Manipulating Video Sequences to Determine the Components of Conversational Facial Expressions

DOUGLAS W. CUNNINGHAM, MARIO KLEINER, CHRISTIAN WALLRAVEN, and HEINRICH H. BÜLTHOFF
Max Planck Institute for Biological Cybernetics

Communication plays a central role in everday life. During an average conversation, information is exchanged in a variety of ways, including through facial motion. Here, we employ a custom, model-based image manipulation technique to selectively "freeze" portions of a face in video recordings in order to determine the areas that are sufficient for proper recognition of nine conversational expressions. The results show that most expressions rely primarily on a single facial area to convey meaning with different expressions using different areas. The results also show that the combination of rigid head, eye, eyebrow, and mouth motions is sufficient to produce expressions that are as easy to recognize as the original, unmanipulated recordings. Finally, the results show that the manipulation technique introduced few perceptible artifacts into the altered video sequences. This fusion of psychophysics and computer graphics techniques provides not only fundamental insights into human perception and cognition, but also yields the basis for a systematic description of what needs to move in order to produce realistic, recognizable conversational facial animations.

Categories and Subject Descriptors: J.4 [**Computer Application**]: Social and Behavioural Sciences—*Psychology*; H.1.2 [**Models and Principles**]: User/Machine Sytems—*Human information processing*; H.5.1 [**Information Interfaces and Presentation (e.g., HCI)**]: Multimedia Information Systems—*Animations*; H.5.1 [**Information Interfaces and Presentation (e.g., HCI)**]: Multimedia Information Systems—*Evaluation/methodology*

General Terms: Experimentation

Additional Key Words and Phrases: Applied perception, facial expressions, human-computer interface, computer graphics, animation

1. INTRODUCTION

During a conversation, information is conveyed through a number of different channels, of which the face is one of the most versatile. Facial motions are very useful in modifying the meaning of what is being said [Bavelas and Chovil 2000; Bull and Connelly 1986; Condon and Ogston 1966; DeCarlo et al. 2002; Motley 1993]. For example, a spoken statement of surprise does not have quite the same meaning when it is accompanied by a look of boredom. Likewise, when producing certain forms of vocal emphasis or stress (e.g., like contrastive stress on the word "large" in the sentence: "No, not the *small* duck, I meant the *large* duck") the face moves to reflect this emphasis. This coemphasis can be seen, to some degree, in the static snapshot shown in Figure 1. The relationship between vocal stress and facial

This research was supported by the IST project COMIC (Conversational Multi-modal Interaction with Computers), IST-2002-32311. For more information about COMIC, please visit the web page (www.hcrc.ed.ac.uk/comic/).

Authors' Address: Douglas W. Cunningham, Mario Kleiner, Christian Wallraven, Heinrich H. Bülthoff, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany; email: {firstname.lastname}@tuebingen.mpg.de.

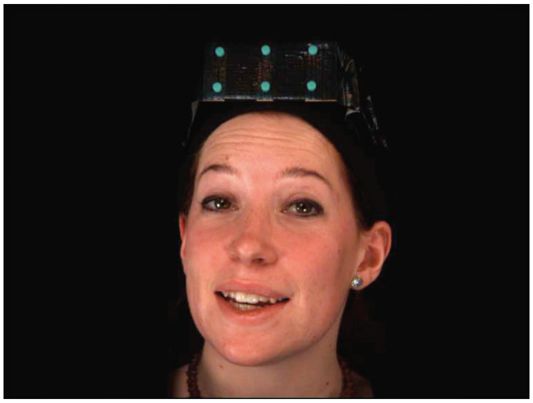


Fig. 1. Facial expression accompanying vocal emphasis. The facial expression that accompanies a vocal emphasis may be sufficient to recognize the emphatic nature of the statement, even in a static snapshot. Note that the actress is wearing a black hat with a tracking target containing six green dots. This target is used in subsequent tracking and model-based image manipulation (see Section 3.1.2.).

emphasis is so strong that it can be exceedingly difficult to produce the proper vocal stress patterns *without* producing the accompanying facial motion. The intimate connection between spoken meaning and facial motion has prompted some to suggest that the two signals together form the basic unit of meaning in speech, rather than providing independent contributions [Bavelas and Chovil 2000].

Facial motion may also be useful in controlling conversational flow [Bavelas et al. 1986; Bull 2001; Cassell and Thorisson 1999; Cassell et al. 2001; Poggio and Pelachaud 2000]. Perhaps the simplest version of this is the use of eye gaze to indicate to whom a request is directed (see, e.g., Cassell and Thorisson [1999]; Cassell et al. [2001]). The influence of eye gaze on conversations can be seen, for example, in the fact that improper or absent eye-gaze information is an oft cited problem with most video-conferencing technology [Isaacs and Tang 1993; Vertegaal 1997]. Facial control of conversational flow can also be quite subtle, especially with the use of "back-channel" responses [Bavelas et al. 2000; Yngve 1970], where a listener informs a speaker what needs to be said next through the judicious use of facial expressions. If a speaker is confronted with a nod of agreement, for example, they will probably continue talking. A look of confusion, disgust, or boredom, however, will almost certainly prompt very different behavior on the part of the speaker [Bavelas et al. 2000; Yngve 1970]. Bavelas et al. [2000] provided a persuasive demonstration of this. They examined storytellers and found that listeners seem to become an active part of the story, reacting as if they were in the situation being described. A lack of such sympathetic responses strongly affected the speaker: The story included less detail, did not last as long, and was often rated as less skillfully told.

Finally, facial motion can, of course, be used to independently express complex ideas and intentions. Emotional expressions (e.g., looking happy or sad) are obvious examples of this and a fair amount is known about the production and perception of the "universal" emotional expressions (happiness, sadness, fear, anger, disgust, contempt, and surprise, according to Ekman [1972]). Interestingly, few experiments have examined the nonaffective expressions which arise during a conversation. Since such "conversational" expressions are of considerable importance not only to daily conversation but also to the design of natural-language human–machine interfaces, the present work focuses primarily on them. Several important points regarding conversational expressions can already be made. First, humans often produce expressions during the course of normal conversations that are misunderstood, leading

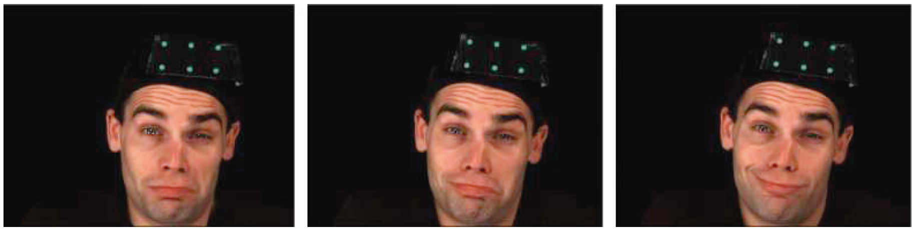


Fig. 2. Small physical changes can produce large perceptual changes. These three snapshots were taken from three separate recordings of a clueless expression. All three expressions were recorded within 5 min. of each other.

to miscommunication. Second, it is possible to produce an expression that is correctly recognized, but is perceived as contrived or insincere. As facial animation improves and is more often employed in the construction of human–machine interfaces (e.g., interface agents), the believability of the animation will likely become a very critical issue. Who would want to continue to talk with, let alone buy anything from, an agent if it is obviously lying or insincere, *regardless* of how good or realistic it looks?

Despite the importance of facial motions, and the frequency with which we use them, the synthesis of proper conversational expressions (e.g., for the production of avatars, interface agents, virtual humans) remains extremely challenging. One reason for this is that humans are amazingly good at recognizing facial expressions and can detect very small differences in both motion and meaning. A second reason can be found in the subject matter itself: The physical differences between an expression that is recognizable and one that is not can be very subtle (see, for example, Figure 2). Moreover, there are a number of different ways that humans express any given meaning and not all of the resulting expressions are easily recognized [Cunningham et al. 2003a, 2003b]. Thus, even if a physically accurate Virtual Human perfectly duplicates all spatial and temporal aspects of facial motion and is driven in real-time from a real human face, there is still no guarantee that the resulting expressions will be understood, let alone be seen as believable.

A systematic description of the necessary and sufficient components of recognizable and believable conversational expressions would, then, prove very helpful in the generation of facial animation. Given the importance of facial expressions, it should not be surprising that they have been the subject of intense study. There is a large literature in computer vision examining a variety of methods for automatically extracting and/or recognizing facial expressions (see Donato et al. [1999] and Pantic and Rothkrantz [2000] for reviews), although many of these methods are not specifically interested in mimicking human perceptual capabilities. There is also a large literature covering facial expressions in computer graphics and in the behavioral sciences (see, e.g., Pelachaud et al. [1994] for a review). Within these fields, an impressive variety of representational systems have been developed to describe facial expressions (see, e.g., Sayette et al. [2001]). Perhaps the most widely used system for describing facial expressions is the Facial Action Coding System (or FACS; Ekman and Friesen [1978]), which segments the visible effects of facial muscle activation into “action units.” Combinations of these action units might then be used to describe different expressions. It is important to note that FACS was designed as a *descriptive system* for representing the elements of facial expressions. Thus, a detailed analysis of which elements actually go together to produce different expressions is external to FACS itself [Sayette et al. 2001].

Most of the descriptive systems developed for facial expressions, including FACS, focus explicitly on static information [Essa and Pentland 1994]. That is, they are primarily concerned with information that is visible at any given instant in time. There is a growing body of evidence, however, that temporal information is of central importance to the perception and recognition of facial expressions [Bassili

1978, 1979; Bruce 1988; Edwards 1998; Kamachi et al. 2001]. Moreover, static and dynamic information for expressions seem to be processed in separate areas of the human brain [Humphreys et al. 1993], suggesting that the temporal information is not merely processed as a collection of static snapshots. Thus, any description of the perceptually necessary and sufficient components of facial expressions should include an examination of the temporal aspects of expressions as well as the static aspects.

Finally, a distinction may be drawn between the motion that *is* present in a *specific recording* of an expression and the motion that *must be* present for that expression to be recognized. The complexity of human faces, the subtle but important differences between different faces, and the range of motion faces exhibit jointly make the recognition of facial expressions one of the most difficult tasks the human visual system can perform [Barton 2003]. In many cases, however, it is neither appropriate nor desired to synthesize all of the myriad details of real facial motion. For example, some have argued that the realism and level of detail of an interface agent should match the capabilities of the human–machine interface [Thorisson 1996]. Likewise, the explicit goal of nonphotorealistic animation is to present abstract, yet nonetheless recognizable, versions of various events and actions [Gooch and Gooch 2001; Strothotte and Schlechtweg 2002]. Thus, in contrast to much of the previous research on facial expressions, we will focus here on *empirically* determining what facial regions *must move* in order for a *conversational* facial expression to be recognized. Providing an empirical basis for such descriptions is, of course, not easy. One reason for this is the lack of stimuli: It is exceedingly difficult and time-consuming to manually alter subregions of a face throughout entire video sequences. Such variations in the presence of different types of motion within a single expressions are required in order to determine the necessity and sufficiency of different facial motion. Here, we present an advanced computer graphics technique that semiautomatically manipulates video recordings of real expressions. The resulting manipulated sequences were used along with the original sequences in two psychophysical experiments, which validate the effectiveness of the technique and provide some initial insights into the components of conversational facial expressions.

The nine conversational expressions that were used in the present work were: agree, disagree, happy/pleased, sad, thinking, confused (i.e., the actor or actress did not understand a question), clueless (i.e., the actor or actress did not know the answer to a question), disgust, and pleasantly surprised (see Figure 3). The nine expressions were recorded, as part of our video database of facial expressions, from six individuals with six synchronized digital video cameras (Section 2.1.1). All the recordings were shown to a number of individuals, who were asked to identify and rate the believability of the expressions (Section 2). In the second experiment (Section 3), the best recordings of each expression for each actor/actress was selected, based on the results of Experiment 1, and then manipulated so that the interior of the face, with the exception of select facial regions, was “frozen” (i.e., replaced with a static snapshot; see Section 3.1.2). In different experimental conditions, different regions were left intact, enabling us to examine the importance of those regions for various facial expressions. The regions examined were the eyes (which included direction of gaze and blinking information), eye and eyebrow region, and the mouth region (see Figure 4). In all conditions, the rigid head motion was left intact.

2. EXPERIMENT 1

The nine conversational facial expressions were recorded from six individuals using a protocol based on method acting. More specifically, a brief scenario was described in detail and the actor or actress was asked to place him/herself in that situation and then react accordingly. A single “take” of an expression consisted of three repetitions of that expression performed in rapid succession, with a neutral expression preceding and following each repetition. The use of multiple repetitions that rapidly followed each other helped the actors and actresses place themselves more deeply into the described scene, which might improve the quality of the expressions. While at least one take per actor/actress was recorded from

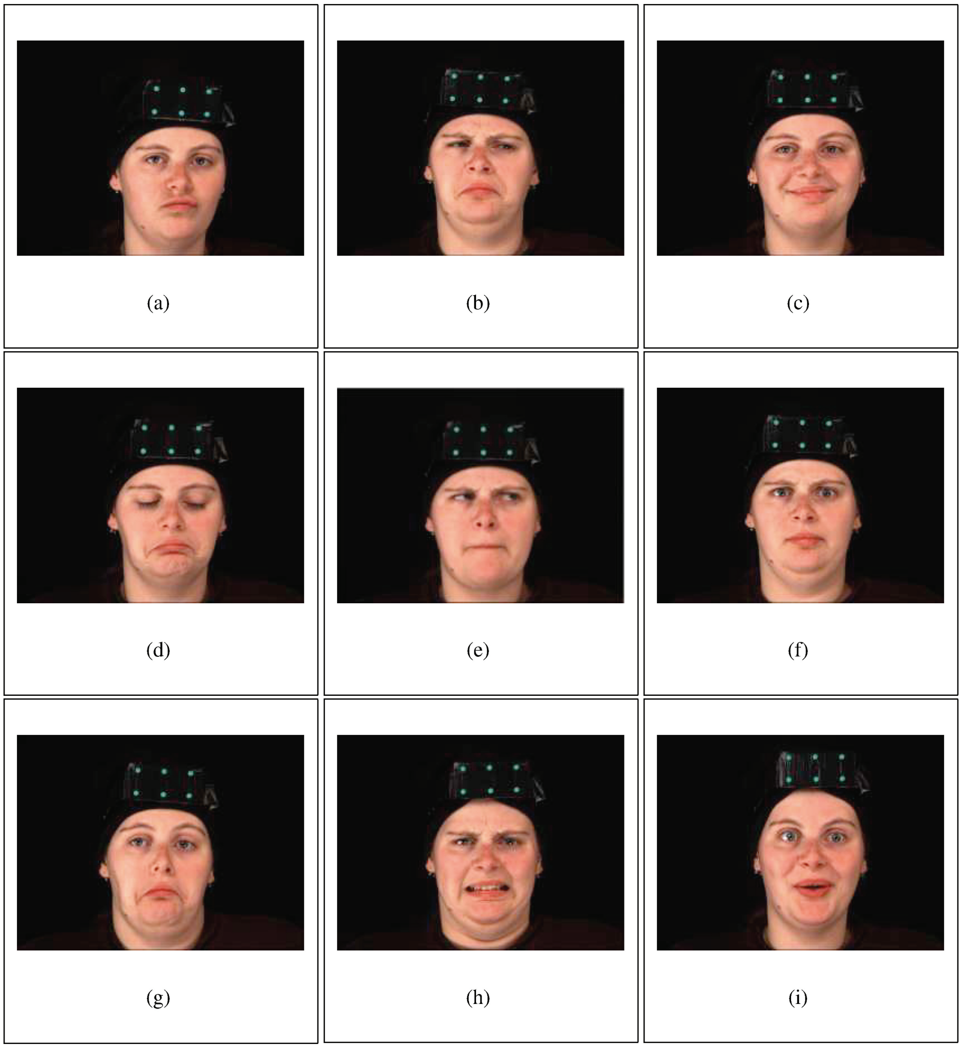


Fig. 3. Snapshots from the nine different expressions. (a) Agreement; (b) disagreement; (c) happiness; (d) sadness; (e) thinking; (f) confusion; (g) cluelessness; (h) disgust; (i) surprise. Some of these expressions can be recognized even in a static snapshot. Other expressions, like agreement and disagreement, seem to rely more heavily on dynamic information.

each of the nine expressions, some difficult expressions had more than one take recorded. In total, 213 individual repetitions were recorded. Naturally, presenting all 213 repetitions in all 6 manipulated versions (see Experiment 2) would produce an uncomfortably long experiment. To keep Experiment 2 at a manageable size, a single repetition for each expression for each actor/actress needs to be selected. To provide an objective basis for selecting the sequences, all 213 recordings were shown to a number of individuals (hereafter referred to as "Participants") in Experiment 1. For each expression from each actor/actress, the repetition that was recognized most often was chosen for use in Experiment 2.

A second goal of the first experiment was to examine the influence of motion context. More specifically, in the original recordings, the actors/actresses sometimes held an expression for an excessively long time, which might make the expression look somewhat unbelievable. Likewise, the return to neutral

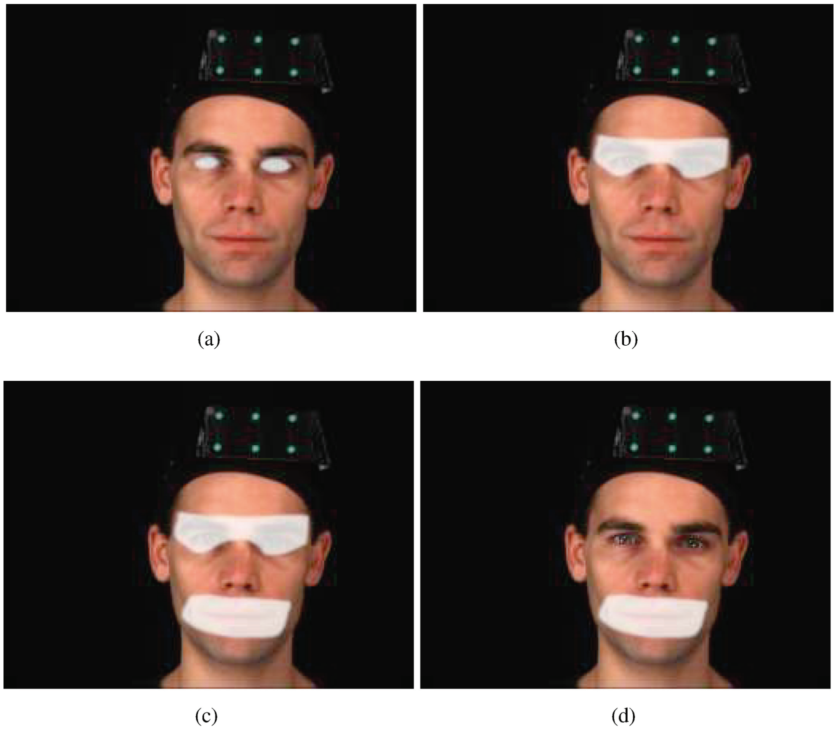


Fig. 4. Sketch of the areas with motion for four of the experimental conditions. A static snapshot is shown here with the area that was allowed to move highlighted in white (the original condition is not shown, since everything was allowed to move and the "rigid head-only" condition is not shown, since no facial areas were allowed to move). This white highlighting is shown here for explanatory purposes only and was never shown to the participants. (a) Condition RwE: This condition had rigid head motion as well as motion of the eyeballs; (b) Condition RwEB: This condition had rigid head, eye region, and eyebrow motion; (c) Condition RwEMB: The head, eyes, eyebrows, and mouth could move; (d) Condition RwM: Only rigid head and mouth region motion were present. An example video sequence can be found at http://www.kyb.mpg.de/˜kleinerm/facefx/.

in some of the recordings is rather abrupt and unexpected, potentially making it more apparent that the sequences were posed. Instead of showing the full video sequence (i.e., from a neutral expression through the peak expression and back to neutral), one might clip it so that it stops after an appropriate amount of time (i.e., so that the sequence merely goes from a neutral expression to the peak expression). Clipping the sequences would remove any potential influence of the relaxation phase. Of course, clipping a sequence has its own potential disadvantages (e.g., the subjectivity of defining the new ending point of the sequence). In Experiment 1, the full sequences and the clipped sequences were presented to two separate groups of individuals, respectively.

2.1 Methods

2.1.1 *Recording Equipment.* The facial expressions were recorded using the Max Planck Institute for Biological Cybernetic's VideoLab (see Kleiner et al. [2004] for a detailed description). The VideoLab setup has six recording units, each of which consists of a digital video camera, a frame grabber, and a computer. Each unit can record up to 60 frames/s of fully synchronized, noninterlaced, uncompressed video in PAL resolution (768 × 576 pixels).

The present recordings were made at 25 frames/s. The exposure time of each camera was set to 3 ms in order to reduce motion blur. At the start of each recording session, an actor or actress was seated in front of the cameras (which were arranged in a semicircle around the actor/actress at a distance of 1.5m) and a homogeneous, black background was placed behind the actor/actress. The actors and actresses were asked to wear a black shawl to hide their shoulders and torso. They also wore a black hat which, in addition to hiding their hair, had a tracking target with six green dots. The tracking target was used in the subsequent manipulation of the image sequences (see Section 3.1.2). To help avoid artifacts and unintended information in the recorded sequences, care was taken to light the actor's and actress's faces as flatly as possible. Special effort was devoted to the avoidance of directional lighting effects (cast shadows, highlights).

2.1.2 *Stimuli.* The expressions were recorded from six different people (three male and three female). One of the individuals (actor 1) was a professional actor. The remaining individuals were amateur actors and actresses. As previously described, the expressions were elicited using a protocol based on method acting. During the recordings, the actors and actresses were free to move in any way they felt appropriate, but were asked to refrain from placing their hands in front of their faces. They were also asked to try not to speak during the expressions, unless they felt that speech was absolutely required to react naturally. Sound was neither recorded nor played during the experiment.

For each of the expressions, at least three repetitions were performed in succession with a neutral expression both preceding and succeeding each repetition. Each of these recordings was split into its constituent three repetitions, which were then edited so that each video sequence began on the frame after the face began to move away from the neutral expression. For the "full group," the video sequences were edited so that they ended on the first frame after the face stopped moving (i.e., after the return to a neutral expression). For the "clipped group," the video sequences were edited so that they ended after reaching the peak of the expression. The resulting video sequences varied considerably in length. There was no apparent simple correlation between expression and duration.

2.1.3 *Psychophysical Methods.* Eighteen individuals participated in the experiment. These participants were randomly assigned to one of the two groups (full or clipped), so that a total of nine participants were in each group. For the present experiment, the size of the images was reduced to 256 × 192 pixels and the participants sat at a distance of approximately 0.5 m from the computer screen (i.e., the images subtended approximately $10 \times 7.5^\circ$ of visual angle).

The order in which the 213 trials were presented was completely randomized for each participant. For both groups, a single trial consisted of a video sequence being repeatedly shown (without sound) in the center of the screen. A 200-ms blank screen was inserted between the repetitions of the video sequence. When participants were ready to respond (which they indicated by pressing the space bar), the video sequence was removed from the computer screen, and the participants had to perform two tasks.

The first task was to identify the expression. This was done by selecting the name of the expression from a list that was displayed on the side of the screen. The list of choices included all nine expressions as well as "none of the above." Since some of the participants were native German speakers and others were not, the expressions were listed in English as well as German (the German names used for the expressions were: zustimmen, nicht zustimmen, glücklich / zufrieden, traurig, nachdenklich, unwissend ["weiss nicht"], verwirrt ["versteht nicht"], angeekelt / abgestossen, and angenehm überrascht).

One might be worried that this type of task (i.e., a nonforced choice task) does not properly reflect identification performance. Frank and Stennett [2001] have shown, however, that this type of task is highly correlated with other identification procedures (e.g., free description of the expressions). Moreover, the nonforced choice methodology offers some advantages over other methodologies, including the avoidance of the inflated accuracy ratings found in the absence of a "none of the above" option (previous work with nonforced choice tasks has shown that people do, in fact, take advantage of the "none of the

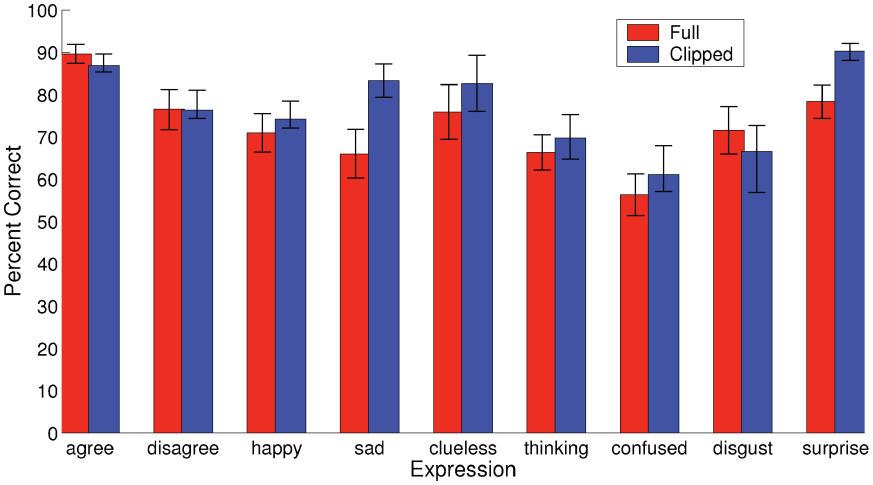


Fig. 5. Overall recognition accuracy. The percentage of the time that each type of expression was correctly identified is shown for each of the two groups. The error bars represent the standard error (SE) of the mean.

above" option; see, e.g., Cunningham et al. [2003a, 2003b, 2004]) and avoiding the subjectivity found when experimenters must categorize and analyze free description results.

The second task was to rate the believability of the expressions. The participants were to enter a value from 1 to 7 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.

2.2 Results and Discussion

Overall, the results are very similar to those found in other experiments using a subset of this database [Cunningham et al. 2003b, 2004] as well as recordings from nontrained individuals [Cunningham et al. 2003a]. In general, the expressions were well recognized (albeit imperfectly—the average recognition performance was 74.6%) and were seen as relatively believable (the average believability rating was 4.91). Clearly, the visual information present in the video sequences is sufficient to specify these conversational expressions.

Interestingly, there were few performance differences between the full and clipped groups, indicating that the motions that occur after the peak of an expression do not play a role in the two tasks used here. Where there are differences (as is the case for the sad and surprise expressions), the clipped sequences have higher recognition rates. Given that the clipped sequences are marginally better (i.e., more recognizable) and noticeably shorter, they will be used in Experiment 2. For each actor/actress, the repetition that was correctly recognized most often was chosen for use in Experiment 2 (in cases where two or more repetitions of an expression from an actor/actress had identical recognition accuracies, the repetition with the highest believability rating was chosen).

2.2.1 *Recognition Accuracy*. Overall, the expressions were recognized well, with little difference between the full and clipped expressions (the average recognition rates were 72.4 and 76.8%, respectively). Consistent with previous research, the participants responded with "none of the above" on 6.6% of the trials, on average (the highest rate was for the confused expressions, where 12.8% of the responses were "none of the above").

For both types of sequence, the recognition rates varied considerably across the different expressions (see Figure 5). The error bars in Figure 5 show that there is also a considerable amount of variation in recognition rates within each expression type. That is, some versions of any given expression were more recognizable than other versions of that expression.

This variation in quality within expressions is also true for the individual actors and actresses (see Figure 6). This variation is significant for several reasons. First, no single individual produced recognizable versions of all expressions. For example, actor 5 produced a version of confusion that was recognized by nearly everyone, but very few people recognized that person's sad expression. Second, each actor/actress showed a high degree of variability for each expression. In other words, regardless of the average quality of any given expression, each actor/actress sometimes produced decent versions of that expression and sometimes produced less recognizable versions (even though the different versions were generally produced within 2m of each other). Jointly, these observations strongly suggest that if one produces facial animations that are perfect copies of any given individual, at least one of the resulting expressions would be unrecognizable (and the quality of the rest will vary).

It is critical at this point to note that for each expression at least one actor/actress produced a version that was recognized by most people. In other words, each of the nine expressions can, in principle, be fully specified by visual motion information. This suggests that one way to produce an interface agent whose expressions are always recognized is to combine expressions from different individuals.

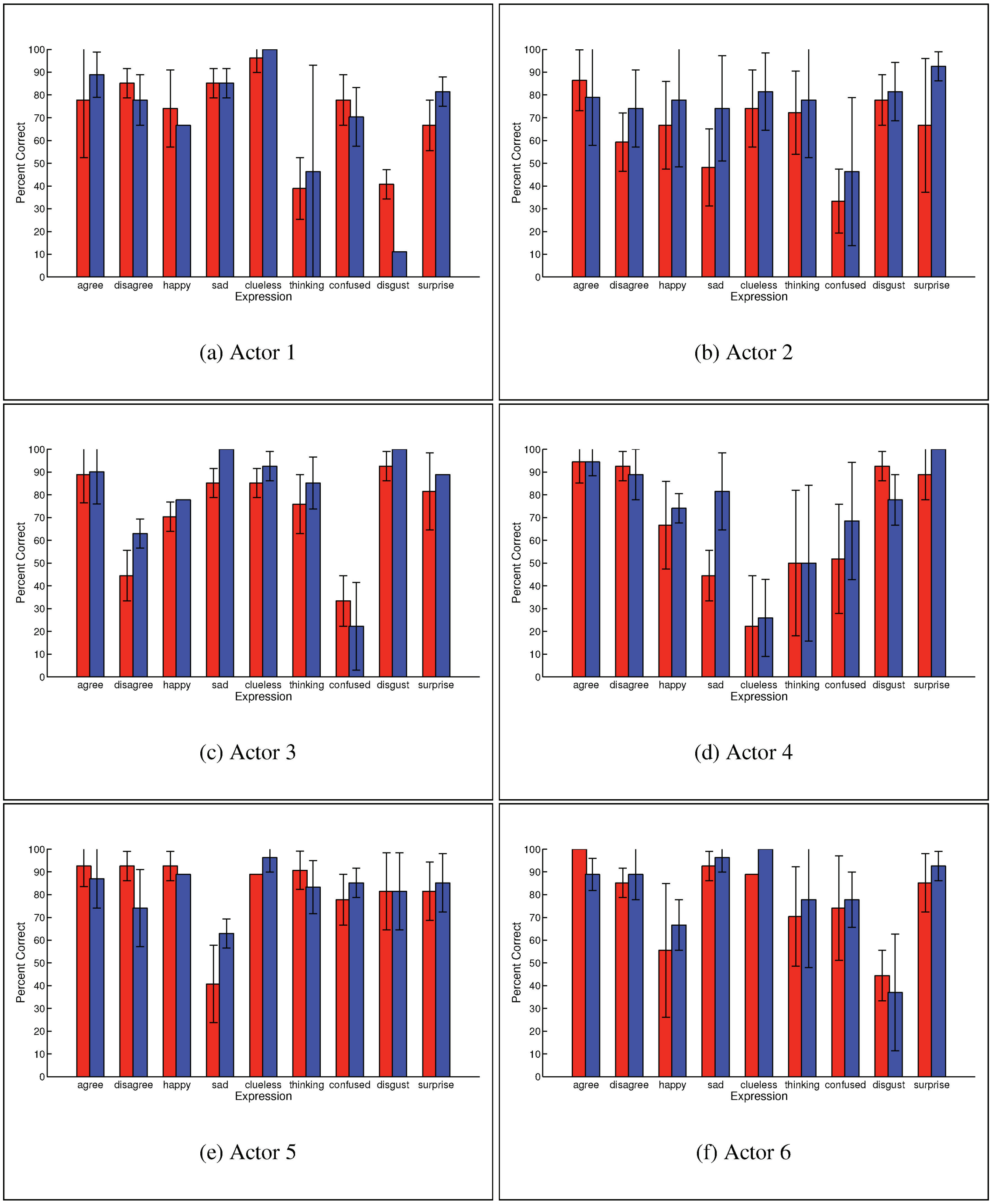
Finally, there was no systematic difference in either recognition performance or believability ratings between the professional actor's expressions and the amateur actors' and actress' expressions (see Figure 6). Furthermore, the data from the present experiment are remarkably similar to those reported in Cunningham et al. [2003a] (where the expressions were recognized 80% of the time), even though the expressions used in that study were elicited with a simpler protocol and untrained individuals were used. Taken together, these results seem to suggest that level of training does not have a strong influence on the recognizability of conversational facial expressions.

2.2.2 *Believability Ratings.* As has been previously found with these and other expressions [Cunningham et al. 2003a, 2003b], the expressions were seen as somewhat believable, but not completely convincing (see Figure 7). Also as previously found, there was little difference between the believability of expressions that were correctly recognized and those that were not.

Figure 7, which plots the believability ratings for the trials where the expression was correctly recognized, shows that there is little systematic difference between the believability of full and clipped expressions (on average, the ratings were 4.94 and 4.88, respectively). The two exceptions to this trend are sadness and thinking. These two expressions were seen as slightly more believable, on average, in the full than in the clipped condition. Since these results come from those trials that were correctly identified, one potential explanation for this might lie in the recognition rates. Indeed, as was mentioned above, sadness was recognized less often in the full than in the clipped group. It is possible, then, that whatever is present in the full condition but removed from the clipped sequences (i.e., the end of the expression) led participants in the full group to decide that the expression was not really sadness. This would mean that, in the full condition, only the really convincing sad and thinking expressions would be left. Additional experiments are needed to fully examine the potential connection between recognition and believability. Since Experiment 2 focuses primarily on the recognition of expressions, the clipped sequences will be used.

3. EXPERIMENT 2

The results of Experiment 1 clearly showed that the conversational facial expressions of unknown individuals can be recognized even without a conversational context. In other words, the visual information contained within the video sequences is, in principle, sufficient to recognize this set of conversational expressions. The results of Experiment 1 also showed, however, that every individual produced at least



(a) Actor 1

(b) Actor 2

(c) Actor 3

(d) Actor 4

(e) Actor 5

(f) Actor 6

Fig. 6. Expression recognition accuracy for the six actors and actresses. The percentage of the time that each expression was correctly recognized is shown for each of the two groups of participants. The results of the full group are shown in red and the clipped group in blue.

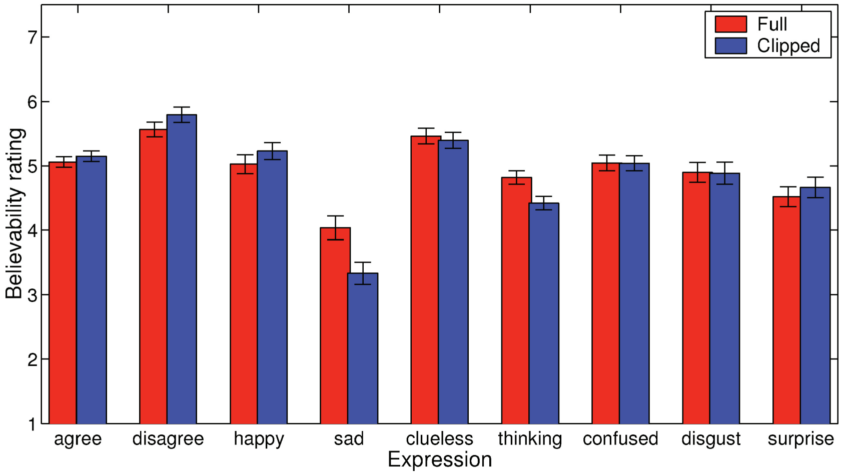


Fig. 7. Average believability ratings for trials where the expressions was correctly identified. A rating of 1 means that the expression was not believable (i.e., the actor or actress was clearly pretending), while a rating of 7 indicates that the expression was seen as genuine.

one expression that was not well recognized. Even a quick glance at the video sequences shows that there are a lot of differences in both structure and motion between those instances of an expression that were recognized and those that were not. Which of the many differences determine whether the expression was recognized or not? What are the *core components* of recognizable, believable facial expressions?

To begin to answer this question, the best expressions from Experiment 1 were manipulated to systematically alter the motion information available in the video sequences. More specifically, the entire face was "frozen" (i.e., replaced with a static snapshot from a neutral expression) and then selective areas of the face were "unfrozen." In this manner we can roughly determine which areas of the face *need* to move in order for an expression to be recognized. For example, in the most extreme condition, the entire face is frozen, leaving only rigid head motion (i.e., translations and rotations of the entire head). If an expression can still be recognized in this condition, then clearly rigid head motion is *sufficient* for the recognition of that expression. For example, both agreement and disagreement can probably be fully specified by nodding or shaking one's head. If an expression cannot be recognized when only rigid head motion is present, then some other facial motion must be necessary (as is probably the case for smiling, for example). Here, we use the image manipulation procedure to examine the necessity and sufficiency of the four types of motion that are most commonly used in facial animation: Rigid head motion, eye motion, eyebrow motion, and mouth motion (see Section 3.1.1 for more information).

3.1 Methods

Experiment 2 is similar to Experiment 1 with several critical differences. In addition to introducing manipulated video sequences (see Sections 3.1.1 and 3.1.2), a new task was added (rating of naturalness; see Section 3.1.3).

3.1.1 *Stimuli.* In addition to the original video footage, five "freeze-face" conditions were shown. To produce the freeze-face sequences, each of the original recordings was subjected to postprocessing (see Section 3.1.2). The postprocessing resulted in video sequences that were nearly identical to the original. The primary difference between the manipulated sequences and the original was that all of the face except select regions was replaced with a static snapshot (rigid head motion was left intact in all conditions). The static snapshot used in freezing the face was from a neutral expression. For each actor or actress, the same snapshot was used to produce all of their frozen expressions, ensuring that the frozen regions carried no expression-specific information. In the first condition (Rigid Only), all of the face was held still. In the second condition (RwE), both the original rigid head motion and the motion of the eyeballs were present (see Figure 4). In the third condition (RwEB), rigid head motion, eye motion, and the region around the eyes and eyebrows (but not the forehead) were present. In the fourth condition (RwEBM), motion of the mouth region was added. Finally, the fifth condition (RwM) contained only rigid head and mouth motion.

3.1.2 *Image Manipulation Technique.* Each of the individuals who were recorded wore a black hat with a tracking target (a black rectangular plate with six green markers; see, e.g., Figure 1). Since the six cameras in the VideoLab are fully synchronized, we are able to utilize a custom, image-based, stereo motion-tracking algorithm to recover the three-dimensional (3D) location of the tracking target (using a single stereo camera pair). The first step in the 3D motion tracking is to find the location of the six green markers in each of the stereo pair images. The algorithm then tracks the image positions of corresponding markers in both images to recover the 3D spatial position of the markers via stereo triangulation. Finally, the algorithm fits a geometric model of the tracking target to the 3D point cloud of markers, thereby recovering position and orientation of the tracking target in space. In order to determine the relative location of the markers to the individual's head, a 3D model of that individual's head is needed. The 3D models for the present experiment were acquired with a Cyberware 3D laser range scanner, and consisted of a 3D polygon mesh of approximately 150,000 triangles, defined by 75,972 vertices with a spatial resolution of approximately 0.1 mm. The 3D scans also yielded a texture map (512 × 512 texels) of the individual. To ensure that the elements of the meshes and texture maps of different models always define the same facial region, all of the models are brought into correspondence with each other (the process is described in detail in Blanz and Vetter [1999]). Since there is a fixed spatial relationship between the rigid motion of the head and the motion of the tracking target,[1] the recovered position and orientation of the target is used to position and orient the 3D shape model of the individual's head. This establishes a point-to-point correspondence between texels in the texture map of the model and image pixels in the video footage.

To selectively freeze parts of the face, the 3D head model was superimposed onto the video footage ("video rewrite"). In face regions where we wished to leave the original recording intact, the corresponding parts of the model mesh were rendered with an alpha value of zero (i.e., the model was fully transparent and, therefore invisible, in these regions). In regions where we wanted to freeze the face the model was rendered with an alpha value of 1.0 (fully opaque) using one of the previously extracted texture maps. Since all of the 3D head models are in correspondence, we were able to define specific facial regions with a single texture mask (specifying the transparency levels) per experimental condition and then apply these manipulations to all the recordings. This greatly reduced the amount of manual work involved in manipulation of a large number of sequences.

3.1.3 *Psychophysical Methods.* Participants in this experiment were asked to perform three tasks. The first two tasks were identical to the two tasks in Experiment 1: Recognize the expression and rate its believability. For the third task, participants were asked to rate the naturalness of the expression. Specifically, participants were asked to indicate if what they just saw is something that people normally do. This task also used a 7-point scale, with a rating of 1 representing expressions or motions that are not natural and a rating of 7 representing motions that are natural. The participants were specifically

[1] The relationship between the target and the head is set up by manual interactive initialization on the first frame of each recorded sequence.

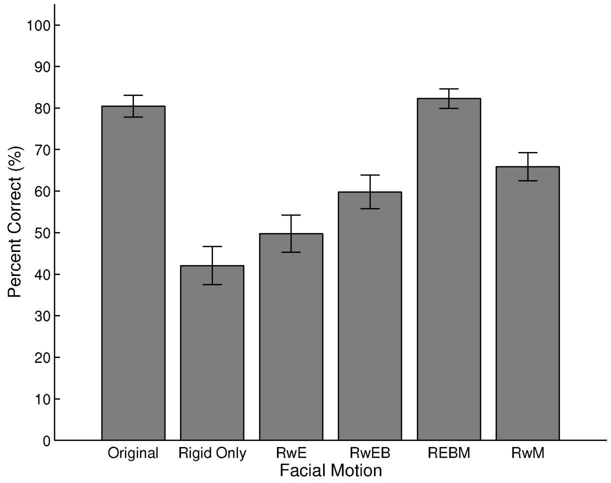


Fig. 8. Overall recognition accuracy. The percentage of the time that the participants correctly identified the expressions is shown for the six freeze-face conditions.

asked to rate as unnatural any expressions that contained noticeable artifacts from the manipulation techniques.

As in Experiment 1, the participants sat at a distance of approximately 0.5 m from the computer screen. Cunningham et al. [2004] have shown that altering the size of the images does not alter recognition of unmanipulated versions of these conversational expression as long as the face subtends at least 1.4° of visual angle. It is possible, however, that since the more subtle facial changes (e.g., the wrinkles on the nose and forehead in a confused expression) are more noticeable at the larger image size, their absence in the freeze-face conditions will be more noticeable. Thus, in Experiment 2, we used an image size of 512 × 384 pixels (approximately 20 × 15° of visual angle).

All of the 324 video sequences (nine expressions crossed with six actors and six freeze-face conditions) were shown to nine new participants in a psychophysical experiment. Since the experiment lasted approximately 2.5 h, participants were given the opportunity to take a break every 40 trials. As in the first experiment, the order in which the trials were presented was completely randomized for each participant. Due to technical difficulties, the data from one participant were not complete and were not included in the analyses. The data from a second participant were not included since the participant failed to follow the instructions for the experiment.

3.2 Results and Discussion

Overall, the participants were very good at identifying the expressions (80.4% in the original condition) even though they did not know the actors and actresses and had no conversational context. Recognition accuracy and the patterns of confusion that the participants made in the original condition were very similar to previous work with these expressions from these actors and actresses [Cunningham et al. 2003b] and with these expressions from other, nontrained individuals [Cunningham et al. 2003a]. The believability ratings (4.93, on average, for the correctly identified trials) were also similar to previous results. Finally, the participants found all of the expressions to be somewhat natural in all conditions (the naturalness rating was 4.86, on average, for the correctly identified trials).

3.2.1 *Recognition Accuracy.* Figure 8 shows, for each of the six freeze-face conditions, the percentage of trials where the expressions were correctly identified. One of the most interesting findings is that, on

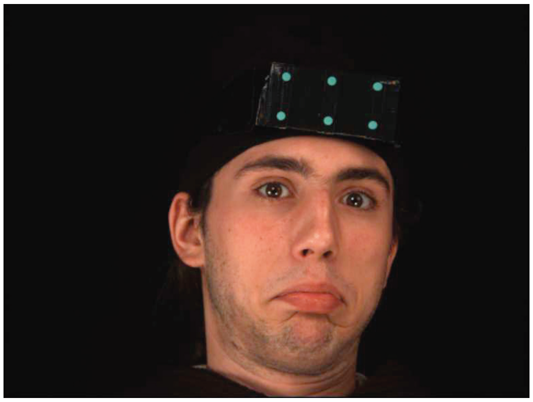
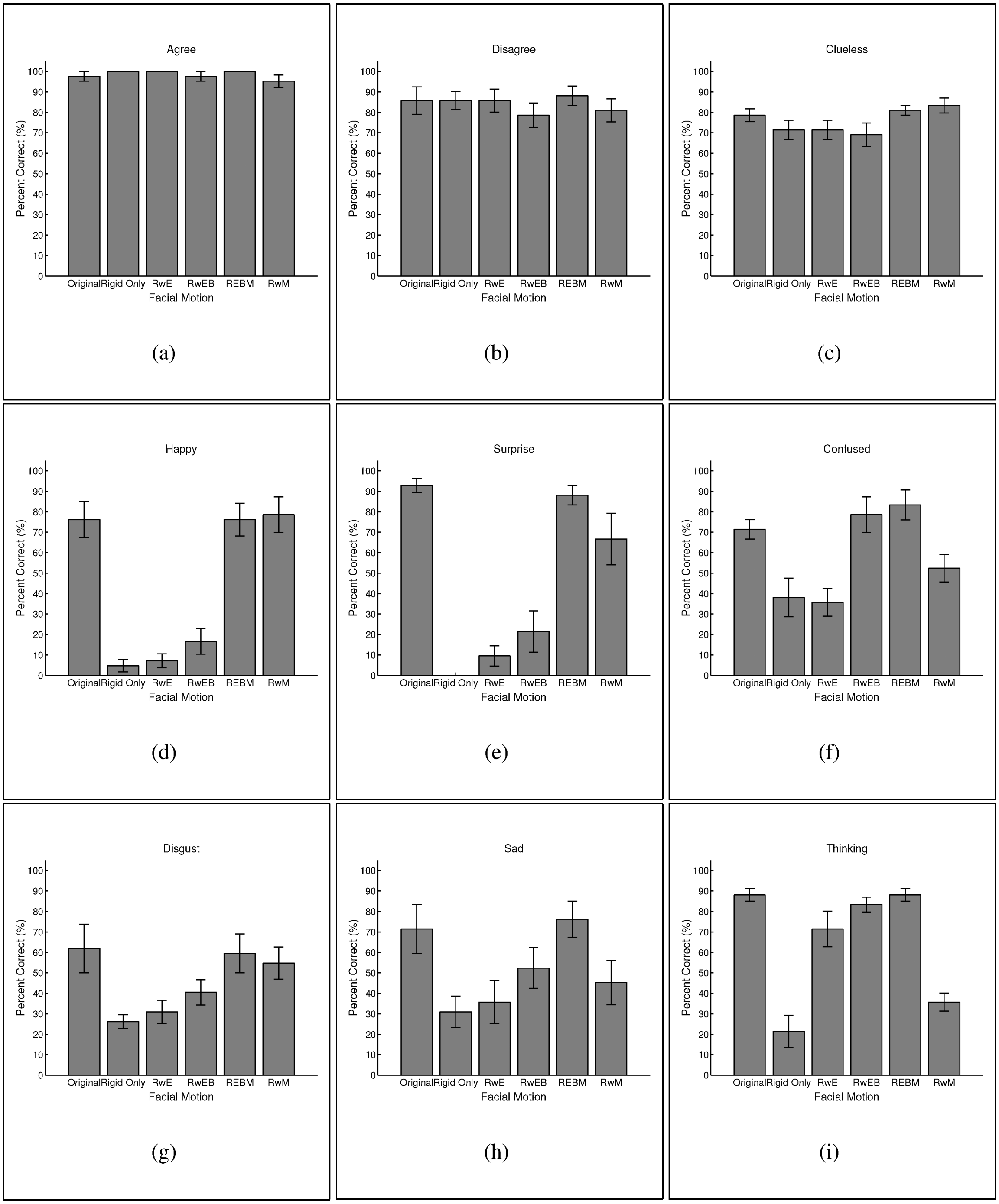


Fig. 9. A single snapshot from one actor's clueless expression. The upward motion of the chin and the down-turned corners of the mouth seen here are typical for an expression where one wishes to say "I don't know."

average, rigid head motion carries a fair amount of information about the expressions. That is, when only rigid head motion is present, participants could still identify the expressions better than would be expected if the participants were blindly guessing. On average, the addition of eye motion (condition RwE) tended to improve recognition accuracy somewhat and the further addition of eyebrow motion (condition RwEB) tended to improve recognition accuracy even more. The mouth seems to carry a considerable amount of information (condition RwM). In short, each of the four types of motion carry some information about the expressions all four together seem to carry enough information to accurately identify all the expressions used here.

As can be seen in Figure 10, different expressions rely on different types of motion to convey their meaning, with most expressions relying primarily on a single region. Agreement and disagreement, for example, are fully specified by rigid head motion. Interestingly, every participant recognized every version of agreement from every actor and actress. This is a higher recognition rate than in the original sequences. It is possible that, in the original sequences, the motion of internal facial regions specifies variants of agreement (e.g., reluctant agreement, enthusiastic agreement, considered agreement) causing participants to occasionally label the original agreement expressions as "none of the above." The rigid head motion alone condition, on the other hand, seems to specify pure agreement. Since the accuracies in the original are already near 100%, however, any improvement of the rigid head-only condition over the original conditions can not be fully measure with the current task.

The apparent sufficiency of rigid head motion for the accurate recognition of cluelessness is a bit surprising, especially since all of the actors and actresses had characteristic mouth motions (see Figure 9): The chin moved upward, the lower lip was pushed out, and the corners of the mouth moved downward. Moreover, careful examination of the video sequences did not reveal any obvious, consistent pattern of rigid head motion for the clueless expressions. One potential explanation for why rigid head motion seems to be sufficient to identify cluelessness is that, although the shoulders were masked with a black shawl, the motion of the shoulders and the effect of such motion on the neck were still visible. Since all of the actors and actresses shrugged their shoulders, it is likely that this motion, and not the rigid head motion, is what participants used to recognize an expression of cluelessness. Additional experiments that manipulate the visibility of the shoulder and neck motion are necessary to fully explore this possibility. If shoulder motion is, in fact, sufficient to specify an expression of cluelessness, this would be clear evidence that any system used to describe conversational expressions must encode shoulder, neck, and rigid head motion as well as facial motion.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Fig. 10. Expression recognition accuracy for the nine different expressions.

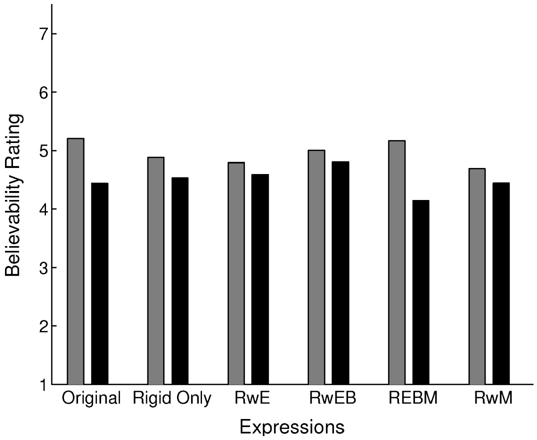


Fig. 11. Average believability ratings. A rating of 1 means that the expression was not believable (i.e., the actor or actress was clearly pretending), while a rating of 7 indicated that the expression was genuine. The grey bars show the believability ratings for those trials where the participants correctly identified the expression. The black bars show the ratings for trials where the expression was not correctly identified.

The results for happy expressions are also quite clear: Rigid head, eye, and eyebrow motion carry little or no information about happiness. When the mouth was allowed to move, participants could recognize happy expressions as well as they could for the original sequence. Thus, mouth motion seems to be sufficient to accurately recognize an expression of happiness.

The results for pleasantly surprised expressions were similar to happiness. Rigid head motion carries little or no information. Eye and eyebrow motion do not, in-and-of themselves, seem to provide much of a basis for identifying this expression. The mouth region is where most of the information is located. Unlike expressions of happiness, however, the addition of eye and eyebrow motion is necessary for proper recognition of pleasantly surprised expressions. It seems, then, that the eyes and eyebrows do provide information about surprise, but that they play more of a supporting role rather than a central role.

The results for confusion are likewise straightforward: There is some information in the rigid head motion and the mouth motion might contribute a small amount. The majority of confusion, however, is specified in the motion of the eyebrows.

The results for sadness and disgust are somewhat more complicated. Each of the four types of motion seems to contribute something to the accurate recognition of these expressions. Although each type of motion, by itself, allows the identification of recognition of these two expressions some of the time, rigid head motion and mouth motion together seem to be sufficient to specify disgust. For accurate recognition of sadness, however, all four types of motion were required.

Finally, each type of motion seems to carry some information about thinking expressions, with eye motion playing the central role. Rigid head motion seems to provide only minimal information. The addition of eye motion allowed the participants to recognize the thinking expressions almost as well as as in the original footage. The further addition of eyebrow motion allows normal recognition of the expression. Interestingly, mouth motion carries some information, but the addition of this motion to an image sequence that already has rigid head, eye, and eyebrow motion does not help much. In other words, the mouth motion can be informative, but does not seem to be necessary.

3.2.2 *Believability Ratings.* Figure 11 shows the average values for the believability ratings. The results are split into the ratings for those trials where the expression was correctly identified (grey) and

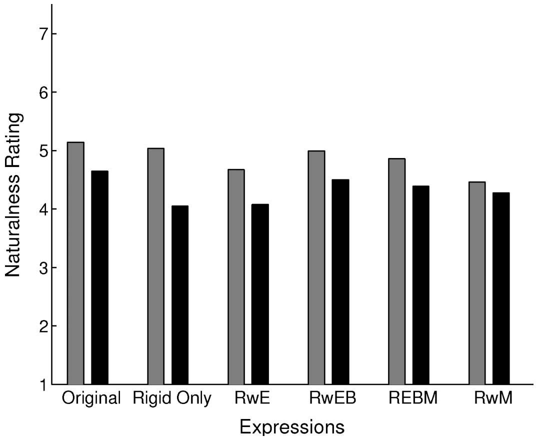


Fig. 12. Average naturalness ratings. A rating of 1 means that the expression was not natural (i.e., something humans do not normally do), while a rating of 7 indicated that the expression was natural. The grey bars show the ratings for those trials where the participants correctly identified the expression. The black bars show the ratings for trials where the expression was not correctly identified.

those where the expression was not correctly identified (black). Several things are immediately clear from the figure. First, as was mentioned previously, the ratings were similar regardless of whether one correctly identified the expression or not. One possible interpretation of this is that even if one does not know what an expression means, it is clear that the emotion underlying the expression was genuine.

3.2.3 *Naturalness Ratings.* Figure 12 shows the average values for the naturalness ratings, separated into correctly and incorrectly identified trials. As was found for the believability ratings, the naturalness ratings were similar regardless of whether one correctly identified the expression or not: Even if one does not understand the intent of an expression, it still looks like something humans normally do. As one can see in the figure, there was no meaningful variation in the naturalness ratings across the freeze-face conditions. Some of the individual expressions show some variation across the freeze-face conditions, but again no meaningful pattern is apparent. Since the participants were explicitly asked to rate expressions that had visible manipulation artifacts as unnatural, these minor variations for the different expressions are most likely due to manipulation artifacts. The fact that, overall, the manipulated conditions were not rated as more unnatural than the original condition suggests that the manipulation technique did not introduce many artifacts.

4. CONCLUSION

The collection of advanced computer graphics techniques presented here produces natural, realistic-looking, artifact-free variations on real video sequences and thus represents a potentially powerful tool in the examination of facial expressions. In additional to validating the effectiveness of the techniques, the present experiments also provided some initial insights into which facial regions must move for various conversational expressions to be recognized. Although humans can and do use a variety of different facial motions to express themselves, there was a remarkable degree of consistency in which motions were needed to specify the nine conversational expressions used here (at least for the actors and actresses used in the present experiment). Most of the expressions seem to rely heavily on a single facial region to convey their meaning. Interestingly, the results show that rigid head motion plays a significant role in identifying expressions (this is especially true for agreement and disagreement). Furthermore, the results seem to suggest that shoulder motion may carry valuable information for some expressions (e.g., shrugging for confusion or drooping in sadness). This highlights the need to include description of the motion of regions external to the face in any system that desires to provide a *complete* description of conversational expressions.

Expressions of happiness and pleasant surprise seem to be primarily specified through mouth motion, although the motion of the eyes and eyebrows are necessary for truly accurate recognition of pleasantly surprised expression. Confusion seems to be mostly defined by eyebrow motion. Thinking relies heavily on eye motion. Sadness and disgust are the two exceptions to the trend of using primarily one region, as they both seem to require all four types of motion (rigid head, eye, eyebrow, and mouth motion). It is clear, however, that these four types of motion are generally sufficient to produce expressions that are as easy to recognize as expressions in the original recordings.

The qualitative description of the conversational expressions produced by the present experiments can already be helpful in synthesizing these expressions. For example, one could focus more resources on the relevant areas for the different expressions. Nonetheless, it still remains to be determined *exactly* how the various areas move. For example, eyeball motion (i.e., direction of gaze) is very important for thinking expressions, but do we direct our gaze upward and to the left and then hold it there? Or, do we move our gaze back and forth horizontally? In our recordings, both of these types of eyeball motion occurred, but it is not clear whether one or both of them are acceptable signals for a thinking expression. The techniques presented and validated in the present work offer a powerful and effective tool for manipulating image sequences, offering a means for producing a more detailed, systematic description of the necessary and sufficient components of conversational facial expressions.

REFERENCES

Barton, J. J. S. 2003. Disorders of face perception and recognition. *Neurologic Clinics 21*, 521–548.

Bassili, J. 1978. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology 4*, 373–379.

Bassili, J. 1979. Emotion recognition: The role of facial motion and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology 37*, 2,049–2,059.

Bavelas, J. B., Black, A., Lemery, C. R., and Mullett, J. 1986. I show how you feel—motor mimicry as a communicative act. *Journal of Personality and Social Psychology 59*, 322–329.

Bavelas, J. B. and Chovil, N. 2000. Visible acts of meaning—an integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology 19*, 163–194.

Bavelas, J. B., Coates, L., and Johnson, T. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology 79*, 941–952.

Blanz, V. and Vetter, T. 1999. A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99 Conference Proceedings*. 187–194.

Bruce, V. 1988. *Recognising Faces*. Lawrence Erlbaum Associates, Mahwah, NJ.

Bull, P. 2001. State of the art: Nonverbal communication. *The Psychologist 14*, 644–647.

Bull, R. E. and Connelly, G. 1986. Body movement and emphasis in speech. *Journal of Nonverbal Behaviour 9*, 169–187.

Cassell, J., Bickmore, T., Cambell, L., Vilhjalmsson, H., and Yan, H. 2001. More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems 14*, 22–64.

Cassell, J. and Thorisson, K. R. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence 13*, 519–538.

Condon, W. S. and Ogston, W. D. 1966. Sound film analysis of normal and pathological behaviour patterns. *Journal of Nervous and Mental Disease 143*, 338–347.

Cunningham, D. W., Breidt, M., Kleiner, M., Wallraven, C., and Bülthoff, H. H. 2003a. How believable are real faces?: Towards a perceptual basis for conversational animation. In *Computer Animation and Social Agents 2003*. 23–29.

Cunningham, D. W., Breidt, M., Kleiner, M., Wallraven, C., and Bülthoff, H. H. 2003b. The inaccuracy and insincerity of real faces. In *Proceedings of Visualization, Imaging, and Image Processing 2003*. 7–12.

Cunningham, D. W., Nusseck, M., Wallraven, C., and Bülthoff, H. H. 2004. The role of image size in the recognition of conversational facial expressions. *Computer Animation & Virtual Worlds 15*, 3–4 (July), 305–310.

DeCarlo, D., Revilla, C., and Stone, M. 2002. Making discourse visible: Coding and animating conversational facial displays. In *Proceedings of the Computer Animation 2002*. 11–16.

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. 1999. Classifying facial actions. *IEEE Trans. Pattern Anal. Mach. Intell. 21*, 10, 974–989.

Edwards, K. 1998. The face of time: Temporal cues in facial expressions of emotion. *Psychological Science 9*, 270–276.

Ekman, P. 1972. Universal and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation 1971*, J. R. Cole, Ed. University of Nebraska Press, Lincoln, NE, 207–283.

Ekman, P. and Friesen, W. 1978. *Facial Action Coding System*. Consulting Psychologists Press, Inc., Palo Alto, California.

Essa, I. and Pentland, A. 1994. A vision system for observing and extracting facial action parameters. In *CVPR94*. 76–83.

Frank, M. G. and Stennett, J. 2001. The forced-choice paradigm and the perception of facial expressions of emotion. *Journal of Personality and Social Psychology 80*, 75–85.

Gooch, B. and Gooch, A. 2001. *Non-Photorealistic Rendering*. A.K. Peters. Natick, Massachusetts, USA.

Humphreys, G., Donnelly, N., and Riddoch, M. 1993. Expression is computed separately from facial identity, and is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia 31*, 173–181.

Isaacs, E. and Tang, J. 1993. What video can and can't do for collaboration: a case study. In "*ACM Multimedia '93*". ACM, New York, 496–503.

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., and Akamatsu, S. 2001. Dynamic properties influence the perception of facial expressions. *Perception 30*, 875–887.

Kleiner, M., Wallraven, C., and Bülthoff, H. H. 2004. The MPI Videolab—a system for high quality synchronous recording of video and audio from multiple viewpoints. Tech. Rep. 123, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Motley, M. T. 1993. Facial affect and verbal context in conversation—facial expression as interjection. *Human Communication Research 20*, 3–40.

Pantic, M. and Rothkrantz, L. J. M. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 12, 1424–1445.

Pelachaud, C., Badler, N., and Viaud, M. 1994. Final report to the NSF of the standards for facial animation workshop. Tech. rep., University of Pennsylvania, School of Engineering and Applied Science, Computer and Information Science Department, Philadelphia, PA 19104-6389.

Poggio, I. and Pelachaud, C. 2000. Perfomative facial expressions in animated faces. In *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. MIT Press, Cambridge, MA, 115–188.

Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., and J., D. 2001. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior 25*, 167–186.

Strothotte, T. and Schlechtweg, S. 2002. *Non-Photorealistic Computer Graphics: Modeling, Rendering, and Animation*. Morgan Kaufmann, San Francisco, CA.

Thorisson, K. R. 1996. Toonface: A system for creating and animating cartoon faces. Tech. Rep. 96-01, M.I.T. Media Laboratory, Learning & Common Sense Section.

Vertegaal, R. 1997. Conversational awareness in multiparty vmc. In "*Extended Abstracts of CHI'97*". ACM, Atlanta, 496–503.

Yngve, V. H. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago, Illinois. 567–578.

Received December 2004; revised April 2005; accepted April 2005