The components of conversational facial expressions

Douglas W. Cunningham*
MPI for Biological Cybernetics

Mario Kleiner
MPI for Biological Cybernetics

Christian Wallraven
MPI for Biological Cybernetics

Heirich H. Bülthoff
MPI for Biological Cybernetics

Abstract

Conversing with others is one of the most central of human behaviours. In any conversation, humans use facial motion to help modify what is said, to control the flow of a dialog, or to convey complex intentions without saying a word. Here, we employ a custom, image-based, stereo motion-tracking algorithm to track and selectively "freeze" portions of an actor or actress's face in video recordings in order to determine the necessary and sufficient facial motions for nine conversational expressions. The results show that most expressions rely primarily on a single facial area to convey meaning, with different expressions using different facial areas. The results also show that the combination of rigid head, eye, eyebrow, and mouth motion is sufficient to produce versions of these expressions that are as easy to recognize as the original recordings. Finally, the results show that the manipulation technique introduced few perceptible artifacts into the altered video sequences. The use of advanced computer graphics techniques provided a means to systematically examine real facial expressions. This provides not only fundamental insights into human perception and cognition, but also yields the basis for a systematic description of what needs to be animated in order to produce realistic, recognizable facial expressions.

CR Categories:

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

Keywords: applied perception, facial expressions, computer graphics, animation

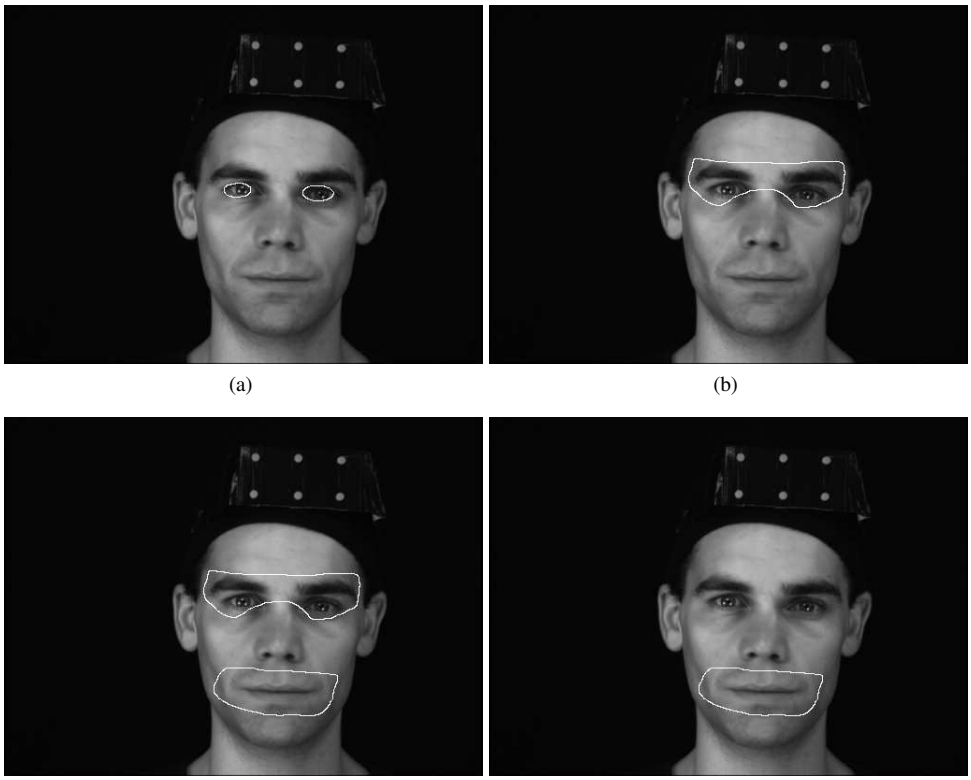
1 Introduction

Communication is simultaneously one of the most important and one of the most complex tasks that humans undertake. A central part of any conversation is, of course, trying to determine what the other individuals in the conversation are trying to communicate. Facial motions play several roles in communication. The can be used to modify the meaning of what is being said [Bull and Connelly 1986; Bavelas and Chovil 2000; Condon and Ogston 1966; Motley 1993; DeCarlo et al. 2002]. For example, a statement of surprise does not have quite the same meaning when it is accompanied by a look of boredom. Facial motion is also useful in controlling conversational flow [Bavelas et al. 1986; Bull 2001; Cassell and Thorisson 1999; Cassell et al. 2001; Poggi and Pelachaud 2000]. This can be done with simple motions, such as using the direction of eye gaze to determine who is being addressed [Cassell and Thorisson 1999; Cassell et al. 2001; Isaacs and Tang 1993; Vertegaal 1997]. Facial motion can also be used to more subtly direct the flow of a conversation. For example, a listener can inform 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, they will probably continue talking. A look of confusion, disgust, or boredom, however, will almost certainly prompt very different behaviour 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.

Since facial expressions can be a very powerful form of communication, it is only natural that they should be used in applied settings, such as Human-Machine interfaces. The synthesis of proper conversational expressions is, however, 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. 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; Cunningham et al. 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.

A systematic description of the necessary and sufficient components of conversational expressions could prove very helpful in the synthesis of conversational agents. 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 [Pantic and Rothkrantz 2000; Donato et al. 1999] for reviews), although many of these methods are not specifically interested in mimicking human perceptual capabilities. There is also a large literature 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 method 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 can 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.

*e-mail: douglas.cunningham@tuebingen.mpg.de



(a) (b) (c) (d)

Figure 1: 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 outlined 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 line 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 eyes. (b) Condition RwEB: This condition had rigid head, eye, 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.

Thus, a detailed analysis of which elements go together to produce different expressions is external to FACS itself [Sayette et al. 2001]. In other words, FACS itself does not tell us what the necessary and sufficient components of facial expressions are.

Many of the descriptive systems developed for facial expressions focus explicitly on statically visible deformations of the facial surface. Given that temporal information seems to be of central importance to the perception and recognition of expressions [Bassili 1978; Bassili 1979; Bruce 1988; Edwards 1998; Kamachi et al. 2001], and that static and dynamic information for expressions seem to be processed in separate areas of the human brain [Humphreys et al. 1993], 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. Providing an empirical basis for such descriptions is, however, not easy since it is exceedingly difficult and time-consuming to systematically alter by hand sub-regions of a face throughout entire video sequences in order to examine the role of different types of facial motion. Here, we used advanced computer graphics techniques to semi-automatically manipulate video recordings of real expressions. The resulting manipulated sequences were used in a psychophysical experiment in an attempt to provide some initial insights into the components of nine conversational expressions: agreement, disagreement, happiness, sadness, thinking, confusion, clueless, disgust, and surprise. These nine expressions were recorded, as part of our video database of facial expressions, from six individuals with six synchronized digital video cameras (Section 2). The recordings were 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). 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 1). In all conditions, the rigid head motion was left intact.

2 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/sec of *fully synchronized* non-interlaced, uncompressed video in PAL resolution (768 x 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, a person was seated in front of the cameras (which were arranged in a semi-circle around the actor at a distance of 1.5 meters) and a homogeneous, black background

was placed behind them. The individual was 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).

3 Methodology

The expressions were recorded from six different people (three male and three female). One of the individuals was a professional actor. The remaining individuals were amateur actors and actresses. The expressions were elicited 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. For the present experiment, nine expressions were recorded: agreement, disagreement, disgust, thinking, pleased/happy, sadness, pleasantly surprised, clueless (as if the actor did not know the answer to a question), and confusion (as if the actor did not understand what was just said). 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 to react without speaking, unless they felt that speech was absolutely required to react naturally. For each of the expressions, three repetitions were performed with a neutral expression both preceding and succeeding each repetition. All of the recordings were used in a pilot experiment to determine the best repetition for each expression for each actor or actress. Each of the recordings was edited so that the video sequence began on the frame after the face began to move away from the neutral expression and ended after reaching the peak of the expression. The resulting 54 video sequences varied considerably in length; The shortest sequence was 17 frames long (0.68 seconds) and the longest lasted 194 frames (7.76 seconds). There was no apparent simple correlation between expression and duration.

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 post-processing (see section 3.1). The post-processing 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 frozen condition (Rigid Only), all of the face was held still (i.e., only the rigid head motion was present). In the second condition (RwE), both the original rigid head motion and the motion of the eyes were present (see Figure 1). 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.

All of the video sequences were shown to nine individuals (hereafter referred to as *participants*) in a psychophysical experiment. 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. For the present experiment, the size of the images was reduced to 512x384. The participants sat at a distance of approximately 0.5 meters from the computer screen. Nine expressions crossed with six actors and six "freeze face" conditions yielded 324 trials. Since the experiment lasted approximately 2.5 hours, participants were given the opportunity to take a break every 40 trials. The order in which the trials were presented was completely randomized for each participant. A single trial consisted of the video sequence being shown repeatedly in the center of the screen. A 200 ms blank screen was inserted between repetitions of the video clip. When participants were ready to respond (which they indicated by pressing the space bar), the video sequence was removed from the computer screen, and the participants had to perform three 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 other were not, the expressions were listed in English as well as German (see [Cunningham et al. 2003b] for a list of the German names of the expressions).

One might be worried that this type of task does not properly reflect identification performance. [Frank and Stennett 2001] have shown, however, that this type of task (a non-forced choice task) is highly correlated with other identification procedures (e.g., free description of the expressions). The non-forced 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 (i.e., in forced-choice tasks) 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.

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

3.1 Image manipulation technique

As mentioned previously, each of the individuals who were recorded wore a black hat with a tracking target (a black rectangular plate with six green markers; see Figure 1). Since the VideoLab has six cameras that are fully synchronized, we were able to utilize a custom, image-based, stereo motion-tracking algorithm to recover the three-dimensional 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 x 512 texels) of the individual. To ensure that the elements of the meshes and

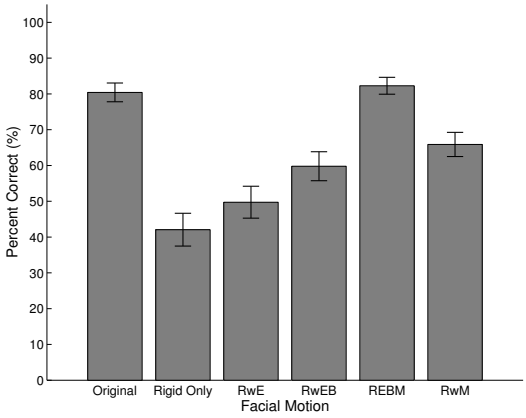


Figure 2: Overall recognition accuracy. The percentage of the time that the participants correctly identified the expressions is shown for the six "freeze face" conditions. The error bars represent the standard error of the mean.

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 re-write"). 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 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.

4 Results and Discussion

Overall, the participants were very good at identifying the expressions 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, non-trained individuals [Cunningham et al. 2003a]. The believability ratings were also similar to previous results: Generally, the participants found the expressions to be somewhat believable, but not overwhelmingly so. Finally, the participants found all of the expressions to be rather natural in all conditions.

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

Figure 3: A single snapshot from one actor's clueless expression. The upwards 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".

4.1 Recognition accuracy

Figure 2 shows, for each of the six "freeze face" conditions, the percentage of trials where the expressions were correctly identified. On 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. The addition of eye motion (condition RwE) tended to improve recognition accuracy some, and the further addition of eyebrow motion (condition RwEB) tended to improve recognition accuracy still 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, and the joint usage of all four seems to be sufficient to identify all the expressions used here. This suggests that motion of other facial regions (e.g., the forehead and cheeks) are not *necessary* to identify these expression (at least not with a 10 alternative, non-forced choice task).

As can be seen in Figure 4, different expressions relied on different types of motion to convey their meaning. Agreement, disagreement, and clueless, for example, all seem to be sufficiently specified by rigid head motion. For agreement and disagreement, the recognition accuracies are near 100%, suggesting a possible ceiling effect. That is, it is possible that the other facial areas contribute to the recognition of these expressions, but the fact that accuracies are so high in the Rigid Only condition prevents us from measuring the contribution of other facial motions. A more difficult task is required to examine the potential contribution of the other types of facial motion to agreement and disagreement. It is also possible that the other facial areas do not contribute to the recognition of agreement and disagreement, but would help to distinguish between variants of these expression (e.g., reluctant agreement, enthusiastic agreement, considered agreement).

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 3): The chin moved upwards, the lower lip was pushed out, and the corners of the mouth moved downwards. The fact every actor and actress had this characteristic motion makes it unclear why people do not use or need this information. One potential explanation for the high success in the rigid head motion condition 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

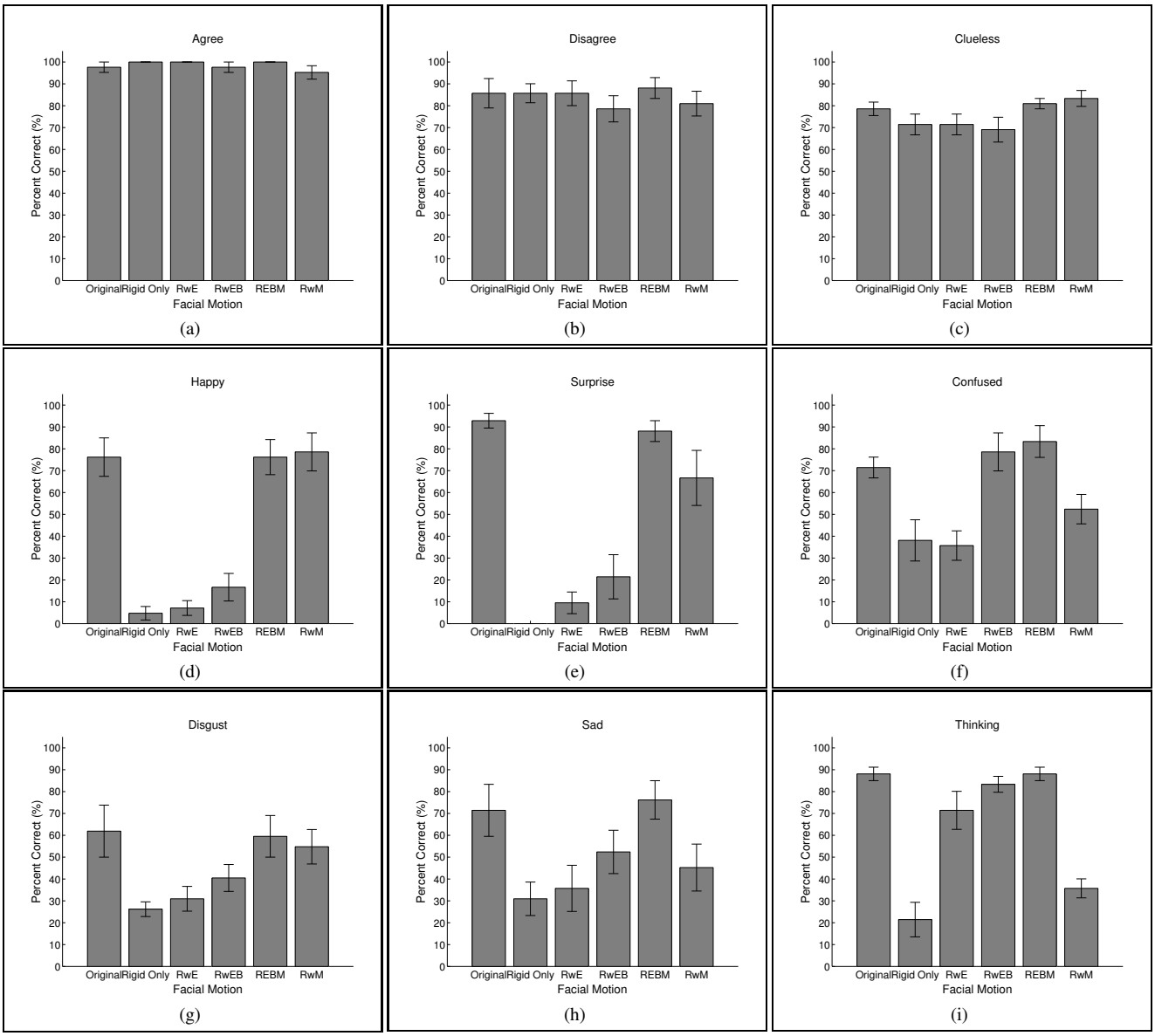


Figure 4: Expression recognition accuracy for the nine different expressions.

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 cluelessnes. Future experiments that eliminate all perceptible shoulder and neck motion would be useful in further explorations of the components of cluelessness.

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. The addition of eye and eyebrow motion is, however, 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 straight-forward: 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 seem 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.

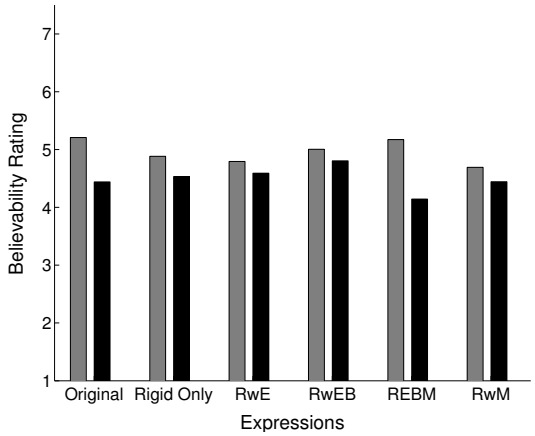


Figure 5: Average believability ratings. A rating of 1 means that the expression was not believable (i.e., the actor or actress was clearly pretending), while a rating of 7 indicated that the expression was genuine. The grey bars show the believability ratings for those trials where the participants correctly identified the expression. The black bars show the ratings for trials where the expression was not correctly identified.

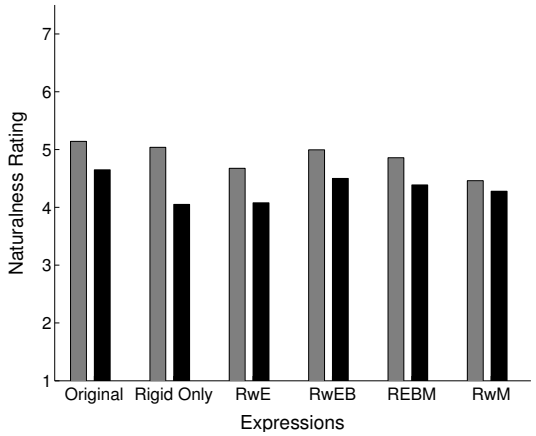


Figure 6: Average naturalness ratings. A rating of 1 means that the expression was not natural (i.e., something humans do not normally do), while a rating of 7 indicated that the expression was natural. The grey bars show the ratings for those trials where the participants correctly identified the expression. The black bars show the ratings for trials where the expression was not correctly identified.

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.

4.2 Believability ratings

Figure 5 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 those where the expression was not correctly identified (black). Several things are immediately clear from Figure 5. First, 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 emotion underlying the expression was genuine. Another, equally plausible, explanation is that participants did not use the believability scale properly. This alternate explanation is supported by the fact that believability ratings were basically the same for all expressions in all conditions. That is, nothing we manipulated altered the ratings at all. This pattern is similar to previous work with these types of expressions ([Cunningham et al. 2003a; Cunningham et al. 2003b]).

4.3 Naturalness ratings

Figure 6 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 face freeze conditions. There were some variations across the freeze face conditions for the different expressions, 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.

5 Conclusion

Using advanced computer graphics and computer vision techniques, we were able to examine the basic components of conversational facial expressions. The techniques, which seem to introduce few noticeable artifacts, allowed us to use real video sequences, selectively adding and removing facial motion to determine which facial areas need to move in order for different expressions to be accurately 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. Agreement, disagreement, and possibly cluelessness seem to only need rigid head motion. Since the motion of the shoulders was not completely eliminated, it is difficult to draw any strong conclusions about which facial motions are necessary for cluelessness. 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 this trend, 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 sufficient to produce expressions that are as easy to recognize as expressions in the original recordings.

While it is now clear what areas need to move to produce recognizable versions of these expressions, it still remains to be determined exactly how these areas move. For example, eyeball motion (i.e., direction of gaze) is very important for thinking expressions, but do we direct our gaze upwards and to the left and then hold it there? Or, do we move our gaze moved 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 qualitative description of the conversational expressions that the present experiment yielded can be useful helping to synthesize these expressions, by allowing one to, for example, focus more resources on the relevant areas for the different expressions. Moreover, the results validate the effectiveness of the image manipulation technique, which offers a means for producing a more detailed, systematic description of natural, recognizable facial expressions.

References

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., 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., 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., 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.

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

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

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.

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.

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. 2003. How believable are real faces?: Towards a perceptual basis for conversational animation. In *Computer Animation and Social Agents 2003*, 23 – 29.

CUNNINGHAM, D., BREIDT, M., KLEINER, M., WALLRAVEN, C., AND BÜLTHOFF, H. 2003. The inaccuracy and insincerity of real faces. In *Proceedings of Visualization, Imaging, and Image Processing 2003*.

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., AND FRIESEN, W. 1978. *Facial Action Coding System*. Consulting Psychologists Press, Inc., Palo Alto, California.

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.

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.

POGGI, 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.

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, 567 – 578.