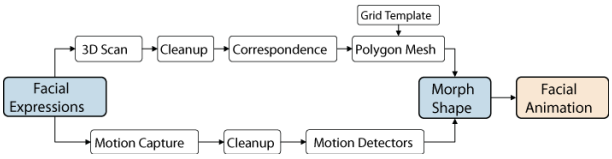

Figure 3: The facial animation system used for the creation of the computer animated avatar.

3.1 3D Face Scans and Motion Capture

In order to record the facial expressions, a dynamic structured light scanner developed by ABW GmbH was used. A 3D measurement is acquired by projected four patterns, each containing vertical stripes, in rapid succession onto the face. The patterns are viewed by two high-speed digital video cameras, located at 22 degrees on the left and right side of the projector. The well-defined sequence of stripes allows one to determine a unique correspondence for each pixel in the two cameras. Since the spatial configuration of projector and cameras is known, the deformation of the stripe patterns provides information about the geometry of the scanned face. With the application of image processing techniques, the exact location of the edges of the stripes in each video image can be detected and accurate 3D data can then be calculated from the images with triangulation. As one scan took about two seconds, the expression needed to be held for just a short time.

In order to obtain temporal characteristics of the facial expressions, facial motion data was captured with an optical Vicon 612 Motion Capture system from the performer who was previously scanned. The system consists of six cameras (running at 120 Hz) arranged in a semi-circle at a distance of 150 cm from the face. 69 reflective markers were attached to the skin of the person's face, 3 more on a target on the person's head for tracking rigid head motion. The markers had a diameter of 2mm and did not affect facial motion as they were not noticed by the performer after a few seconds. The person was then asked to perform the same expressions as during the scanning process.

3.2 The Animated Avatar

Three dimensional scans of a neutral expression and each peak expression were obtained. The individual scans were cleaned and put into correspondence with each other using a manually designed control mesh in order to create the morph shapes. Second, motion capture data were recorded. To determine the facial motion, rigid head motion was temporarily removed. Then the distance between previously defined markers was used to transfer the real-world-coordinates into morph animation channels, using one channel for each morph shape. The weights for the different morph channels were derived by simple linear detectors, and thus produce morph animation based on the amplitude and timing of marker motion. Rigid head motion corresponding to the original motion capture data was subsequently added to the final avatar. An exact geometric match between the movements during the scanning and the movements during the motion capturing process is not essential since the motion capture data is used for timing and qualitative analysis only.

3.3 Recording Protocol

Seven facial expressions were performed by an amateur actor: confusion (as if something said was not understood), disgust, fear, happy, pleasently surprised, sad, and thinking (as if solving a problem – see Figure 2). Before the recordings were started, a brief scenario was described in detail and the actor was asked to put himself in that situation and to react with the appropriate facial expression. Furthermore, the actor was asked to react without speaking, but was encouraged to emit nonverbal sounds that helped him to sympathize with the described situation.

4 Psychophysical Experiments

Overall, the following three experiments attempt to elucidate how much information different facial regions, and combinations of regions, carry for different expressions. More specifically, the first experiment examined the sufficiency of individual regions, the second experiment the role of pairwise combination of regions. The third experiment examined all possible three-way combinations. Each

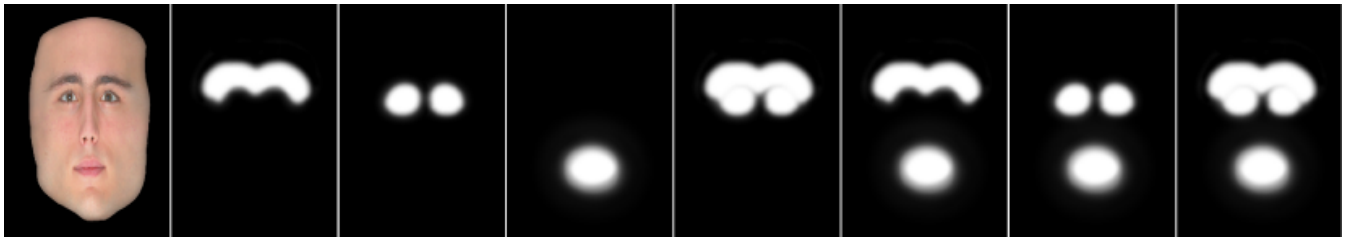

Figure 4: Image masks applied to determine facial regions for the animations. The white areas define the regions and their intensity for the animation in 3ds Max. From left to right: Original texture map (Orig), eyebrows (B), eyes (E), mouth (M), eyebrows and eyes (BE), eyebrows and mouth (BM), eyes and mouth (EM), eyebrows, eyes, and mouth (BEM).

experiment also contained the original animation condition, as a baseline.

The individual expressions and the single facial regions were selected based on previous experiments [Nusseck et al. 2007], who used video recordings of real faces of different actors and actresses. A few of the specific conditions used here were also used in these previous studies. A comparison of the present results with previous results will yield insights into the quality of the animated avatar and the usability of the applied manipulation technique. Additionally, the same actor that was used to create the avatar also was also part of the previous study which allows for a direct, more detailed comparison of the results.

4.1 Facial Regions

The individual basic facial regions, here also referred to as *information channels*, are the eyes, eyebrows, and the mouth. To examine the degree to which the rest of the face contains important information for recognition, the movement of all single regions together was compared to the original motion. It is important to note that eyeball motion carries information about certain expressions, and is also decisive for realistic appearance [Wallraven et al. 2005]. [1] Altogether, the following 16 combinations, also referred to as *conditions*, were chosen for the psychophysical experiments:

Orig: This is the original animation. All facial motions, rigid head motion, as well as eyeball motion are present.
R: This animation contains rigid head motion only. The face is kept still and the eyeballs are fixed to the head. Thus, for the observer the eyeballs seem not to move relatively to the head.
B/RB: The eyebrows and closely surrounding area are animated. The animations are created with and without rigid head motion. In both cases the eyeballs move relative to the head.
E/RE: The eyes and closely surrounding area (without eyebrows) are animated. The movement is created both with and without rigid head motion. If rigid head motion is present, the eyeballs are fixed to the head. In the animation without rigid head motion the eyeballs show head-relative motion.
M/RM: The mouth and surrounding area are animated. The animations are created both with and without rigid head motion. The eyeballs are fixed to the head and thus show no head-relative motion.
BE/RBE: The areas of eyebrows and eyes are animated. The animations are created both with and without rigid head motion. If rigid head motion is present, the eyeballs are fixed to the head and show no head-relative motion. In the animation without rigid head motion the eyeballs move relative to the head.

[1] Unfortunately, the present animation system does not include eye-tracking information, and thus the eyeball motion in the avatar does not correspond exactly to the actor's real movements.

BM/RBM: Eyebrows and mouth are animated. The animations are created both with and without rigid head motion. In both cases the eyeballs move relative to the head.
EM/REM: Eyes and mouth are animated. The animations are created both with and without rigid head motion. If rigid head motion is present, the eyeballs are fixed to the head. In the animation without rigid head motion the eyeballs move.
BEM/RBEM: All single regions, eyebrows, eyes, and mouth, are animated. The animations are created both with and without rigid head motion. If rigid head motion is present, the eyeballs do not move relative to the head. In the animation without rigid head motion the eyeballs show head-relative motion.

Experiment 1 contained 6 conditions: Orig, R, B, E, M, and RBEM. Experiment 2 had 7 conditions: Orig, RB, RE, RM, BE, BM, and EM. Experiment 3 had 5 conditions: Orig, RBE, RBM, REM, and BEM. Across all three experiments, we therefore investigated the interplay of facial information channels at different levels of recognition.

4.2 Creation of Animations

The animation of the individual motions for the video sequences was created with the animation program *3ds Max*. As some parameters in 3ds Max, such as the falloff for a soft-select region, cannot be set independently for the marked vertices of the morph shapes, tools of an image editing program *Adobe Photoshop* were used to create gray-scale image masks corresponding to the conditions in 4.1 (see Figure 4). With this technique, any desired shape, size, and intensity of the regions to be animated could be defined. This also includes the ability to produce irregular fall-offs or different fall-offs for combined regions. The image masks were loaded into 3ds Max to select and mark contiguous and discontiguous vertices of the morph object. This selection of vertices was then affected by the animation according to the gray-values of the loaded mask and could be transferred onto other morph objects using MAXscript. When defining the individual regions, care was taken to choose small and precise selections to allow an accurate examination. For this reason, artifacts appearing in the transition zone of the moving and non-moving regions were deemed acceptable.

4.3 Experimental Procedure

In each of the three experiments, video sequences were presented in randomized order at 512×512 pixels on a black computer screen with a resolution of 1024×768 pixels. Each experiment had 10 participants, with a different set of 10 people being used for each experiment. The participants sat in a darkened room, in front of the screen at a distance of roughly 0.5 meters (the face on the monitor subtended a visual angle of 11.4 degrees). Each video sequence was shown in the middle of the computer screen and disappeared when its end was reached. A list of all seven expressions and the

additional option "none of the above" (an 8-alternative-non-forced-choice task, see [Cunningham et al. 2005]) was displayed on the left side of the screen both in English and German. Before the experiments started, the participants were given detailed instructions about the experimental setup. Their task in each experiment was to identify each expression by selecting one of the options from the list displayed on the side of the screen and press the appropriate button on the keyboard. Experiment 1 had 7 expressions and 6 manipulations, each of which was shown 6 times, yielding 252 trials. Experiment 2 had 7 expressions, 7 manipulations, and 6 repetitions, yielding 294 trials. Experiment 3 had 7 expressions, 5 manipulations, and 6 repetitions, yielding 210 trials. The experiments lasted between 25 and 40 minutes each.

4.4 Results and Discussion

The recognition accuracy and reaction time results for each experiment were analyzed separately using standard "analysis of variance" (ANOVA) methods with Expression, Manipulation, and Repetition as within-participants factors. In the follwing, the significant statistical effects for each measure will be discussed.

4.4.1 Recognition Accuracy

All three experiments showed a significant main effect of Expression (all F's > 6.9, all p's < 0.0001), indicating that some expresions where recognized better then others. Furthermore, overall, all expressions achieved a recognition accuracy significantly above chance level (defined as the rate that would be produced by blindly guessing – 12.5% in this case). *Disgust*, *happy*, and *sad* were very reliably identified (accuracies of 97%, 94%, and 95%, when averaged across the three experiments, respectively). *Confusion*, *pleasant surprise*, and *fear* were recognized reasonably well (61%, 73%, and 49%, respectively). For *thinking*, the recognition rate was only about 29%.

All three experiments also had a significant main effect of Manipulation (all F's > 25.4, all p's < 0.0001), indicating that the expressions were easier to recognize in some manipulations than in others. The main effect of Repetition was only significant in Experiment 1 ($F(5,45) = 2.9632$, $p < 0.05$ for experiment 1; all other F's < 1.9, all other p's > 0.12), showing a mild or non-existent improvement in recognition over the course of the repetitions. The Manipulations by Expression interaction was significant in all three experiments (all F's > 12.9, all p's < 0.0001), indicating that some manipulations affected some expressions more then others (see Figure 6). The Expression by Repetition interaction was significant for the first two experiments (all F's > 1.5, all p's < 0.05), but not for the third ($F(30,270) = 0.9487$ $p > 0.05$), indicating that when the expressions are defined by individual regions and pair-wise combinations of regions, some expressions are easier to recognize after they have been seen a few times. Neither the Manipulation by Repetition, nor the three way interaction were ever significant (all F's < 1.56, all p's > 0.068).

It has previously been shown that rigid head motion is central for certain expressions [Wallraven et al. 2005]. The present results are consistent with this: Looking across the three experiments, animations containing rigid head motion yielded higher recognition accuracies than those without (58% versus 37%, respectively) (see Figure 6). Interestingly, the original motions (condition ORIG) did not achieve higher recognition rates than the manipulated faces. This effect is probably due to the fact that the original facial expression iteself is not unambigious. Especially for the expression *disagree* motion is not clearly definable (for details see [Nusseck et al. 2007]).

A comparison of the results from the avatar with those from a real video of the same actor [Nusseck et al. 2007] shows that the avatar performs comparatively well for the expressions *confusion*, *disgust* and *surprise*. A slight decrease of 8% in correct recognition of the animated avatar is observable for *happy*, and a remarkable decrease of 46% for *thinking*. This decrease is probably due to the prominent role of eye-motion in thinking. In contrast, the expression *sad* of the animated avatar shows an increase of 38% (Figure 5). One might well consider the videos examined by [Nusseck et al. 2007] to be the "ground truth" since they show a real person. Thus, lower recognition accuracy for the animated avatar is probably caused by the animation (and differing manipulation techniques). In contrast, the extremely high recognition accuracy for the animated version of *sad* in comparison to the real version shows that real faces are not always better than animated faces.

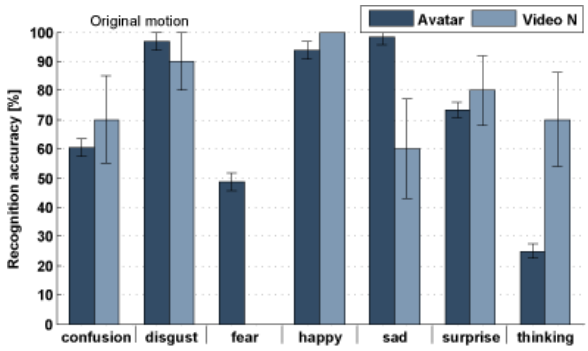

Figure 5: Recognition accuracy of original facial expressions. **Avatar** denotes the results of these experiments, **Video N** of [Nusseck 2007], for which "*fear*" was not tested.

Another possible explanation for the differences in recognition accuracy across experiments is the varying interpretation of the names describing the expressions. Since no context for the expressions was given, participants had to choose their own definitions for the terms. Thus, they might interpret the expression *confusion* as "not understood", "not knowing" or "feel unsure". Likewise, *thinking* might be interpreted as "reflecting", "dreaming" or "trying to remember". Another potential expanation is related to the fact that expressions occasionally appear in combination: *confusion* is often accompanied by a look of disgust, *fear* by surprise, and *pleasant surprise* naturally encloses happiness. Similarly, *thinking* usually contains a certain amount of confusion (and, depending on the underlying thought, maybe even disgust or happiness). Previous experiments have shown that, e.g., the expression *thinking*, reaches a much higher recognition accuracy if the peak expression is shown only [Wallraven et al. 2005], indicating that other expressions occur while the face moves away from the neutral to the peak expression. Finally, the actor's emotional state and performance ability play a role when facial geometry, motion capture data and videos were recorded. As the actor had to perform the expressions for the video recordings, the face scanning, and the motion capture data separately and hold them up to 4 seconds, expressions with the same semantic meaning might slightly differ or appear as "unnatural".

Despite the low recognition accuracy for some of the expressions, there is a clear indication which information channels drive the recognition of facial expressions. Here, the major characteristics of the seven expressions are described (see Figure 6).

Confusion Rigid head motion itself is sufficient to recognize this expression (recognition accuracy of 90%). All animations containing rigid head plus eyebrow or/ and eye motion (conditions RB, RE, RBE, RBM, REM and RBEM) were recognized as well as, or even better than, the original expression. This demonstrates that the recognition of *confusion* for this avatar is driven by rigid head motion.

Disgust In all animations that contain mouth motion and motion in at least two additional regions (conditions RBM, REM, BEM and RBEM), the percentage of correct identification is comparable to the rate of the original expression. This indicates that mouth motion is necessary, yet not sufficient to recognize *disgust*. Rigid head, eyebrow and eye motion seem to contribute important information for the recognition of this expression.

Fear All animations containing eye motion or rigid head plus mouth motion (conditions E, EM, BE, BEM, RE, RM, RBE, RBM, REM and RBEM) showed recognition rates as high as for the original expression. Thus, *fear* seems to be driven by eye motion but also by rigid head and mouth motion in combination.

Happy When mouth motion is present, the rate of correct identification rose to the rate of the original motion. In contrast, eyebrow and eye as well as rigid head motion do not contribute important information for recognizing this expression. This clearly shows that movement of the mouth is necessary and sufficient to recognize *happy*.

Sad All animations that include rigid head plus eye or mouth moution (conditions RE, RM RBE, RBM, REM and RBEM) achieved a level of recognition accuracy comparable to the original expression. Whereas eyebrow motion seems not to contribute important information, rigid head motion is necessary for the identification of *sad*.

Surprise Rigid head, eye, and mouth motion in combination (conditions REM and RBEM) are necessary to recognize this expression. Although rigid head motion itself contributes almost no information to recognize *surprise*, it is important as it plays a supporting role.

Thinking Eyebrow and eye motion themselves are sufficient to identify this expression as well as or even better than the original expression. As soon as rigid head, eyebrow and mouth motion in combination are present, the recognition rate drops considerably. This indicates that important information for identifying *thinking* is located in the region of the eyes and eyebrows. The movement of the mouth seems not be important for *thinking*.

4.4.2 Reaction Times

The overall rection time (see Figure 7) values for the original expression (condition Orig) were the same in experiment 1 and 2 (2.8 seconds), but a little higher in experiment 3 (3.4 seconds). In general, reaction times showed a decrease of about 2 seconds from the first to the fifth repetition, indicating that the participants became more familiar with the computer animated avatar and the experimental setup during the experiment. This is reflected in a significant main effect of Repetition. In fact, all three main effects were significant in all three experiments (all F's > 2.7, all p's < 0.02). The Manipulation by Expression interaction was also significant in all three experiments (all F's > 3.0, all p's < 0.0001). The Expression by Repetition interaction was only significant in Experiment 2 ($F(30,270)=1.6519$, $p < 0.051$). Likewise the three-way interaction was only significant for Experiment 2 ($F(180,1620)=1.2267$, $p < 0.05$). The Manipulation by Repetition interaction was never significant.

It is interesting to note that, on average, participants responded fastest for animations that had the highest recognition accuracies. The graphs in Figure 7 clearly shows that reaction times increased as recognition accuracy decreased. For Experiment 1 and 2, animations containing rigid head motion had lower reaction times than animations without rigid head motion (3.6 seconds vs. 3.1 seconds on average). This highlights the importance of rigid head motion for facial expression recognition. The combination of all individual facial regions (condition RBEM) resulted in reaction times that are comparable to the original expression (2.8 seconds). The same effect is observable for the combination of rigid head, eye and mouth motion (condition REM) with 3.4 seconds. In both experiments, these animations reached the same recognition accuracy as the original expression (74% and 70%, respectively). In contrast, the combinations RM and EM, which showed reaction times similar to the original expression (2.7 seconds), achieve lower recognition rates than the original expression (59% and 49% vs. 68%). This effect clearly demonstrates that recognition accuracy as well as reaction times have to be taken into account to judge the overall quality of animations.

4.4.3 Summary

The experiments have shown that the expressions of the computer animated avatar are recognized with varying degrees of accuracy. The expressions *disgust*, *happy* and *sad* achieve extremely high recognition rates (more than 90%). *Surprise* reached 70%, *confusion* 60% and *fear* 50%. Recognition accuracy for *thinking* was rather low (roughly 25%). The recognition results have also shown that, for the some expressions, motion of a single facial region can be sufficient. For example, *confusion* seems to be mostly driven by rigid head motion and *happy* by mouth motion. For other expressions, however, the combination of several regions is necessary. For example, *sad* relies mainly on rigid head plus eye or mouth motion and *thinking* on eyebrow plus eye motion. For *disgust* the combination of mouth, eyebrow, eye and rigid head motion is required, for *fear* eye, mouth and rigid head motion and for *surprise* rigid head, eye and mouth motion. It is important to note that these results are just for a single actor. Previous research has shown that, at least with real video sequences, recognition accuracy depends on the actors and actresses who performed the expressions. Some actors or actresses are better at some expressions than others. For most expressions, fortunately, there is nonetheless a strong degree of consistency regarding which regions are important. Thus, while the guidelines for the role of facial motion information in the recognition expressions derived here are no doubt to some degree specific to this one actor, there is every reason to believe that they will be true, at least qualitatively, for many other people.

The evaluation of the reaction times showed a correspondence between reaction times and recognition accuracy: The better an animation is recognized, the faster it will be recognized. When both reaction time and recognition rates are considered, it is clear that the joint usage of the information channels eyebrows, eyes, mouth, and rigid head motion (condition RBEM) as well as the combination of eye, mouth, and rigid head motion (condition REM) are optimal for effective animation of facial expressions for all seven expressions.

5 Conclusion

The results of the these psychophysical experiments have shown that the computer animated avatar is a useful tool for investigating

the importance and interplay of facial regions in expression recognition. The animation system is good enough to support recognition of some of the computer animated avatar's facial expressions with high accuracy. To achieve good performance for all expressions, however, parts of the animation system need be improved. Most critically, it seems, would be the addition of accurate eyeball motion.

Clearly, when transforming real-world, 3D motion into a 2D projection, certain information about shape, texture and motion might disappear. Therefore, it is crucial to capture the movements of the important facial regions with a high degree of accuracy. To reach this goal, more markers for deriving the motion capture data could be positioned in the regions of the mouth, eyes and eyebrows. Moreover, the polygon mesh underlying the animated faces could be subdivided further in these regions, which allows a more precise representation of these areas. Another possibility to improve the accurate representation of the facial movements might be the application of so-called "action units" instead of semantically defined regions. Action units mostly correspond to small, individual and independent facial muscle activations, and thus can be used to model facial expressions. For example, to raise the eyebrows, different parts (pars medialis and lateralis) of the same muscle (frontalis) are activated. The complex of different action units might lead to more realistic and exact movements as small facial areas can be animated independently. To this end, future research might use the action-unit animation system by Curio et al. (for details see [Curio et al. 2006]). Consistent with this approach, Schwaninger et al. examined the recognition of identity from human faces in psychophysical studies, and suggest that humans encode face parts (component information) as well as the spatial interrelationship of facial features (global configural information)[Schwaninger et al. 2006].

Despite the low recognition performance for certain expressions, the computer animated avatar is very helpful when investigating face recognition and in particular the significance of facial regions in expression recognition. A major advantage of the animated avatar is the possibility of manipulating spatiotemporal properties in a highly controlled manner. The manipulation technique used here has proven to be a powerful and effective tool to produce a more detailed systematic description of the important regions of facial expressions. Despite some visible artifacts caused by this technique, accurate facial regions could be defined, resulting in flexible yet realistic animated faces.

While the present results provide some detailed insights, they are limited to a few expressions. In order to investigate facial expressions and their subtle movements more fully, we are currently recording a wider variety of facial expressions. Moreover, these expressions are being recorded at two intensity levels. This not only allows the investigation of the perception of intensity, but also the degree to which intensity affects which facial components make up any given expression. In addition, several new experimental methodologies are currently being tested, in order to more closely determine how expression recognition functions in everyday life situations. Furthermore, these approaches would allow us to test the generalizability of the results so far and might provide an indication how the computer animated avatar can be applied in future.

References

BASSILI, J. 1978. Facial motion in the perception of faces and the emotional expression. *Journal of Experimental Psychology 4*, 373379.

BLANZ, V., AND VETTER, T. 1999. A morphable model for the synthesis of 3d faces. In *Proceedings of the SIGGRAPH 1999 annual conference on Computer graphics*, 187–194.

BREIDT, M., WALLRAVEN, C., CUNNINGHAM, D., AND BÜLTHOFF, H. 2003. Combining 3d scans and motion capture for realistic facial animation. In *Proceedings der Eurograph, (Eds.) Julian and F. and P. Cano and The Eurographics Association*, 63–66.

CUNNINGHAM, D., KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. 2005. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception 2(3)*, 251-269.

CURIO, C., BREIDT, M., KLEINER, M., VUONG, Q., GIESE, M., AND BÜLTHOFF, H. 2006. Semantic 3d motion retargeting for facial animation. *New York ACM Press 2(3)*, 251-269.

DARWIN, C. 1872. *The Expression of Emotion in Man and Animals*. London, John Murray.

EKMAN, P., AND FRIESEN, W. 1978. Facial action coding system. *Consulting Psychologists Press*.

JOSHI, P., TIEN, W., DESBURN, M., AND PIGHIN, F. 2003. Learning controls for blend shape based realistic facial animation. In *Proceedings of the 2003 ACM SIGGRAPH/Eurythmics sumposium on Computer animation and ACM Press*, 187–192.

MARTIN, J., NIEWIADOMSKI, R., DEVILLERS, L., BUISINE, S., AND PELACHAUD, C. 2006. Multimodal complex emotions: gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics 3*, 269–291.

NUSSECK, M., CUNNINGHAM, D., WALLRAVEN, C., AND BLTHOFF, H. 2007. The contribution of different facial regions to the recognition of conversational expressions and (in preparation).

PARKE, F., AND WATERS, K. 1996. Computer facial animation. *Wellesly MA USA: A.K. Peters and Ltd*.

SCHWANINGER, A., WALLRAVEN, C., AND BÜLTHOFF, H. 2006. Computational modeling of face recognition based on psychophysical experiments. *Swiss journal of psychology (Swiss j. psychol.) ISSN*, 1421-018.

SIFAKIS, E., NEVEROV, I., AND FEDKIW, R. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. In *ACM Transactions on Graphics, SIGGRAPH Proceedings, TOG*, vol. 24, 417–425.

WALLRAVEN, C., BREIDT, M., CUNNINGHAM, D., AND BÜLTHOFF, H. 2005. Evaluating the perceptual realism of animated facial expressions. *ACM Transactions on Applied Perception*.

WATERS, K. 1987. A muscle model for animation three-dimensional facial expression. *ACM SIGGRAPH Computer Graphics 21*, 4, 17–24.

WILLIAMS, L. 1987. Performancedriven facial animation. *ACM SIGGRAPH Computer Graphics 24*, 4, 235–242.

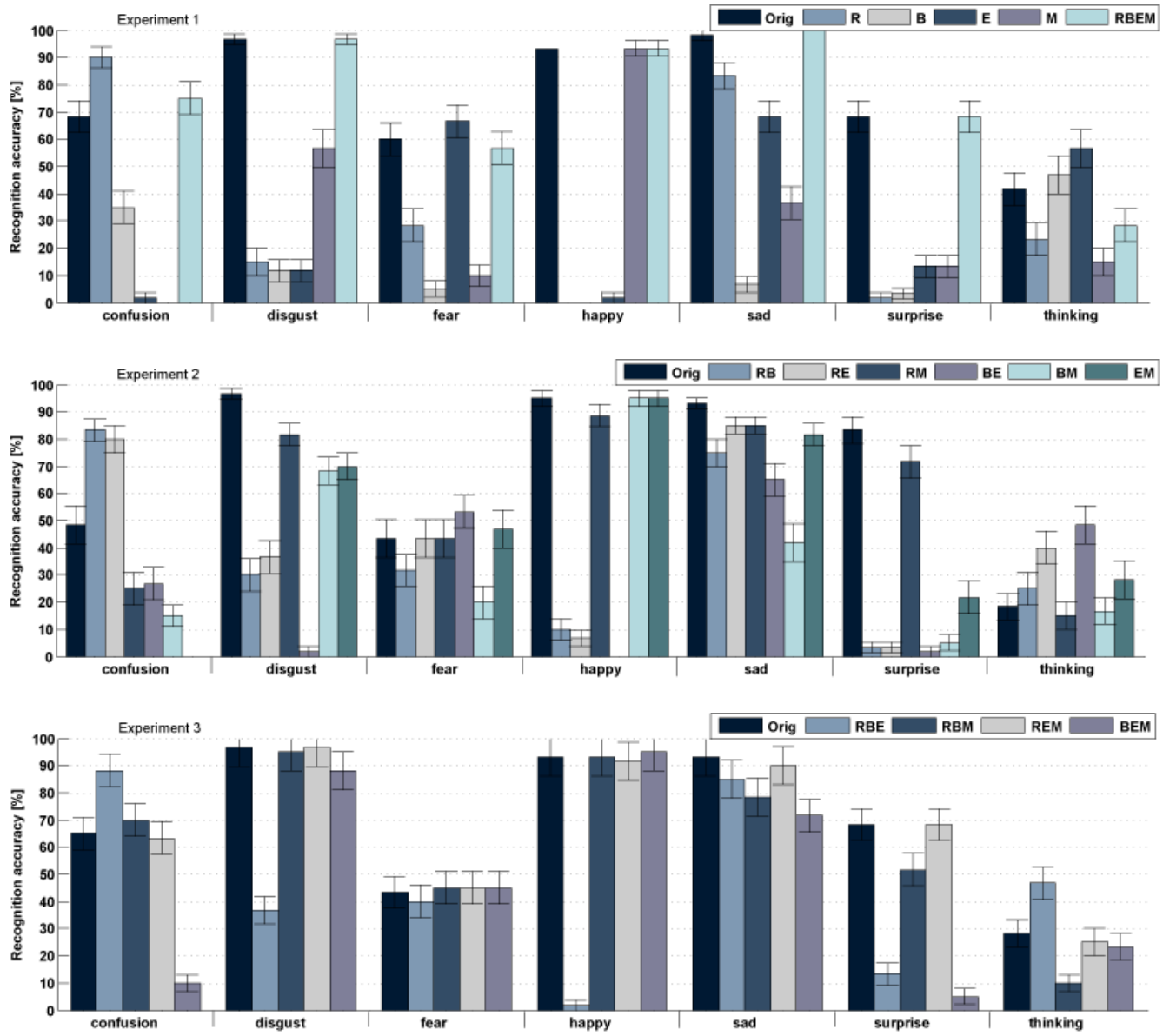

Figure 6: Recognition accuracy of the conditions in experiment 1,2 and 3 (see also color-plate).

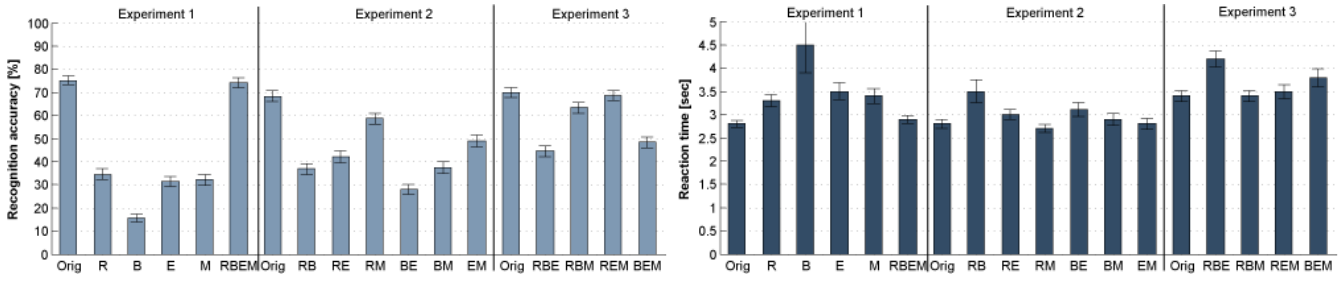

Figure 7: Recognition accuracy (left) and reaction times (right) of the conditions in experiment 1, 2 and 3 (see also color-plate).

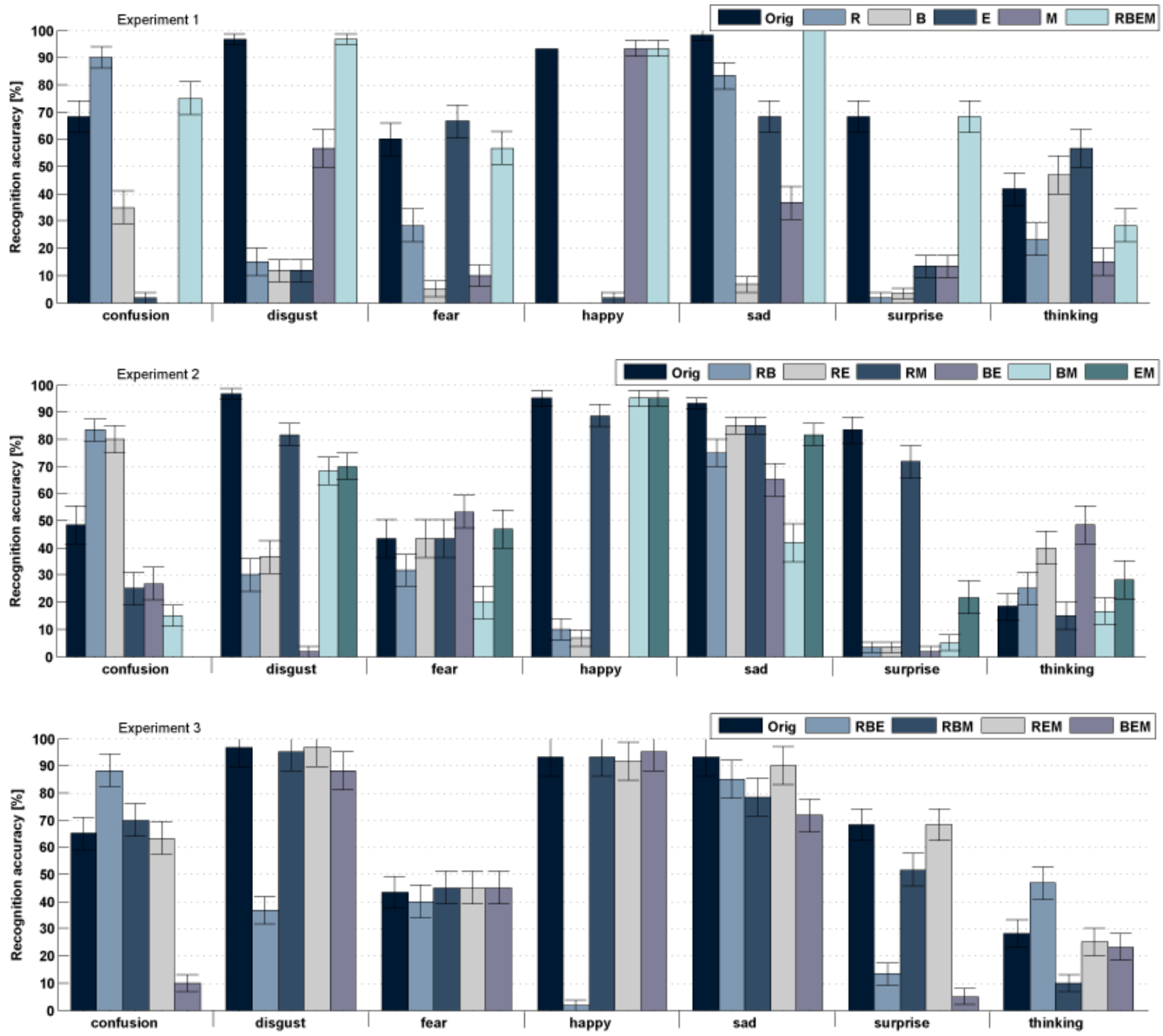

Figure 6: Recognition accuracy of the conditions in experiment 1,2 and 3 (see also color-plate).

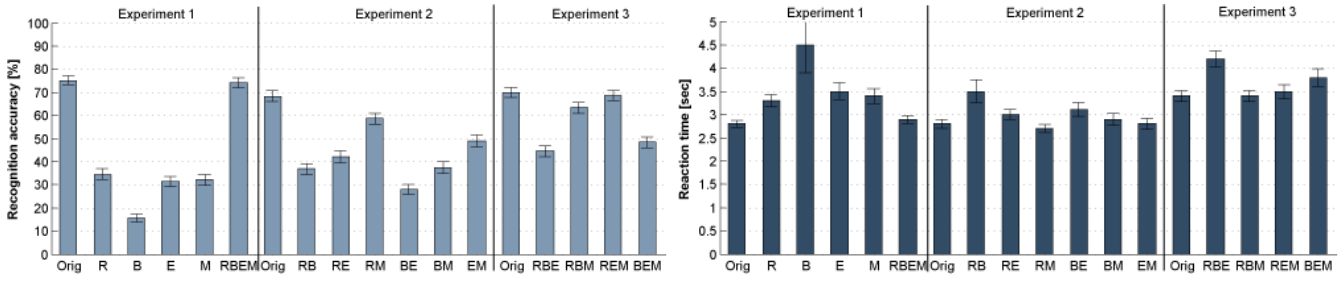

Figure 7: Recognition accuracy (left) and reaction times (right) of the conditions in experiment 1, 2 and 3 (see also color-plate).