The contribution of different facial regions to the recognition of conversational expressions

Manfred Nusseck
Max Planck Institute of Biological Cybernetics, Tübingen, Germany
WSI/GRIS, University of Tübingen, Tübingen, Germany

Douglas W. Cunningham
Max Planck Institute of Biological Cybernetics, Tübingen, Germany

Christian Wallraven
Max Planck Institute of Biological Cybernetics, Tübingen, Germany

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

The human face is an important and complex communication channel. Humans can, however, easily read in a face not only identity information but also facial expressions with high accuracy. Here, we present the results of four psychophysical experiments in which we systematically manipulated certain facial areas in video sequences of nine conversational expressions to investigate recognition performance and its dependency on the motions of different facial parts. The results help to demonstrate what information is perceptually necessary and sufficient to recognize the different facial expressions. Subsequent analyses of the facial movements and correlation with recognition performance show that, for some expressions, one individual facial region can represent the whole expression. In other cases, the interaction of more than one facial area is needed to clarify the expression. The full set of results is used to develop a systematic description of the roles of different facial parts in the visual perception of conversational facial expressions.

Keywords: faces, conversational expressions, facial motions, expression recognition


Introduction

The face is one of the most ecologically relevant entities for visual perception. Humans are astonishingly good at identifying and recognizing faces and decoding facial expressions. The faces around us change expressions constantly and in a variety of complex ways, yet one can easily tell different expressions apart within a glance. The moods, attitudes, and emotions of a person can be determined from very subtle changes in their face, making processing of facial information one of the most relevant perceptual skills in everyday life.

Faces are particularly important for communication, where facial features can be used to control the flow of a conversation (Bull, 2001). Cassell and Thorisson (2001) showed that head and gaze direction are useful in managing turn-taking in face-to-face conversation. More complex control is also possible through “back-channel” responses (Yngve, 1970). Bavelas, Coates, and Johnson (2000), for example, examined storytellers and found that without emphatic responses from the listeners, the speakers shortened the story and omitted details. Facial gestures can, of course, also help to disambiguate or modify the meaning of spoken words (Bull, 2001; Cohn, Schmidt, Gross, & Ekman, 2002).

How we recognize and distinguish faces and facial expressions is the subject of a large field of research. Scientific studies on the nature of facial expressions arguably began with Darwin (1872). He gathered evidence that some resulting facial expressions seem to have certain relations to emotional states. Later, several studies on the perception of facial expressions showed explicitly that different facial areas are responsible for the recognition of different facial expressions (for an overview, see Adolphs, 2002; Calder & Young, 2005; Russell & Fernandez-Dols, 1997; Schwaninger, Wallraven, Cunningham, & Chiller-Glaus, 2006).

The cultural dependency of facial expressions was investigated by Ekman and colleagues (Ekman, 1972, 1989, 1992, 2004), in which they found that some expressions were recognized very accurately across different cultures. These expressions, which are happiness, surprise, fear, sadness, anger, and disgust, are often referred to as “universal emotional expressions.”

Although the possibility of directly observing a mental state is still a heavily debated topic in the “theory of
mind” literature, some claim that emotional facial expressions are direct links to inner states. Baron-Cohen, Wheelwright, and Jolliffe (1997) suggested that this is also the case for non-emotional expressions and classified all facial expressions into “basic” emotions and “complex” mental states. For example, averting the gaze to look upward and to the left or right, even when there is no apparent object in view, is clearly recognized as a state of thinking about something. Pelachaud and Poggi (2002) examined the communicative functions of facial expressions, including performatives, turn-taking signals, and deictic elements. They developed five groups for the classification of expressions:

1. expressions that convey the location and/or properties of concrete or abstract objects or events;
2. expressions that relate to the belief in an expressed statement;
3. expressions for intonation;
4. expressions that communicate affective states; and
5. expressions that provide metacognitive information about the person’s mental actions.

For the present experiments, we decided to focus on facial expressions that are important for, and easily observable in, a conversation in a Western culture. The set of nine expressions used (for static illustrations from one actor, see Figure 1) contains a combination of both basic emotional (happiness, sadness, surprise, and disgust) and complex mental state (agreement, disagreement, confusion, thinking, and clueless) expressions. Note that the non-emotional expressions are concentrated in groups 4 and 5 of Pelachaud and Poggi’s (2002) classification scheme.

Traditionally, one central goal of facial expression research is to describe the specific facial deformations that occur for a given expression. In the most well-known example of this, Ekman and Friesen (1978) developed a descriptive system for individual facial deformations called the facial action coding system (FACS). FACS starts with the acknowledgment that the activation of specific facial muscles produces specific, visually detectable distortions in the face. FACS uses these visible distortions, referred to as “action units” (“AU”), as the basic building blocks of facial expressions. Theoretically, any facial expression can be described by the correct combination of action units. Numerous studies have used this system. Kohler et al. (2004), for example, examined the presence of AUs in four emotional expressions (happy, sad, angry, and fear). This highlights an important aspect of FACS: It does not, in-and-of itself, say which AUs go together to form specific expressions. Furthermore, it also does not address the question of whether these AUs must be present for the expression to be recognized. The goal of the present work is to see which facial areas must move for normal recognition of conversational facial expressions. Since we are interested in the perceptual processing of faces, we chose to examine perceptually intuitive and salient facial areas: the mouth region, the eyes, the eyebrows, and the rigid head motion. These are, in fact, the regions that are most often examined in previous research and are the areas that are most heavily relied upon in most forms of computer-based facial animation, regardless of the software or facial descriptive system used (Kshirsagar, Escher, Sannier, & Magnenat-Thalmann, 1999).

A majority of studies on facial expressions have restricted their research to static displays. There are many potential reasons for this. For example, static images clearly provide enough information for us to recognize facial expressions quite well. Additionally, static stimuli are much easier to describe, manipulate, and analyze than dynamic stimuli. Nonetheless, faces in the everyday world are rarely static. Since the goal of the present experiments is to examine the relative importance of different facial areas in the everyday perception of expressions, we chose to use natural-looking stimuli and thus focus solely on dynamic video recordings. Here, it is important to note that while FACS has been very successfully applied to static images, its application to video sequences is very limited. More specifically, FACS was not designed for
and is not adept at describing the dynamic aspects of facial expressions (Essa & Pentland, 1994; Sayette, Cohn, Wertz, Perrott, & Parrott, 2001).

One might be tempted to ask whether dynamic stimuli are any different from static stimuli. After all, if they are not, then why should one go through all of the extra effort of using dynamic stimuli? At the extreme, one might go so far as to suggest that dynamic sequences are nothing more than a collection of static snapshots. Indeed, the application of FACS to video sequences explicitly requires this assumption, as it has no method of encoding spatiotemporal information. There is, however, a lot of evidence demonstrating that dynamic stimuli are fundamentally different from static stimuli in general (e.g., Gibson, 1979) as well as for facial expressions in specific. Possibly the first evidence that there is some temporal information specific to facial expressions comes from Bassili’s (1978, 1979) seminal work using point-light displays. Subsequent research has clearly shown that dynamic information can improve the recognition of facial expressions even in more natural displays (Ambadar, Schooler, & Cohn, 2005; Constantini, Pianesi, & Prete, 2005; Cunningham & Wallraven, 2008; Ehrlich, Schiano, & Sheridan, 2000; Harwood, Hall, & Shinkfield, 1999; Katsyri, Klucharev, Frydrych, & Sams, 2003; Wallraven, Breidt, Cunningham, & Bültthoff, 2007; Wehrle, Kaiser, Schmidt, & Scherer, 2000; Weyers, Mühlberger, Hefele, & Pauli, 2006). A recent series of perceptual experiments shows that not only are dynamic expressions recognized more easily than static expressions, but also that no explanation based solely on static information can explain the difference (Cunningham & Wallraven, 2008). That is, there is some characteristic information for specific expressions that is only available over time. Finally, there is anatomical work suggesting that static and dynamic facial expressions are processed differently in the human brain, possibly involving completely different brain structures (Adolphs, 2002; Humphreys, Donnelly, & Riddoch, 1993; LaBar, Crupain, Voyvodic, & McCarthy, 2003; Schwaninger et al., 2006). Regardless of whether dynamic expressions are different from static expressions or not, it remains true that dynamic expressions represent the normal case in everyday life and thus are the focus of the present work.

The stimuli used here are inspired by previous research, which used a sophisticated video manipulation technique to systematically replace motions in specific facial areas (Cunningham, Kleiner, Wallraven, & Bültthoff, 2005). The resulting manipulated stimuli looked just as realistic as the original recordings and thus satisfy the desire for naturalness. In their study, Cunningham et al. (2005) found that, for some expressions, one single area is sufficient, while other expressions required a combination of different areas. The specific area required was different for the different expressions. The video sequences used by Cunningham et al. included some non-facial areas, such as portions of the ears, the shoulders, and the neck. Based on some rather surprising results (e.g., that rigid head motion is sufficient for fully normal recognition of Clueless), Cunningham and colleagues suggested that the non-facial areas might be playing a stronger role than at first expected.

Cunningham et al.’s (2005) experiment focused primarily on the sufficiency of certain regions (i.e., the expression can be recognized well even if nothing except this region moves). To build a more complete picture of expression recognition, however, one must also examine the issue of necessity (i.e., the expression cannot be properly recognized if the region’s motion is not present). Thus, after replicating and extending Cunningham et al.’s sufficiency experiment (with the critical difference that all non-facial areas were removed from the video sequences), the present work examined the necessity of facial regions using the same manipulation technique.

It is important to note that, in the first two experiments (Experiments 1 and 2), rigid head motion was always present. Rigid head motion, however, is an important signal especially in a conversation. Not only does rigid head movement provide the observer with a moving stimulus and with more views of the head than a static head would (O’Toole, Roark, & Abdi, 2002), but individual head movements occur frequently in expressive gestures. For example, the head often raises and turns toward someone at the beginning of a smile (Cohn & Kanade, 2006). Likewise, Keltner (1995) found that before an embarrassed smiling was expressed, the head and the gaze direction were directed initially down and away. Furthermore, rigid head motion functions as a turn-taking signal, a deictic gesture. Indeed, Wallraven et al. (2007) found that rigid head motion can play a central role in the recognition of facial expressions. Thus, the presence of rigid head motion in every condition may mask the contribution of individual facial areas. Therefore, Experiments 3 and 4 replicate Experiments 1 and 2, respectively, but without rigid head motion. Figure 2 shows a schematic overview for the four experiments.

Figure 2. Schematic diagram of the four experiments. The eyes, the eyebrows, and/or the mouth region were either the only parts in the face that moved (Experiments 1 and 3) or were the only parts that were frozen (Experiments 2 and 4). Rigid head motion was present in Experiments 1 and 2 but eliminated in Experiments 3 and 4.
General methods

Recordings

The recordings were done with the Max Planck Institute for Biological Cybernetics’s VideoLab (for more details, see Kleiner, Wallraven, & Bülthoff, 2004), which is a custom-designed setup with six digital cameras arranged in a semicircle around the actor/actress. These cameras are fully synchronized and have a PAL video resolution of 768 × 576. The actors/actresses were filmed at 25 frames/s. Sound was not recorded.

We recorded six different individuals (two male and four female), all of whom speak German as their native language. Each individual performed nine different expressions as if the central camera was a second person. With one exception, all individuals were amateur actors/actresses. There is some general concern that results using “posed” facial expressions (especially those detached from any context) cannot be generalized to everyday life (Russell, 1993). Unfortunately, obtaining “real” versions (e.g., a candid recording of a conversation) of some expressions presents some serious difficulties. Ignoring the ethical considerations involved in putting people into the situations required to force them to really experience certain emotions (e.g., fear, anger, disgust, etc.), recordings of real situations leave one in the uncomfortable situation that one is not 100% sure what the expression was supposed to be. Indeed, in a fully unconstrained situation, one cannot be sure that the desired expression will be emitted at all, let alone know what intensity it will have. As a compromise, we chose to use a method acting protocol.

The resulting expressions are as close to natural as possible while still retaining complete knowledge about what the expression is supposed to be. Specifically, the actors/actresses were told a little background scenario describing a specific situation. They were then asked to remember a similar situation in their life, imagine that they were in that situation again, and to act normally (with the exception that they were asked not to use their hands and not to speak). To aid in the post-processing of the video sequences, the actors/actresses wore a black hat with six green dots on it, which worked as head tracking markers.

Each expression was recorded for each actor/actress three times in a row with a little pause between repetitions. The best repetition of each expression for each actor was empirically determined in previous work (Cunningham et al., 2005). Since no difference in recognition accuracy, reaction time, or believability was found between a “clipped” version of the video sequences (where only the frames between the initial neutral expression and the expression’s peak were shown) and the full version (neutral to peak and back to neutral), the present studies used the clipped version. Likewise, since there seems to be no difference in performance between viewing the sequence multiple times and only once (Cunningham, Breidt, Kleiner, Wallraven, & Bülthoff, 2003; Cunningham, Kleiner, Wallraven, & Bülthoff, 2004), the sequences here were therefore shown once.

The length of the sequences varies considerably: The shortest sequence is 15 frames long (0.6 s) and the longest has 298 frames (11.9 s). There are no general differences between the actors/actresses in the duration of the expressions and no straightforward correlation between expression and duration.

Finally, in order to control for potential cultural effects, all actors/actresses were German, and all participants were either German or had lived in Germany for a long time and came originally from a similar culture (specifically, from a Western culture).

Facial manipulation technique

The traditional focus, in perception research as well as in facial animation, on the eyes, eyebrows, mouth, and rigid head motion should not be too surprising since these areas are frequently used in conversations. For instance, head movements like nodding or shaking are often associated as a sign of agreement or disagreement, as well as for greeting someone or forbidding something non-verbally. Furthermore, head orientation can be used to point at something or can reflect affect (e.g., a downward direction of the head is associated with sadness; Pelachaud, Badler, & Steedman, 1996). Eye gaze and eye contact are highly conversational signals (Argyle & Cook, 1976) and often related to thinking (Cunningham et al., 2005). The eyebrows can be used to, among other things, emphasize utterances (Ekman, 1989).

The areas were selected in the present work with a focus on perceptually salient units and thus end at “natural” boundaries in the face (e.g., the edges of the eyelid, the folds near the mouth, etc.). It should be noted that these “perceptual units” do not always overlap with the parsing performed by FACS, FACS++, etc. In more detail, the “eyes” manipulation refers to the eyeballs and sometimes the eyelids (e.g., during a blink). According to the FACS notation (Ekman & Friesen, 1978), this part roughly corresponds to AU5, AU7, AU41–46, and AU61–64, which are related to specific eye and eyelid motions. The “eyebrows” region includes both eyebrows and eyes—covering all of the above AUs as well as AU1–3 and AU9. For the mouth, we distinguish between the lips and a larger mouth region, which includes areas such as right below the nose and small parts of the cheeks (roughly AU10–28). Note that where Experiments 1 and 3 focused only on the presents of these areas, Experiments 2 and 4 addressed also the facial parts in between them. A complete analysis of each expression for each actor using FACS or any of the other existing descriptive system would necessitate manually examining each frame in the stimulus set (which is an overwhelmingly large number of images) and is beyond the scope of this paper. In this
paper, we rather aim to determine which of the motions that are present in recordings of (largely) natural expressions are perceptually necessary and/or sufficient.

We post-processed the video sequences using an image-based, stereo motion-tracking algorithm (for a detailed description, see Cunningham et al., 2005; Kleiner, Wallraven, Breidt, Cunningham, & Bülthoff, 2004). This technique facilitates the selective manipulation of certain areas of a face in the video sequence. It allows us to replace parts of the face with other images, for example from a different video sequence. Briefly, the head’s 3D position is automatically determined by the location of the green markers in several of the cameras. Since the relation between these markers and different facial regions is roughly constant, we are able (using a 3D model of the individual actor’s/actress’s head) to overlay a different movie onto select, targeted regions of the original footage. This enables us to replace any facial part in the expression with another sequence of the same facial part. To delete all expression related information from specific facial parts, we used a single frame (from the neutral expression) of the appropriate actor/actress as the source of the overlay. The same static picture was used in every manipulation of every expression for the respective actor/actress. Note that by neutral, we mean that the static frame is from a face that is completely at rest and has no expression specific information in-and-of itself. Of course facial areas might not be processed fully independently, and thus these “neutral areas” may interact with the non-frozen regions in the video stimuli and therefore indirectly provide information about an expression. For example, shaking the head with no facial motion might signify a “disagree” expression, whereas the same head motion with a particular mouth motion might be seen as “disgust.” Furthermore, the presence of a non-moving area next to a moving area may alter the perception of motion in the non-frozen areas (e.g., through motion contrast, induced motion, etc.). Thus, it is entirely possible that by freezing an area, we are changing how the non-frozen areas are processed. It is clear, however, that if recognition performance is the same when a region moves normally and when it is “frozen,” then one can conclude that the motion of the frozen region is not critical to the perception of that specific expression, and that the motion of the remaining regions is. One might be tempted to try to avoid this problem by completely removing the frozen regions. Thus, in the eye motion only condition, a pair of eyes would be seen floating in an otherwise empty screen. It should be clear, however, that such displays present precisely the same problem (there may be an interaction between the moving area and the now-empty neighboring region). Moreover, the complete absence of a context (i.e., the face) certainly changes the how the remaining areas are processed. Since the only remaining presentation style would be to present a non-manipulated face, we chose to use a static context (i.e., “neutral” face”). This allows us to present natural stimuli (the faces still look like real faces). This also allows us to address the following real-world question: If, in a real face, specific regions did not move, would the resulting expression still be recognizable?

Figure 3 outlines an example of the technique. On the input side is the original video recording. Using individual image masks, we first cut out the face from non-facial parts, such as the hat, the ears, the neck, and the shoulders. A second mask was used to define the edges of the targeted facial region. Hard edges from the masking were eliminated through a gradual falloff of the influence of the mask at its boundaries (alpha blending). The appropriate facial region was then taken from the overlay and rendered into the proper position. The resulting stimulus is the original footage with a frozen facial area. This was done automatically for all the images in the video sequence.

Note that this technique allows us, in principle, to remove the rigid head motion so that the head was fixed in one position for the whole sequence. The green dots on the hat mark the head position that is used to focus a virtual camera on one particular location related to the face. Since the virtual camera followed the head, it looks like the head is fixed. This, however, produces some slight artifacts, particularly with respect to the lighting. Sometimes, shadows move over the now motionless face. This illumination effect can possibly be used to perceptually recover the original rigid head motion.

Psychophysical methods

The task

In all four experiments, the participant’s task was to identify the expressions. They were asked to select the name of the expression from a list that was displayed on the side of the screen. This list, therefore, included all nine expressions and was displayed in English and in German.

Figure 3. Post-processing an expression. The input is the recorded video sequence (here the happy sequence). Non-facial regions were removed and a weight mask was used to replace a facial location. The output stimulus therefore showed the expression but without any movement of the masked region (here, happiness without mouth motion).
(see Figure 4). On the left side was the English description: agree, disagree, happy/pleased, sad, clueless (‘doesn’t know’), thinking, confused (‘doesn’t understand’), disgusted/repulsed, and pleasantly surprised. On the right was the German list: zustimmen, nicht zustimmen, glücklich/zufrieden, traurig, unwissend (‘weiss nicht’), nachdenklich, verwirrt (‘verstehe nicht’), angeekelt/abgestoßen, and angenehm überrascht. Forcing the participant to choose from a list of expressions can artificially inflate the accuracy rates and possibly produce other performance artifacts. Russell (1993) has shown that forced-choice designs sometimes yield consensus on the wrong answer. This can be eliminated by providing a “none of the above” (“nichts davon”) option (Frank & Stennett, 2001). Therefore, we included this option at the bottom of the list of expressions.

### Procedure and design

Each trial had the following order: Participants were presented with a black screen and were asked to press the space bar to start the video sequence. The participants were able to enter their answer as soon as the sequence started. As soon as the participant entered a response, the screen went black. If the participant did not enter a response before the sequence was shown once, the screen went black and the program waited for the participant’s response. Once a response was entered, the request to press the space bar to start the next trial was once again displayed. The order of the trials was completely randomized for each participant. The whole experiment lasted approximately 1 hour. The program offered a possibility to take a break every forty trials.

Participants were given a few practice trials before the real experiment in order to allow them to familiarize themselves with the expression list. These practice trials were presented using video sequences from an actor who was not used in the real experiment. At no time was feedback given to the participants about the accuracy of their responses. Similarly, the participants were not told about the manipulations.

### Experiment 1

Several researchers have examined the components of facial expressions in static photographs. Hanawalt (1944), for instance, discovered that the mouth is the primary information source for happiness, while the eyes specify fear and surprise. Kohler et al. (2004) determined more
precisely that *happiness* is specified by the combination of raised inner eyebrows, cheeks, and upper lips or by the combination of tightened lower eyelids and upward turned lip corners. Here, we examine the issue not of what does move in an expression but what must move. That is, the experiments are designed to determine the relative importance of the different areas. The first experiment replicates and extends Cunningham et al. (2005) by removing all non-facial regions (see step one of Figure 3). It also provides a more complete description of the relative contribution of the different areas.

**Method**

The task, procedure, participants, design, and equipment are described in the General methods section.

**Stimuli**

The original video sequences of the facial expressions were altered (see Facial manipulation technique section) to produce five different conditions. In the first condition, the whole face was held still and only the rigid head motion was visible (RigidOnly). The remaining four conditions were based on this condition. In the second condition, the rigid head motion and the eyes moved normally (RwE). In the third condition, the eyes and the eyebrows were both added to the head motion (RwEB). The motion of the mouth region (RwMR) was added to the rigid head motion in the fourth condition. Finally, all of these regions were visible in the fifth condition (RwEBMR). The combination of six stimulus conditions (five manipulations and the original video sequence) for nine expressions from six actors/actresses yielded a total of 324 trials.

**Results and discussions**

Overall, the original video sequences of the expressions were well recognized (on average ~75%) despite the absence of a conversational context. These results are similar to those found in other experiments using the same set of video recordings (Cunningham et al., 2004, 2005; Cunningham, Nusseck, et al., 2004). This validates the reliability of the previous results and demonstrates that the face can, even without the neighboring body areas, provide enough information to recognize these facial expressions. The results are examined in more detail in the following sections.

**Effect analysis**

A three-way analysis of variance (ANOVA) was run for the within-participants factors manipulation, expression, and actor. Significant main effects were found for manipulation (\(F(5,45) = 96.14, p < 0.001, \eta^2 = 0.91\)) and expression (\(F(8,72) = 30.22, p < 0.001, \eta^2 = 0.77\)). A marginally significant effect was found for the factor actor/actress (\(F(5,45) = 2.27, p < 0.06, \eta^2 = 0.2\)). These findings show that, overall, the different manipulations altered the video sequences differently, that some expressions were recognized more often than others, and that performance across the different actors/actresses was relatively consistent.

There was a significant interaction between actor/actress and expression (\(F(40,360) = 6.23, p < 0.001, \eta^2 = 0.41\)), indicating that each actor/actress was “good” at some expression and “bad” at others, and the set of preferred expressions differed across actors/actresses. There was also a significant interaction between expression and manipulation (\(F(40,360) = 15.32, p < 0.001, \eta^2 = 0.63\)), indicating that different expressions rely on different facial areas. There was, however, no significant interaction between actor/actress and manipulation (\(F(25,225) = 1.39, p = 0.11, \eta^2 = 0.13\)), indicating that no actor/actress consistently showed any special emphasis on, or avoidance of, a particular facial area. The three-way interaction was also significant (\(F(200,1800) = 2.48, p < 0.001, \eta^2 = 0.22\)), suggesting that there is more than one way to produce a given expression.

**Average recognition performance**

Figure 5 shows the percentage of correctly identified expressions for each manipulation. Note that while these averages are clearly not representative of the individual expressions, they do demonstrate an interesting trend. On average, the performance for the condition in which only the rigid head motion was present (RigidOnly) had the lowest performance, but there was still a recognition rate of 30%. This means that subjects could identify expressions significantly better than blind guessing\(^1\) (\(t(9) = 11.34, p < 0.001\)).\(^2\) Furthermore, each area seems to

![Figure 5. Recognition performance in percentage for each manipulation. The error bars depict the standard error of the mean. The meaning of the abbreviations can be found in Table 2.](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933530/)
contribute some information, with the mouth region being the most critical. Finally, the condition in which all four regions were present (RwEBMR) is slightly but significantly lower than the original condition ($t(9) = 2.69, p < 0.03$). This suggests that there are still some expressions that require additional information. Interestingly, Cunningham et al. (2005) did not find this difference. Since their video sequences contained visible non-facial motion, it is possible that these areas can help to compensate for missing information in other facial regions.

**Analysis of individual expressions**

Recognition performance averaged over all expressions provides a general systematic description of the influence of the particular parts. To find out more about how important these facial areas are for the individual expression, we will perform an analysis for each expression. We will do this using the following three analyses:

1. **Recognition performance for each expression**: As can be seen in Figure 6, the manipulations differentially affected the expressions. A detailed look at the individual conditions can highlight the relative importance of the different areas for an expression.

2. **Confusion matrix**: The confusion matrix, which presents systematic interpretation errors, allows us to determine which facial areas are related to which expression, both for the displayed expression and the expression chosen by the participants. The patterns of confusion for the original conditions of all experiments are similar to each other and also to those found in previous works. Participants frequently used the none-of-the-above option, suggesting that other errors were not forced but reflect actual confusions.

![Figure 6. Recognition accuracy for each expression.](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/933530/ on 10/26/2018)
3. **Qualitative behavior description:** Determining that an individual facial part played a role in the recognition of an expression does not indicate what that part was doing. For example, the **agreement** and the **disagreement** expression both rely heavily on head motion but actual head motions are very different. Thus, we qualitatively collected all motions that are visible in the face of each actor/actress. This analysis, which was largely absent in Cunningham et al.’s (2005) original work, helps to interpret and deepen the findings since it not only contains the presence or absence but also detailed descriptions of the movements in the faces.

**Happiness**

As expected, **happiness** clearly depends on mouth motion. Rigid head motion (Rigid only), the eyes (RwE), and the eyebrows (RwEB) contributed little to the recognition of this expression, whereas the conditions in which the mouth region was involved (RwMR and RwEBMR) were nearly equal to the original condition. Head motion was barely used by the actors/actresses, but if they moved their head, they tended to tilt it just a little bit to the side.

**Thinking**

For **thinking**, the primary region seems to be the eyes. Every condition in which the eyes were present (RwE, RwEB, and RwEBMR) reached the accuracy of the original sequences. These findings are consistent with both anecdotal descriptions of thinking and previous studies (Cunningham et al., 2005; Wallraven, Fischer, Cunningham, Bartz, & Bülthoff, 2006). Nevertheless, it still remains unclear exactly how the eyes move. The individual gaze behavior of the actors/actresses differs considerably. Some look up and to the right where others look up and to the left. All of them, however, avert their gaze and look up. This special eye motion behavior seems to be sufficient to recognize this expression.

Interestingly, in some conditions, **happiness** and **sadness** are often misinterpreted as thinking. This response did not appear in the confusion matrices of the original and the combined condition (RwEBMR) nor did it appear in the mouth region condition (RwMR) for **happiness**. In other words, as long as the primary regions for **happiness** or **sadness** are moving, they are not mistaken as **thinking**. If, however, little or nothing moves, these expressions are confused for **thinking**. The mistake seems to derive from the eye motion or rather the lack thereof. For both **happiness** and **sadness**, the eyes did not move but stared directly at the cameras. It seems that staring straight at someone can be a clue for a thinking behavior as well.

**Surprise**

The recognition of **surprise** seems to depend mostly on mouth motion. Many studies (Constantini et al., 2005; Ekman, 2004; Pelachaud & Poggi, 2002) claimed that **surprise** is expressed through a combination of eyebrow, eyes, and mouth motion, mostly due to a rapid opening of the eyes and raised eyebrows. Here, although all but one actor performed this special eye widening behavior, the eyes and the eyebrows do not seem to be sufficient nor required. This is consistent with recent results from Smith, Cottrell, Gosselin, and Schyns (2005) using static displays. Most of the actors/actresses move their head rapidly backward. Whether it is the mouth motion alone or the combination with head motion that is critical will be explicitly tested in Experiment 3.

**Clueless**

The results for **clueless** are quite similar to **thinking**: Eye motion seems to be sufficient. The most common eye behavior, a widening of the eyes, is quite similar to **surprise** and possibly explains why **surprise** was often mistaken for **clueless**. The **surprise** confusion matrices for the eyes (RwE) and the eyebrows (RwEB) conditions confirm that there are more ratings for **clueless** than for the **surprise** expression.

It should be noted that these results differ from the findings of Cunningham et al. (2005). Without the dominating non-facial areas, the recognition of the original sequences dropped from ~80% down to ~50% and allowed us to determine which facial regions are critical. This suggests that while the eyes provide a fair amount of information for this expression, full recognition requires the non-facial areas.

**Agreement**

The recognition accuracy of **agreement** is very high in all conditions, reflecting a strong reliance on rigid head motion. Nearly all actors/actresses moved their head rapidly and uniformly up and down. It is hard to tell what the other areas contribute to the recognition due to a possible ceiling effect.

**Disagreement**

Like **agreement**, this expression shows a very high recognition rate for all conditions. In contrast to **agreement**, the motion of the head is not clearly definable. Sometimes the motion is fast, other times it is slow. It is always, however, a left to right motion.

Interestingly, performance in the original sequence and the condition in which all parts were present (RwEBMR) was slightly but significantly lower than in the rigid only condition ($t(9) = 2.23, p < 0.05$). The head motion seems to clearly determine this expression, but as soon as there is information from other parts it gets harder to recognize it as **disagreement**. A closer look at the individual actors/actresses revealed that some of them knitted their eyebrows together and opened their mouth some (the upper lip rapidly moved upward). This behavior could be interpreted as related to a more confused expression, and indeed the confusion matrix shows that this expression was sometimes labeled as **confusion**. Moreover, **clueless** and **confusion** are systematically misinterpreted as **disagreement**. For both of...
these expressions, this misinterpretation is bigger in the rigid only condition than in the other conditions. The rigid head motion present in the clueless and confusion expressions is very different for each actor/actress. Some tilt their head a bit to one side, some move it back or forward, whereas others made combination of both. Rarely can a clear right/left pattern be seen. This suggests that a statement of disagreement does not necessarily have to be related to a clear right/left movement.

The pattern of results for the last three expressions is more complex. In all three expressions, the rigid head motion provides only minimal information. The addition of the eyes (RwE) and the eyebrows (RwEB) increases the performance but still does not reach the original condition. Similarly, the performance for the mouth motion (RwMR) is better but not as good as the original sequence. This suggests that no individual region is fully sufficient. The combination of all four types of motion together, however, does support accurate recognition performance.

**Sadness**

Surprisingly, only two actresses performed the expected gaze lowering behavior. The others stared straight into the camera. The mouth motion also varied quite a lot. Some of the actors/actresses open it a bit whereas others “pout.” Four of the actors/actresses performed the typical mouth corner down motion, but only in a very small and barely visible way.

**Confusion**

While the eyebrow motion (a rapid jerk to the middle) and the mouth motion (slight raising of the upper lip) are similar for all actors/actresses, the pattern of head and eye motion is quite diverse. Interestingly, the results for confusion differ drastically from the results of Cunningham et al. (2005). The accuracy of all manipulation conditions decreased, on average, by about 20%. This might be due to the missing non-facial areas. Nevertheless, the most surprising difference is the huge performance drop for the eyebrows (RwEB) condition (from ~80% down to ~25%). It is possible that the non-facial parts contain some expression-related information, which only works if the eyebrows and the head motion are present.

**Disgust**

The characteristic rigid head motion (a rapid turning away that was performed by all actors/actresses) is not by itself sufficient. The addition of the eyes (RwE) and the

Figure 7. Recognition performance for the original video sequences for each actor/actress.
eyebrows (RwEB) turns this expression into clueless. The mouth motion is quite similar for all actors/actresses: the mouth opens rapidly so that the closed teeth are visible. Interestingly, participants often chose the happiness option in the mouth only condition.

Actor/actress

As can be seen in Figure 7, while each actor/actress did well on average, none performed all nine expressions well. Agreement, disagreement, happiness, and disgust are the only expressions to show a relatively constant accuracy for all actors/actresses.

Conclusions

Overall, the results are quite similar to Cunningham et al. (2005). In general, three facts stand out. First, all four facial areas seem to contribute to the recognition of the expressions. Second, rigid head motion provides a relatively low amount of information overall. Third, the combination of all four areas (RwEBM) is almost as good as the original condition. Unlike Cunningham et al. (2005), the combined condition does not reach the performance accuracy of the original condition. This indicates that there are expressions, primarily clueless, which need information from the non-facial areas.

Also consistent with Cunningham et al. (2005), the results show that some expressions seem to rely on only one individual facial part, while other expressions seem to require all four facial regions to be recognized as such. Table 1 summarizes the descriptions from all four experiments. As one might expect, most of these descriptions are quite obvious. For example, happiness seems to be mostly related to mouth motion (Ekman, 2004; Smith et al., 2005), and thinking seems to rely on an eye gaze behavior.

The results clearly show that some motions are present in more than one expression. Shaking the head, for instance, seems to be not only related to disagree, but is also found in the confusion and clueless expressions. Likewise, happiness and disgust have similar mouth motions, and clueless and disgust have similar eye motions. As can be seen in the confusion matrices, some facial motions carry multiple potential meanings, and the intended expression only becomes clear when combined with information from other regions. Just as some motions are found in multiple expressions, some expressions have multiple types of characteristic motion. For example, an upward motion of the eyes is present in nearly all the thinking recordings and seems to be a clear sign for this expression. The pattern of confusions for other expressions, however, shows that staring or looking down was also associated with thinking.

A facial expression can be seen as a conglomerate of a variety of different more or less subtle motions spread over the whole face. As Experiment 1 showed, there are dominant, obvious features. There are also some observable movements in other, non-traditional areas. Pelachaud and Poggi (2002) suggest, however, that we can only perceptually distinguish head, eyebrow, eyes, and mouth motions in facial expressions. They also say, more anecdotally, that the nose is a critical component because we are able to wrinkle it, but we do so rather rarely in facial expressions.

In this experiment, we inverted the manipulation used in Experiment 1: Everything in the face moves normally and the “canonical” facial regions were systematically frozen. If recognition is still possible when all the canonical facial areas are frozen, then there must be some information in non-canonical areas that support the recognition of these expressions. Furthermore, the experiment examined the perceptual necessity of the canonical regions.

Method

The task, procedure, design, and equipment are the same as in the previous experiment.

Stimuli

In the first condition, everything moved normally with the exception of the eyes, which were replaced by a static, neutral texture (AeE—all except eyes). The second condition froze the eyebrows and the eyes (AeEB). For a finer grained manipulation than was used in the previous experiment, we added a new mouth area condition: In the third condition, we froze the mouth so that only the lips (and enclosed areas) were frozen but adjacent areas (such as the cheeks, the jaw, and the parts underneath the nose) were still moving (AeL). In the fourth condition, the mouth and its surrounding areas were frozen, so that everything was moving except the mouth region (AeMR).

The last condition was, as in Experiment 1, the combination of all parts. In other words, everything moved normally except the eyes, the eyebrows, and the mouth region (AeEBMR). The combination of six types of video sequence, nine expressions, and six actors/actresses yielded 324 trials.

Results and discussions

Effect analysis

The three-way ANOVA showed significant main effects for manipulation ($F(5,45) = 35.38, p < 0.001, \eta^2 = 0.79$),
expression ($F(8,72) = 16.76, p < 0.001, \eta^2 = 0.65$), and actor/actress ($F(5,45) = 9.16, p < 0.001, \eta^2 = 0.51$). There was a significant interaction between actor/actress and expression ($F(40,360) = 6.99, p < 0.001, \eta^2 = 0.44$) as well as between expression and manipulation ($F(40,360) = 11.01, p < 0.001, \eta^2 = 0.55$) but no interaction for the actor/actress and manipulation interaction ($F(25,225) = 1.78, n.s., \eta^2 = 0.17$). These effects are basically the same as found in Experiment 1.

**Average recognition performance**

Just as in Experiment 1, the original video sequences of the expressions were well recognized (on average ~ 79%). A closer look at the performances for the manipulations across all expressions (Figure 8) shows a few general trends. First, freezing the eyes (AeE) yielded a significant drop in the recognition rate ($t(9) = 4.75, p < 0.001$), as did the removal of lip motion (AeL; $t(9) = 6.72, p < 0.001$). The performance in this condition does not differ significantly from the eyes (AeE) and the eyebrows (AeEB) conditions ($t(9) = 0.97, n.s.$). Freezing the mouth and its surrounding regions (AeMR) produced a drastic reduction in recognition performance ($t(9) = 5.72, p < 0.001$), indicating that there seems to be a lot of critical expression-related information in the parts close to the mouth. Freezing all three facial areas frozen (AeEBMR) reduced accuracy by half, clearly showing that the canonical areas are necessary for some expressions. If, however, these areas were the only information carrying locations, then performance here should be the same as the RigidOnly condition in Experiment 1. Since Rigid-Only is much lower than the AeEBMR, it seems that the non-canonical areas carry expression-related information for some expressions.

**Analysis of individual expressions**

This section examines the individual expressions (see Figure 9 for accuracy rates) in more detail using the three analysis tools (see Analysis of individual expressions section). Since the original sequences were identical to those used in Experiment 1, it is worth noting that both recognition rates and confusion matrices are also very similar across the two Experiments.

**Thinking**

The thinking expression showed a large performance drop when the eyes were frozen (AeE). When the eyes were intact but the mouth motion was blocked (AeL and AeMR), recognition was similar to the original sequence. This suggests that eye movements are indeed not only sufficient but also necessary for this expression.

**Happiness**

When the eyes and the eyebrows were frozen (AeE and AeEB), performances for happiness remained similar to the original sequence. When lip motion is removed (AeL), accuracy fell only by a small amount. Removal of the mouth and its surroundings (AeMR), however, produced a large drop in performance. Since the surrounding mouth regions are heavily influenced by the mouth motion itself, their motions could possibly carry some hints for some original mouth motions, e.g., smiling cheeks need motions from the mouth.

Surprisingly, the accuracy of the condition in which everything moves except the three major areas (AeEBMR) was quite high. This suggests that some non-canonical facial areas carry happiness-related information (e.g., the nose regions or some wrinkles below the eyes). Furthermore, the equivalent performance between the condition where just the mouth region was frozen (AeMR) and where all major regions were frozen (AeEBMR), combined with the lack of an effect for the eyes and eyebrows in Experiment 1, shows that the eyes and the eyebrows do not provide any useful information for this expression.

**Agreement and disagreement**

The results for these two expressions are extremely similar to Experiment 1.

**Surprise**

Freezing the eyes and the eyebrows (AeE and AeEB) did not alter performance for surprise. Likewise, replacing the lips (AeL) did not alter performance. Only the blocking of the mouth region (AeMR) reduced accuracy, albeit only by a small amount. This suggests that the mouth itself is not necessary, but the neighboring regions might be. For most of the actors/actresses, the only motion observable in the mouth region is that of the chin, which typically moved downward a small amount. It seems that this motion (together with the head motion) provides a fair amount of information for this expression. Thus, the mouth region seems to be sufficient but not necessary. Surprisingly, performance when all three areas are frozen (AeEBMR) showed the highest accuracy of all conditions. It seems that freezing the mouth and the surrounding regions removed the possibility for misinterpretation.
Clueless

Performance when the eyes are frozen (AeE) is surprisingly high, but performance drops a lot when the eyebrows are frozen (AeEB). This, together with Experiment 1, suggests that the eyes are sufficient but not necessary whereas the eyebrows are necessary but not sufficient. The extremely low recognition rates when all three parts were frozen (AeEBMR) suggest that rigid head motion and the non-canonical facial regions do not carry sufficient information.

Confusion

Similar to clueless, the eyes are not required but the eyebrows are. Unlike clueless, freezing the mouth and the mouth regions (AeL and AeMR) also strongly reduced performance. Performance was again lowest when all parts were frozen (AeEBMR).

Sadness

Recognition performance for sadness varies little across the conditions. The eyebrows and the mouth do not seem to be required. Only when all regions are frozen (AeEBMR) does performance drop. Overall, the combination of canonical areas is important, but there is some partially sufficient information in the non-canonical regions.

Disgust

Neither the eyes nor the eyebrows are necessary. Only the freezing of the lips (AeL) and the mouth region (AeMR) produced a noticeable reduction in recognition rates. The very low accuracy in the condition in which all parts are frozen (AeEBMR), however, suggests that there is important information in the interaction of these parts.

Conclusions

Overall, freezing all three facial areas reduced recognition performance by nearly half. There was also a general, but small, difference between this condition and the condition in Experiment 1 where only rigid head motion was present.
(RigidOnly), indicating that there is some expression-related information in the non-canonical facial areas.

For some expressions, the areas found in Experiment 1 to be sufficient were also necessary. For other expressions, the sufficient areas were not necessary. Finally, as was found in Experiment 1, some expressions rely on a combination of regions (this is the case for sadness, confusion, and disgust). For these expressions, only the joint information when all facial areas are present is sufficient and is also necessary.

**Experiment 3**

The findings of the previous experiments were based on stimuli in which the original rigid head motion was always present. Some expressions were still perfectly recognizable if only the head motion was present. This suggests that the rigid head motion does contain at least some expression-related information. Moreover, since it was present in each sequence and might have interacted with whatever facial motion might have been present, it was not possible to precisely determine the relative contribution of the individual facial regions. Thus, in this experiment, we eliminated the rigid head motion and repeated Experiment 1.

**Method**

The task, procedure, design, and equipment are the same in the previous experiments.

**Stimuli**

The stimuli were identical to the first experiment, with one exception: In all manipulation conditions, the virtual camera was fixed to the head in the 3D reconstruction, effectively removing rigid head motion. In the first condition, we let everything in the face move normally (noRigid). For the other conditions, the face was completely replaced with a neutral picture of the actor/actress and certain facial areas were allowed to move. The second condition therefore had the eyes (no rigid motion, with eyes—noRwE). The third condition was with eyes and eyebrows (noRwEB). The fourth condition was with the mouth region (noRwMR). In the last condition, the combination of these regions was unfrozen (noRwEBMR).

**Results and discussions**

**Effect analysis**

The results of the ANOVA were similar to Experiments 1 and 2 with significant main effects for manipulation 

\[(F(5,45) = 170.57, p < 0.001, \eta^2 = 0.95),\]  

expression \[(F(8,72) = 18.19, p < 0.001, \eta^2 = 0.67),\]  

and actor/actress \[(F(5,45) = 9.64, p < 0.001, \eta^2 = 0.52).\]

Similar to the previous experiments, there were significant interactions between actor/actress and expression \[(F(40,360) = 6.35, p < 0.001, \eta^2 = 0.42)\]  

as well as between expression and manipulation \[(F(40,360) = 14.75, p < 0.001, \eta^2 = 0.62),\]  

but not for the actor/actress and manipulation interaction \[(F(25,225) = 1.33, \text{n.s.}, \eta^2 = 0.13).\]

The three-way interaction was also significant \[(F(200,1800) = 2.17, p < 0.001, \eta^2 = 0.19).\]

**Average recognition performance**

The original video sequences of the expressions were well recognized (≈75%). On average, removal of the head motion reduced recognition rates a small but significant amount (see Figure 10; \(t(9) = 5.89, p < 0.001).\) When only eye motion was present (noRwE), few expressions were recognized correctly. In fact, performance was not significantly different from random guessing \((t(9) = 2.01, \text{n.s.}).\) This divergence from the pattern of results found in Experiment 1 suggests that it was not the eye motion that was important in Experiment 1 but the combination of eye and rigid head motion. Adding eyebrow motion to the recordings (noRwEB), increased performance significantly \((t(9) = 3.73, p < 0.005).\) This is also different from Experiment 1. While this also reinforces the suggestion that the constant presence of rigid head motion was masking the true effect of the different facial areas, it also reinforces the conclusion from Experiment 2 that the eyebrows do carry expression-relevant information. The mouth condition (noRwMR) also showed relatively high performance. The presence of all three facial regions (noRwEBMR) yielded accuracies that were high but still lower than presence of everything except rigid head motion \((t(9) = 5.02, p < 0.001).\) This difference must be due to non-canonical facial areas.

Although none of the conditions reached the recognition level found for the equivalent conditions in Experiment 1

---

**Figure 10. Results of the third experiment.**
(the rates are about 20% lower in the present experiment), the overall pattern of results is very similar.

**Analysis of individual expressions**

Figure 11 shows the recognition rates for each expression. In general, the results are similar to Experiment 1. Not surprisingly, the rigid head motion seems to be the most necessary source for agreement and disagreement. While most of the other expressions showed no significant difference between the original sequences and the condition in which the head was fixed, they did show an effect of rigid head motion in the manipulation conditions. It seems, therefore, that rigid head motion does not directly provide expression-related information but instead functions as an underlying source for the interaction with other facial parts to help clarify the recognition. In the following, we will highlight the major differences to Experiment 1 (note that no critical differences were found for happiness, thinking, sadness, or confusion).

**Agreement and disagreement**

As expected, neither of these expressions is really recognizable without rigid head motion. The fact that the recognition rates are not zero seems to be due to the visible shadow artifacts (which imply the original head motion).

**Surprise**

Performance when only the mouth region was present (noRwMR) was high, confirming the central role of the mouth in this expression. Performance in this condition was, however, much lower than the original sequence, suggesting that rigid head motion and the mouth region interact.

**Clueless**

Unlike Experiment 1, accuracy was not high when only the eyes move (noRwE). The further addition of eyebrows (noRwEB) increases recognition some. Only when these areas are combined with the rigid head motion (RwEB, Experiment 1) or with the mouth motion (noRwEBMR) was recognition performance similar to the original sequences.

Figure 11. Recognition accuracy in percentage for each expression.
Disgust

While the eyes by themselves (noRwE) do not support recognition of this expression, the addition of rigid head motion greatly increases their influence (RwE, Experiment 1). The presence of just eyes and the eyebrows (noRwE and noRwEB) often lead to confusions with clueless.

Conclusions

In this experiment, we investigated the sufficiency of the facial areas in the absence of rigid head motion. In general, rigid head motion was not very dominant by itself but instead played a facilitator role (by interacting with the motion of specific areas). The pattern of facial regions that is sufficient for the recognition of the individual expression was largely the same as for Experiment 1.

Experiment 4

Experiment 3 showed that rigid head motion interacts with the different facial regions in determining the sufficiency of facial areas. Experiment 4 examines the necessity of these areas in the absence of rigid head motion.

Method

The task, procedure, design, and equipment are the same as in the previous experiments.

Stimuli

The stimuli were identical to Experiment 2, with the sole except that rigid motion was removed in all manipulation conditions. In the first condition, everything moved normally except the eyes (noReE). The second condition saw the removal of eye and eyebrow motion (noReEB). The mouth (strictly speaking the lips) and the mouth region were frozen in the third and fourth conditions, respectively (noReL and noReMR). All four areas were eliminated in the last condition (noReEBMR).

Results and discussions

Effect analysis

Consistent with the previous experiment, the ANOVA shows significant main effects for manipulation ($F(5,45) = 102.66, p < 0.001, \eta^2 = 0.92$), expression ($F(8,72) = 7.47, p < 0.001, \eta^2 = 0.45$), and actor/actress ($F(5,45) = 5.01, p < 0.001, \eta^2 = 0.36$). There are also significant interactions between actor/actress and expression ($F(40,360) = 8.26, p < 0.001, \eta^2 = 0.48$) and between expression and manipulation ($F(40,360) = 9.49, p < 0.001, \eta^2 = 0.51$) but just a marginally significant effect for the interaction between actor/actress and manipulation ($F(25,225) = 2.05, p < 0.03, \eta^2 = 0.19$). The three-way interaction was also significant ($F(200,1800) = 2.09, p < 0.001, \eta^2 = 0.19$).

Average recognition performance

The original sequences were still well recognized (~75%), and the overall pattern of responses for each manipulation (see Figure 12) was very similar to Experiment 2, suggesting that the facial areas seem to have the same influence with or without rigid head motion. Recognition rates, however, are generally lower than in Experiment 2, confirming that rigid head motion is needed for some expressions.

The performance in the eyebrows condition (noReEB) was significantly lower than then eyes condition ($t(9) = 4.08, p < 0.003$). This effect was not found in the second experiment, suggesting that the head motion seems to influence the recognition through the eyebrows. The fact that performance was generally lower in the frozen eye motion condition (noReE) than in the condition where everything except the head moved (noRigid, Experiment 3) suggests that eye motion plays a role for some expressions. Finally, very few expressions were correctly recognized when the three canonical areas were frozen (noReEBMR). Performance here was only slightly better than blind guessing ($t(9) = 2.89, p < 0.06$). This suggests that the effect of the non-canonical areas found in Experiment 2 requires the presence of rigid head motion.

Analysis of individual expressions

Figure 13 shows the accuracy results for the different expressions. Since most of the findings are similar to the
previous experiments, we will focus on the major differences. The results for agreement, disagreement, happiness, thinking, confusion, and disgust are as one would expect from the previous experiments.

Clueless

No manipulation condition for the clueless expression reaches the performance of the original sequence. While the eyebrows are relatively important (noReEB), all facial areas seem to contribute a bit. Non-canonical areas seem not to be necessary.

Sadness

The results confirm the suggestion that the high performance in the condition in Experiment 2 where every area was frozen (AeEBMR) was due to rigid head motion.

Surprise

All conditions for the surprise expression showed a relatively consistent recognition level, which is nearly as high as for the original sequence. The suggestion from Experiment 2, that this is related to rigid head motion, seems not to be confirmed. Another possible information source could be the chin. Most of the actors/actresses rapidly opened their mouth, causing the chin to move as well. This motion is present in the scene, even without the mouth being opened.

Conclusions

The absence of rigid head motion did not change the pattern of necessity much. Additionally, it was confirmed that disgust is mostly related to mouth motion, but information from the other facial areas is needed in order for it not to be confused with a smiling mouth belonging to happiness. Sadness is still quite difficult to define, since the freezing of the facial parts did not influence the recognition performance much. This might be due to the duration of this expression. Edwards (1998) and Kamachi et al. (2001) found that some expressions seem to have a
characteristic speed behavior, e.g., sadness was better recognized in slower than in faster sequences. Long staring at someone seems to be an indicator of a sad state. On the other hand, this expression was also quite often misinterpreted as thinking, for which the staring behavior seems to be also a good information source.

**General conclusions**

In a series of four experiments using video sequences of real people, we systematically altered the presence or the absence of motion in specific facial regions to investigate their relative contribution to recognition of nine dynamic, conversational expressions. To keep the number of stimuli down to a manageable level, we focused mainly on the three canonical facial expression areas: the eyes, the eyebrows, and the mouth. Each of these areas was replaced with the appropriate static texture from a neutral expression. The first two experiments always contained rigid head motion, while rigid head motion was completely removed for the last two experiments. In the first and third experiment, we examined the sufficiency of the facial regions by freezing the entire face and systematically allowing specific facial regions to move. In the second and the fourth experiment, we investigated the necessity of these facial regions by selectively freezing them and allowing the rest of the face to move normally. The video sequences were edited to remove all non-facial regions, such as ears, neck, shoulders, and hair. Despite the total absence of familiarity or conversational context, the expressions were easily recognized.

By combining three different forms of analysis, we attempted to provide a systematic description of the necessary and sufficient components of these expressions. First, we examined the pattern of correct responses. Second, we looked at the pattern of confusions. This not only allowed us to determine which expressions contained related information but also provided some insights into alternate ways to specify specific expressions. Finally, we collected a qualitative description of the visual deformations for each actor/actress for each expression in order to see what types of motion were occurring in the different facial regions.

Consistent with previous work, we found that some expressions rely on a single facial region, while other expressions required the joint motion of several areas. Table 1 provides a summary of the findings of all experiments split into the sufficient and necessary components. It provides an abstracted, qualitative description of the behavior of the different facial areas. Note that this

<table>
<thead>
<tr>
<th>Expression</th>
<th>Sufficient for recognition</th>
<th>Necessary for recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree</td>
<td>Rigid head motion</td>
<td>Rigid head motion</td>
</tr>
<tr>
<td></td>
<td>Quick up and down with small amplitudes</td>
<td>Quick up and down with small amplitudes</td>
</tr>
<tr>
<td>Disagree</td>
<td>Rigid head motion</td>
<td>Rigid head motion</td>
</tr>
<tr>
<td></td>
<td>A left/right motion of the head</td>
<td>The appropriate motion is not clear. A tilting to the side or to the front can also trigger this expression</td>
</tr>
<tr>
<td>Happy</td>
<td>Mouth motion</td>
<td>Mouth motion including surrounded regions</td>
</tr>
<tr>
<td></td>
<td>Lip corners go upward and baring teeth</td>
<td>Smiling cheeks, squinting the nose and other parts close to the mouth</td>
</tr>
<tr>
<td>Sad</td>
<td>Eyes, eyebrows, and mouth</td>
<td>Eyes, eyebrows, and mouth</td>
</tr>
<tr>
<td></td>
<td>The eye gaze goes down, the eyebrows knit, the mouth corners go down a bit</td>
<td>Same interactions</td>
</tr>
<tr>
<td>Confusion</td>
<td>Eyes, eyebrows, and mouth</td>
<td>Eyes, eyebrows, and mouth</td>
</tr>
<tr>
<td></td>
<td>The eyes squint a bit; the eyebrows go together; the mouth opens a bit</td>
<td>Same interactions</td>
</tr>
<tr>
<td>Thinking</td>
<td>Eyes</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>The eye gaze goes up and left/right or just a staring at the listener</td>
<td>Doing nothing can trigger this expression as well</td>
</tr>
<tr>
<td>Clueless</td>
<td>Eyes, eyebrows, and mouth</td>
<td>Eyebrows</td>
</tr>
<tr>
<td></td>
<td>The eyebrows rise rapidly; the eyes open a bit; the mouth opens to a small slit</td>
<td>The eyebrows rise rapidly</td>
</tr>
<tr>
<td>Surprise</td>
<td>Mouth motion</td>
<td>Possibly the chin</td>
</tr>
<tr>
<td></td>
<td>A small, rapid opening of the mouth,</td>
<td>The rapid mouth opening causes the chin to go quickly downward</td>
</tr>
<tr>
<td>Disgust</td>
<td>Eyes, eyebrows, and mouth</td>
<td>Eyes, eyebrows, and mouth</td>
</tr>
<tr>
<td></td>
<td>The eyes squint; the eyebrows go together; the mouth opens to bare the teeth</td>
<td>Same interactions</td>
</tr>
</tbody>
</table>

Table 1. Summary of the facial areas involved in recognizing these expressions. The table provides a collection of qualitative descriptions of the motion of the facial areas for most of the actors/actresses.
Our findings show that for agreement and disagreement, rigid head motion is both sufficient and necessary. For agreement, the head motion (tilting up and down) is clear and unambiguous: It was never misinterpreted as any other expression. The head motion for disagreement, however, was not that obvious. Although a left/right head motion made this expression clearly recognizable, other rigid head motions such as tilting the head to the side or to the front also seemed to imply disagreement. The expressions happiness and surprise seem to be primarily specified through mouth motion. Whereas the typical upward motion of the lip corners seems to be sufficient and necessary for the happy expression, a rapid mouth opening for a surprised gesture is sufficient, but not necessary. To perform a thinking expression, the eyes are sufficient. Several types of eye motion seemed to be used. First, one could look away (generally upward). One might also stare straight ahead and remain very still. Since this complete lack of motion is also sufficient to specify thinking, no area can truly be claimed to be necessary. Confusion, sadness, and disgust all require a more or less complex interaction of several kinds of motion. Clueless proved to be rather interesting in that Cunningham et al. (2005) had previously suggested that this expression was strongly influenced by shoulder motion. We found that, even in the absence of shoulder motion, this expression can still be well recognized. As long as rigid head motion is present, eye motion seems to be sufficient but not necessary, whereas eyebrow motion seems to be necessary but not sufficient. This changes rather drastically when no rigid head motion is present. In the absence of rigid head motion, no single area was sufficient, but all regions were somewhat necessary. The expression was only recognizable through the combination of all inner facial movements.

Baron-Cohen et al. (1997) suggested that, in contrast to emotional expressions, expressions of “complex” mental states were less recognizable through the mouth than through the eyes (which included the eyebrows) and that they retain a privileged position in the recognition. We found, however, that only by the interaction between both regions were these expressions properly recognized, and that emotional and non-emotional expressions were recognized equally well.

For all expressions except for agreement and disagreement, the combination of the three canonical facial parts is sufficient to produce normal recognition rates. This confirms the generally held belief that these facial areas are indeed the most informative regions for the recognition of these expressions. Nevertheless, a few expressions (namely, happiness, sadness, and surprise) can still be recognized, albeit not often, in the complete absence of motion in these regions. It is possible that motion in non-canonical regions is being used to infer the motion in the canonical areas (either because the latter directly causes the former, or they are merely well correlated).

Interestingly, sometimes two expressions share highly similar motions, differing only in one single region. It seems that there are only a few major types of motion for any given facial area, which are by necessity shared.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>The original video sequence</td>
</tr>
<tr>
<td>Rigid only</td>
<td>The face is completely frozen but the head motion</td>
</tr>
<tr>
<td>RwE</td>
<td>Nothing moves except the rigid head motion and the eyes</td>
</tr>
<tr>
<td>RwEB</td>
<td>Nothing moves except the rigid head motion, the eyes, and the eye brows</td>
</tr>
<tr>
<td>RwMR</td>
<td>Nothing moves except the rigid head motion and the mouth region</td>
</tr>
<tr>
<td>RwEBMR</td>
<td>Nothing moves except the rigid head motion, the eyes, the eyebrows, and the mouth region</td>
</tr>
<tr>
<td>AeE</td>
<td>Everything (all) move except the eyes</td>
</tr>
<tr>
<td>AeEB</td>
<td>Everything (all) move except the eyes and the eyebrows</td>
</tr>
<tr>
<td>AeL</td>
<td>Everything (all) move except the mouth (almost the lips)</td>
</tr>
<tr>
<td>AeMR</td>
<td>Everything (all) move except the mouth region</td>
</tr>
<tr>
<td>AeEBMR</td>
<td>Everything (all) move except the eyes, the eyebrows, and the mouth region</td>
</tr>
<tr>
<td>noRigid</td>
<td>Everything move in the face but the head is fixed</td>
</tr>
<tr>
<td>noRwE</td>
<td>Nothing moves, even no rigid head motion, except the eyes</td>
</tr>
<tr>
<td>noRwEB</td>
<td>Nothing moves, even no rigid head motion, except the eyes and the eyebrows</td>
</tr>
<tr>
<td>noRwEBMR</td>
<td>Nothing moves, even no rigid head motion, except the eyes, the eye brows, and the mouth region</td>
</tr>
<tr>
<td>noReE</td>
<td>Everything move except the rigid head motion and the eyes</td>
</tr>
<tr>
<td>noReEB</td>
<td>Everything move except the rigid head motion, the eyes, and the eyebrows</td>
</tr>
<tr>
<td>noReL</td>
<td>Everything move except the rigid head motion and the mouth (almost the lips)</td>
</tr>
<tr>
<td>noReMR</td>
<td>Everything move except the rigid head motion and the mouth region</td>
</tr>
<tr>
<td>noReEBMR</td>
<td>Everything move except the rigid head motion, the eyes, the eye brows, and the mouth region</td>
</tr>
</tbody>
</table>

Table 2. Abbreviation table for the four experiments.
between expressions. The specification of distinct expressions, then, lies in the combination of these motions. For example, one actor performed the expressions of confusion and disgust with nearly the same squinting of the eyes and a tilting of the head. For disgust, the head was tilted forward and for confusion, it was tilted backward. Only the interaction of these two signals resulted in clearly recognizable expressions. Likewise, disgust and happiness seem to be related. Disgust was often confused for happiness, especially when only the mouth region was allowed to move (with or without rigid head motion). Many of the actors/actresses stretched their mouth open in the disgust expression so that their teeth were visible. Participants probably considered this motion to be similar to the smile found in some happy expressions.

Some facial motions, especially in combination with others, actually reduced the recognizability of an expression. This, for instance, was shown for disagreement where pure rigid head motion (freezing the entire face) led to higher recognition performance than the original video sequences. Similar patterns can be seen for happiness and disgust, where some of the original motions detracted from good recognition performance. This indicates that “normal” or original sequences are not always the best or clearest exemplars of an expression. It also shows that one must make a distinction between what does move in a given recording of an expression and what must or should move for that expression to be recognized. This point is reinforced by the surprised expression. Traditionally, surprised is described as an upward motion of the eyebrows, an opening of the mouth, and a rapid backward motion of the head. While these motions did occur in many of the video sequences (especially the eyebrow motion), our results show that the mouth motion presented by itself is perceptually sufficient, and that the addition or subtraction of the eyebrow motion did not influence recognition performance much. Having determined which areas are perceptually critical for certain expressions, future work will focus on a more detailed description of the perceptually critical spatiotemporal properties of the motion in these areas. The first step would be to acquire a more exacting description of the motion in these areas in each sequence. This can be done either by manually applying an existing descriptive system (e.g., FACS) or by a semi-automatic categorization of the 3D motion trajectories of the surface regions (as recovered, for example, from motion capture recordings). This later has the advantage that, in a further step, one can create highly realistic, targeted computer animated expressions containing systematic variations on the actual motions (for a similar approach, see Hill, Troje, & Johnston, 2005).

In summary, the results provide new insights into how we use the motion and the deformation of different facial areas (or “perceptual units”) to recognize various conversational expressions. The pattern of participants’ responses indicates which facial areas must be present for specific facial expressions to be recognized (i.e., are necessary) and which areas are enough for an expression to be recognized (i.e., are sufficient). This combined with a qualitative description of the specific pattern of motions present in those areas allowed us to systematically specify the perceptual composition of nine specific expressions. Although dynamic facial expressions involve an intricate synchronization of various changes and transformations, humans are amazingly adept at quickly recognizing the intent behind a given expression, even when the speaker is unknown and the expression is shown completely out of context. Interestingly, despite the amount and complexity of information available in dynamic facial expression, most expressions seem to be adequately specified by the interaction between a few major types of motion. The speed and ease with which not only emotional expressions but also conversational expressions are processed is no doubt critical to their ability to assume a central role in many different aspects of a conversation.

These results can be considered as preliminary guidelines and recommendations for the development of facial animations, particularly in a human–computer interaction context. For instance, one goal in creating facial animations for natural and believable looking human agents is to render facial expressions as effectively and as realistically as possible with little computational cost. The list of what is sufficient and necessary for recognizing conversational facial expression might help to focus on important information sources in the synthesis process and guide future developments in this area.

Acknowledgments

This work was supported by the European Network of Excellence ENACTIVE Interfaces (FP6IST 002114) [http://www.enactivenetwork.org].

Commercial relationships: none.
Corresponding author: Manfred Nusseck.
Email: manfred.nusseck@tuebingen.mpg.de.
Address: Spemannstrasse 38, 72076 Tübingen, Germany.

Footnotes

1The expected performance level for purely random or blind guessing in this experiment is 10%.
2The analysis was a two-tailed dependent measures t test (α = .05). Further t tests in this paper are always done in this manner.
3A non-parametric Wilcoxon test to compare if the recognition of the non-rigid conditions of all expression (exclusive of agreement and disagreement) are lower than the conditions with head motion shows a significant difference (p < 0.001, z(62) = 7.81).
References


Hanawalt, N. (1944). The role of the upper and lower parts of the face as the basis for judging facial expressions: II. In posed expressions and “candid camera” pictures. Journal of General Psychology, 31, 23–36.


