Dynamic information for the recognition of conversational expressions

Douglas W. Cunningham

Max Planck Institute for Biological Cybernetics, Tübingen, Germany, &
Department for Graphical Systems, Brandenburg Technical University, Cottbus, Germany

Christian Wallraven

Max Planck Institute for Biological Cybernetics, Tübingen, Germany, &
Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

Communication is critical for normal, everyday life. During a conversation, information is conveyed in a number of ways, including through body, head, and facial changes. While much research has examined these latter forms of communication, the majority of it has focused on static representations of a few, supposedly universal expressions. Normal conversations, however, contain a very wide variety of expressions and are rarely, if ever, static. Here, we report several experiments that show that expressions that use head, eye, and internal facial motion are recognized more easily and accurately than static versions of those expressions. Moreover, we demonstrate conclusively that this dynamic advantage is due to information that is only available over time, and that the temporal integration window for this information is at least 100 ms long.

Keywords: facial expressions, dynamic information, communication, perception

Citation: Cunningham, D. W., & Wallraven, C. (2009). Dynamic information for the recognition of conversational expressions. *Journal of Vision, 9*(13):7, 1–17, http://journalofvision.org/9/13/7/, doi:10.1167/9.13.7.

Introduction

Communication is a central aspect of everyday life, a fact that is reflected in the wide variety of ways that people exchange information—with words, pictures, and even using their face and body. The face and body are potentially very powerful channels for communication. Facial information[1] plays a number of roles, either alone or in combination with other communication channels. For example, it can play a critical role in directing conversational flow, particularly turn-taking (Bavelas, Black, Lemery, & Mullett, 1986; Bavelas, Coates, & Johnson, 2000; Bull, 2001; Cassell, Bickmore, Cambell, Vilhjalmsson, & Yan, 2001; Cassell & Thorisson, 1999; Isaacs & Tang, 1993; Poggi & Pelachaud, 2000; Vertegaal, 1997; Yngve, 1970). A speaker can inform a crowd of listeners at whom a question is directed by a properly directed change in gaze. Listeners can provide a wealth of information to the speaker without ever saying a word (this is referred to as "back-channel" signals; Yngve, 1970). For example, a properly timed nod of agreement can tell the speaker to continue speaking, while a look of confusion at the same point in time indicates that the speaker should stop and try to explain the last point again. Such signals can also be used to indicate the location and intensity of prosodic emphasis (Nusseck, Cunningham, De Ruiter, & Wallraven, under review). Facial information can additionally serve as deictic gestures, indicating what the referent for a spoken statement is (e.g., looking at a particular desk when saying "Please place it over there."). It can also be used to directly modify the meaning of spoken sentences: A spoken statement of surprise combined with a neutral expression is fundamentally different from the same statement accompanied by a surprised expression. Indeed, in situations where the meaning conveyed by the face differs from that in another communication channel, the face tends to be considered more important (Carrera-Levillain & Fernandez-Dols, 1994; Fernandez-Dols, Wallbott, & Sanchez, 1991; Mehrabian & Ferris, 1967). Finally, facial information can independently signify meaning. This includes not only traditional emotional expressions such as happiness or anger, but also more "cognitive" or conversational expressions such as thinking, agreement, confusion, or cluelessness (e.g., Baron-Cohen, Wheelwright, & Jolliffe, 1997; Cunningham, Kleiner, Wallraven, & Bülthoff, 2005; Nusseck, Cunningham, Wallraven, & Bülthoff, 2008; Pelachaud & Poggi, 2002).

A growing body of literature is dedicated to the study of the production and perception of facial expressions. The majority of work in this field has focused on emotional

doi: 10.1167/9.13.7 Received March 12, 2009; published December 7, 2009 ISSN 1534-7362

expressions, usually selected from the seven "universal" expressions (these are happiness, sadness, fear, anger, disgust, contempt, and surprise according to Ekman, 1972). Few experiments have examined the other forms of expression, many of which play a more central role in normal conversations. For example, someone is much more likely to express agreement, thinking, or confusion than fear, contempt, or anger in a normal conversation. Moreover, the absence of conversational expressions in a face-to-face dialog strongly affects the quality of the conversation by impairing comprehension and conversational flow. Thus, the present experiments focus primarily on conversational expressions (for other work with conversational expressions, see Baron-Cohen et al., 1997, Boyle, Anderson, & Newlands, 1994; Cunningham et al., 2005; Nusseck et al., 2008; Stephenson, Ayling, & Rutter, 1976; Wallraven, Breidt, Cunningham, & Bülthoff, 2008).

A further tendency in expression research is to focus on static photographs, generally using only the apex or peak of an expression. Faces in the real world, however, are rarely static. Since our visual system has learned to detect and decode expressions in dynamic situations, it is reasonable to assume that it is optimized for dynamic expressions. While it might be tempting to argue that any dynamic situation is merely a collection of static snapshots, there is considerable evidence that some information is only available over time[2] (Gibson, 1979). Consistent with this, several studies point to the fact that static and dynamic facial expressions are processed, at least partially, in different brain structures (Adolphs, Tranel, & Damasio, 2003; Humphreys, Donnely, & Riddoch, 1993; LaBar, Crupain, Voyvodic, & McCarthy, 2003; Schultz & Pilz, 2009; Schwaninger, Wallraven, Cunningham, & Chiller-Glaus, 2006). These studies include reports from patients who are completely unable to recognize expressions from static pictures or static descriptions but have normal recognition performance for both dynamic expressions and descriptions of dynamic expressions or actions. Psychophysically, it has been shown that the temporal pattern of facial motion is *sufficient* for recognition of facial expressions (Bassili, 1978, 1979). Bassili placed a series of luminescent makers on the faces of several individuals, and then showed just the markers (i.e., Johansson point-light displays; for more on this experimental technique, see Johansson, 1973). While the stimuli were not even recognizable as faces when only a single frame of the video sequence was shown, the expressions were readily recognized in the dynamic version.

In situations where some static information for facial expressions is available, dynamic expressions still seem to be more easily recognized than static expressions (Ambadar, Schooler, & Cohn, 2005; Harwood, Hall, & Schinkfield, 1999; Wallraven et al., 2008; Wehrle, Kaiser, Schmidt, & Schere, 2000; Weyers, Mühlberger, Hefele, & Pauli, 2006). Although the stimuli used in most of these studies contained more static information than point-light displays, some of the static information was seriously degraded. Some of those studies have explicitly shown that dynamic information can compensate for the loss of static information (Cunningham & Wallraven, 2009; Ehrlich, Schiano, Scheridan, 2000; Kaetsyri, Klucharev, Frydrych, & Sams, 2003; Wallraven et al., 2008). Wallraven et al. (2008), for example, systematically degraded the shape, texture, and motion in a series of computer animated facial expressions. Not only did the dynamic sequences produce higher recognition rates than the static sequences, but the presence of dynamic information completely eliminated the otherwise deleterious effect of degrading either shape or texture information.

Just because one *can* use dynamic information does not, however, mean that one *normally* uses it. Moreover, while a few studies have shown a dynamic advantage for video sequences of real individuals, the dynamic and static stimuli always differed not only in terms of the presence of motion, but also along a number of other dimensions, such as the number of images and facial poses. Thus, it is unclear whether the dynamic advantage is due to motion (or the pattern of changes over time in general) or something simpler. To help determine whether the dynamic advantage is simply due to the presence of more images, Ambadar et al. (2005) compared recognition performance for a static expression, a dynamic version of that expression, and a condition where a 200-ms Gaussian noise mask was interspersed between the frames of the dynamic sequence (thus masking the motion information). The second condition (normal dynamic expression) showed better performance than the other two conditions, which did not differ from each other. While the mask between the frames in the third condition did in fact eliminate the perception of motion, the type of mask used can inhibit the processing of static information (see, for example, the literature on backward masking: Averbach & Coriell, 1961; Kahneman, 1968; Stigler, 1910; Williams, Breitmeyer, Lovegrove, & Gutierrez, 1991; or change blindness: Becker & Pashler, 2002; Rensink, O'Regen, & Clark, 1997; Simons & Levin, 1998). Moreover, the stimuli contained only the early phases of an expression (i.e., the first three to six frames). This means that the pose used in the static condition was not a peak expression and therefore does not necessarily contain all of the static information that is present in other studies with static expressions. Thus, it is not clear that the dynamic advantage found in Ambadar et al.'s experiment would generalize to peak static expressions.

In sum, it is clear that there is some form of characteristic dynamic information for facial expressions, that this information is sufficient for the recognition of facial expressions, and that this information is partially processed in different areas of the human brain than static expressions are. It is also clear that the addition of

dynamic information improves the recognizability of expressions and can compensate for the loss of static information. It is not clear, however, whether the advantage of dynamic stimuli over static, peak expressions is due to some form of information that is only available over time or to some simpler explanation (such as the number of images, different views, poorly selected static frame, etc.). Finally, if information is truly integrated over time for facial expression recognition, the length of the temporal integration window would be of interest. Here we present a series of experiments designed to address these issues conclusively. In Experiment 1, we directly compare dynamic and static peak versions of nine conversational expressions and find that seven of them show a dynamic advantage. Experiment 2 shows that the presence of spatiotemporal information is important and that several simpler, static explanations (based on the number of images present) cannot explain the results. By scrambling the order of the frames in the dynamic expression, Experiment 3 shows that the mere presence of face-appropriate dynamic information is not sufficient; there is some specific information present in the normal temporal development of an expression. Since some forms of dynamic information are non-reversible (e.g., watching in reverse a film of a building exploding is a perceptually odd experience), Experiment 4 examined the effect of playing an expression backward (so that the last frame is seen first, and the first seen last). The results show a small but significant advantage of normal over backward films, indicating a high sensitivity of expression perception to dynamic information. Finally, Experiment 5 estimates the length of the temporal integration window.

Experiment 1: Static versus dynamic

In this experiment, we compared the accuracy with which people can recognize nine conversational expressions in static and dynamic sequences. Following the classification of Baron-Cohen et al. (1997), the expressions used here include both "basic" emotional expressions (e.g., happy, sad) as well as more "complex" expressions (e.g., agreement, confusion, pleasantly surprised). In an attempt to ensure that the results generalize to real-world situations while still retaining accurate knowledge about which expressions the actors and actresses were trying to communicate, the expressions were recorded using a method-acting protocol. The video recordings of the expressions began at the last neutral facial expression before the face started to move and ended at the last frame before the face started to head back to neutral after reaching the peak. The frame used in the static conditions was the last frame of each dynamic sequence (i.e., the "peak" expression).

Methods

A series of video recordings of facial expressions is shown to participants, who were asked to identify the expression using a 10-alternative non-forced-choice task. The methodology and stimuli are the same as those used in Cunningham et al. (2005).

Stimuli

The stimuli consisted of recordings from the Max Planck Institute for Biological Cybernetic's Facial Expression Database, which were recorded using the MPI VideoLab (see Kleiner, Wallraven, & Bülthoff, 2004). The database consists of nine expressions: agree, disagree, happy, sad, clueless (as if the actor or actress did not know the answer), thinking, confused (as if the actor or actress did not understand the question), disgust, and pleasantly surprised. They were recorded from six individuals (4 females, 2 males; 1 professional and 5 amateur actors/actresses). Prior to the arrival of the actors/actresses, a set of scenarios was developed from real-world situations that are known to elicit specific emotions (for example, in normal life, someone who accidentally places their hand in a slimy, smelly piece of food generally produces a disgusted expression). The actors/actresses were seated in front of the cameras and asked to remember a similar situation in their own life, and how they reacted at the time. They were then asked to imagine that they were in that situation and to react normally. They were asked to react as naturally as possible, while refraining from speaking or using their hands. To help avoid potential cultural differences in the production or perception of expressions, we selected actors and actress who were born in Germany and grew up in Germany. In addition, all participants either also grew up in Germany or had lived in Germany for a long time. For more information on the expression database, see Cunningham et al. (2005).

Only the videos from the central camera of the recording setup were presented, so that all videos consisted of a frontal view of the face. The video sequences were edited so that only the face of the person in the video was visible. Each sequence was edited so that it began one frame before the face started to move (and was thus in a neutral expression) and ended on the frame just as the face started to move back to neutral. This last frame was used as the peak frame in the static condition. Previous research with these and similar sequences has shown little or no difference in either identification accuracy or perceived believability between such sequences and sequences that went from a neutral expression through the peak and back to a neutral expression (Cunningham et al., 2005). The length of the videos ranged from 27 to 171 frames. No straightforward correlation between expression and length was found. Sound was not recorded and thus not played during the experiment. The videos were recorded and

shown at 25 Hz. The exposure time of each camera was set to 3 ms in order to reduce motion blur. To help avoid artifacts and unintended information in the recorded sequences, care was taken to light the actor's and actress's faces as flatly as possible. Special effort was devoted to the avoidance of directional lighting effects (cast shadows, highlights). For an example, see Movie 1.

All images were scaled from the original size of 768 × 576 pixels to 256 × 192 pixels. Previous research has shown the two image sizes to produce identical patterns of recognition accuracy and response time (Cunningham, Nusseck, Wallraven, & Bülthoff, 2004). Since the participants sat at a distance of approximately 0.5 m from the computer screen, all images subtended approximately 10 by 7.5 degrees of visual angle.

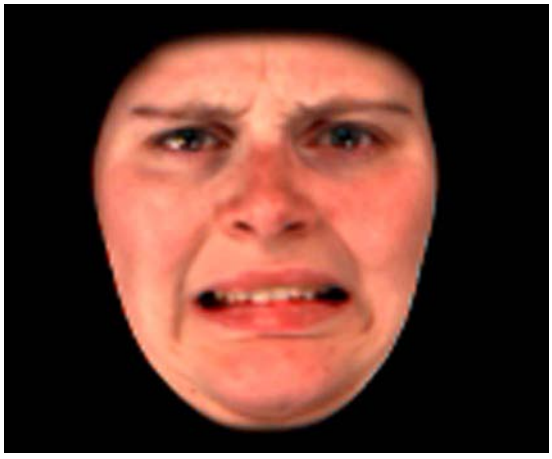
Procedure

Ten individuals participated in the experiment for financial compensation at standard rates. The participants had the nature and procedure of the experiment explained to them. They were then seated in a small, enclosed room. The room lights were dimmed but not completely turned off (in order to enable the participants to still be able to see the keyboard).

The participants were asked to recognize the expression using a 10-alternative non-forced-choice task. More specifically, they were to select the name of the expression from a list that was displayed on the side of the screen. The list of choices, which included all nine expressions as well as "none of the above," was displayed at all times. More information on this type of task and its relationship to forced-choice tasks can be found in Cunningham et al. (2005) and Frank and Stennett (2001).

At the beginning of each trial the participant was prompted to press the spacebar. After the participant pressed the spacebar, a single expression was shown in the center of the screen. As soon as the participant felt that they knew the answer, they were to enter the corresponding number. In the dynamic condition, the individual images in the recording sequence were shown one after another in order at 25 Hz—in other words, as a movie. The entire movie was shown once. In the static condition, the peak frame was shown for the same amount of time as the dynamic sequence would have been shown. If participants had not responded before the end of the allotted display time, a black screen was shown and the participant was prompted to enter their decision.

Crossing six actors, nine expressions, and two conditions yielded 108 trials. Each trial was shown 5 times. Each trial was shown, in a random order, once before any trial was shown (in a different random order) again. Participants were given the chance to take a small break halfway through the 540 trials.



Movie 1. The original sequence of one actress's disgust expression.

Results and discussion

Overall, the advantage of dynamic over static expressions (78% versus 52% accuracy, respectively) also holds for real, normal intensity, video-recorded expressions and peak static expressions. Nearly every expression in the current experiment showed the dynamic advantage (see Figure 1). A two-way ANOVA was performed on the results with condition (static and dynamic) and expression as within-participants factors. All effects were significant. The significant main effect of condition ($F(1,9) = 270.21$, $p < 0.001$) confirms that, on average, the dynamic sequences were recognized significantly better than the static sequences. The significant main effect of expression ($F(8,72) = 18.78$, $p < 0.001$) shows that some expressions are more recognizable, on average, than others. This is consistent with previous experiments, with this database as well as others (Cunningham, Breidt, Kleiner, Wallraven, & Bülthoff, 2003a, 2003b; Cunningham et al., 2005, 2004).

The significant expression by condition interaction ($F(8,72) = 63.17$, $p < 0.001$) indicates that some expressions have a greater dynamic advantage than others. First, it is clear that agree and disagree depend strongly on dynamic information. In fact, they do not seem to be recognizable at all in the static condition. The second group (sad, disgust, clueless, confused, and surprised) shows a mild dynamic advantage. Finally, two expressions (happy and thinking) do not show a dynamic advantage. Interestingly, the happy expressions show a small static advantage. This is consistent with the findings of Ambadar et al. (2005), who also found that of their 6 expressions, only happiness did not show a dynamic advantage.

In sum, there is a reliable advantage of dynamic over static expressions. Note that if participants were merely using the peak image in the dynamic condition (i.e., the last frame), then performance should have been similar in the two presentation conditions, or even better in the static

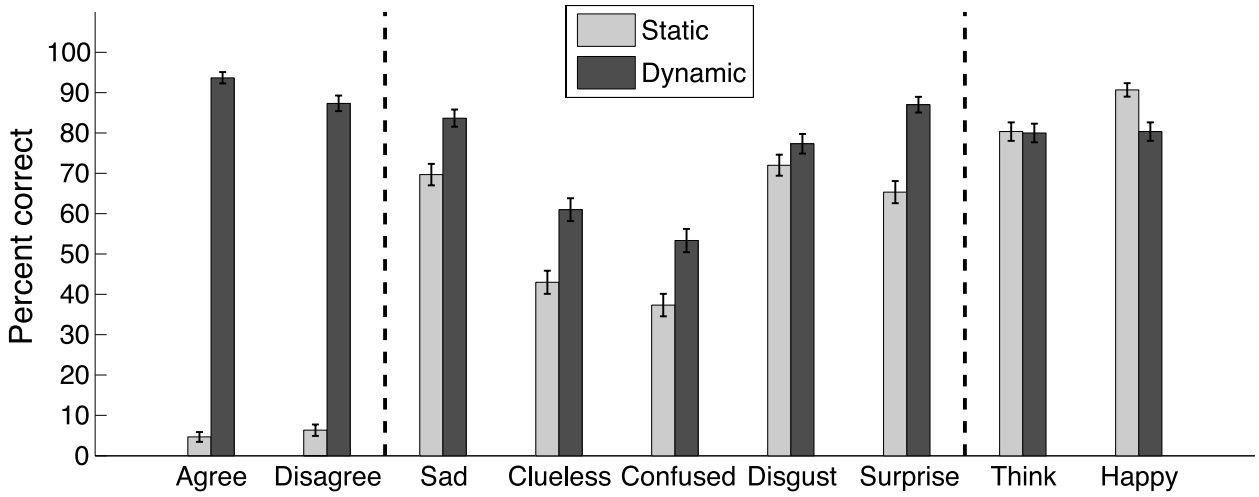


Figure 1. The accuracy scores for the two presentation conditions in Experiment 1 as a function of expression. The two vertical, dotted lines divide the expressions into three different effect size groupings. The error bars represent the standard error of the mean.

condition (since the participants would have had longer to view the critical frame). The pattern of results suggests, instead, that there is some information available in the dynamic sequences that is not available in the peak frame. The exact nature of this effect will be explored in more detail in the following experiments.

Experiment 2: Static array

Experiment 1 confirmed that an expression is more accurately recognized when it is depicted by a dynamic sequence rather than as a static photograph of the expression's peak. There are, of course, many potential reasons for this. The simplest reasons derive from the observation that the dynamic condition consists of a number of frames and the static condition has only one. The first possibility is that the frame chosen in Experiment 1 for the static condition (i.e., the last peak frame before the offset of the expression) was not optimal. That is, maybe some frame other than the last one held a clearer depiction of the expressions. Under this explanation, the participants could still be treating the dynamic condition as a series of static photographs.

A slightly more sophisticated version of this hypothesis can be constructed if one considers that the perception of faces is, at least partially, based on its component parts (for an overview of the debate between the holistic and component theories of facial processing, see Maurer, Grand, & Mondloch, 2002; Schwaninger et al., 2006; see also Mondloch & Maurer, 2008; Tanaka & Farah, 1993). Moreover, it has recently been shown that people can easily integrate separate components of a face over time, at least for the recognition of identity (Anaki, Boyd, & Moscovitch, 2007). Anaki et al. (2007) presented a series of faces either intact or divided into three regions (top including eyes, middle including nose, and bottom including the mouth). The three regions were presented for 17 ms with a blank interval between them (ranging from 17 ms to 700 ms). The images were presented either normally or upside down. Recognition rates as well as matching rates show a similar-sized inversion effect (which is taken as indirect evidence for the usage of the information contained in the relationship between the parts) for the intact and segmented faces—as long as all parts of the face were presented within 500 ms of each other. This is consistent with people being able to integrate the parts of a face over a large period of time in order to see a full face. Anaki et al. (2007) suggested that this process occurs in iconic memory, although visual short-term memory could not be completely ruled out.

Thus, it is possible that people are treating the dynamic condition as a series of still photographs and are merely picking and choosing the best parts from different frames. These temporally distant parts are integrated into a static whole, and it is this static whole that is used to recognize the expression. While this would represent an integration of information from different frames, it does not require motion information. Since this hypothesis would require the identification of and subsequent attentional and intentional selection of object parts, along with the suppression of other similar parts in identical retinal areas, it would seem to require a more sophisticated integration process than is generally attributed to iconic memory. It might, however, be within the capabilities of visual short-term memory, which has a limited capacity (3–4 objects) and location independent representation (Irwin, 1991).

If either hypothesis is correct, then presenting all images simultaneously, side by side, should produce recognition rates that are at least as high as in the dynamic condition. Experiment 2 explicitly tests these two hypotheses using five conditions. The first two, dynamic and

static, are the same as in the previous experiment. The repetition of these conditions tests the robustness and reliability of the effects found in Experiment 1. The remaining three conditions employ only the last 16 frames of each recording. The third condition, dynamic 16, shows the last 16 frames as a movie. Since an expression does not spring immediately into existence, but develops over time, it is possible that the early phase of an expression contains the most important information. Edwards (1998) provided some evidence that is consistent with this. Edwards printed out each frame from dynamic expressions and scrambled the order of the pictures. Participants were asked to place the pictures back into the correct order. Edwards found that humans are differentially more sensitive to temporal order of information that is present in the early phases of an expression than in the later phases. If the critical information for the recognition of an expression is in the early phases, then performance in the dynamic 16 condition should be lower than in the dynamic condition. Thus, a comparison of the two dynamic conditions will test the appropriateness of using just the last 16 frames.

In the fourth condition, referred to as static 16 ordered, the 16 frames are displayed simultaneously in a 4 × 4 grid. Note that the images here are exactly the same as the images shown in the other conditions; they are merely shown all at once. To ease the participants' search through the images, the temporal order of the frames is spatially represented: The first frame is shown at the top left, and subsequent frames are ordered similar to words in a book—left to right and top to bottom. Thus, the last frame is at the bottom right. The final condition, referred to as static 16 scrambled, also shows the frames in a grid but in a scrambled order. This might impair the search through the individual frames. Comparison between the dynamic 16 and the two static array conditions will indicate whether the dynamic advantage was due to the mere presence of more pictures in the dynamic condition.

Methods

Nine new individuals participated in the experiment. The video sequences were identical to those used in Experiment 1. The dynamic and static conditions were identical to those conditions in Experiment 1. The remaining three conditions used only the last 16 frames of each video sequence. These 16 frames were either shown as a movie (dynamic 16), in an ordered 4 × 4 array (static 16 ordered), or in a scrambled 4 × 4 array (static 16 scrambled). A new random order was used for each occurrence of the scrambled array. Only one repetition of each expression was used. Otherwise, the stimuli and procedure were identical to those in Experiment 1. Crossing six actors, nine expressions, and five conditions yielded 270 trials, which were shown in random order.

Results and discussion

Overall, as can be seen in Figure 2, the two dynamic conditions produced higher recognition rates than the three static conditions. The results for the dynamic and static conditions (76% and 56%, respectively) are remarkably similar to those from Experiment 1 (78% and 52%, respectively). This not only confirms the dynamic advantage found in Experiment 1 but demonstrates the reliability of the effect.

A two-way ANOVA was performed with condition and expression as within-participants factors. All main effects and interactions were significant (all F's > 19, all p's < 0.001). The main effect of condition shows that at least one of the presentation styles is different from the others, which seems to be largely due to a general static versus dynamic effect. More specifically, performance is significantly worse in all three static conditions than in either dynamic condition (all t's(485) > 4.68, all p's < 0.001).[3] Second, the two dynamic conditions do not, on average, differ significantly from one another (t < 1). Note that this suggests that the last 16 frames can be safely used in the static array conditions without loss of generality. While the two static array conditions do not statistically differ from one another (t(485) < 1), they are both slightly, but significantly, higher than the static condition (all t's(485) > 2.05, all p's < 0.05). This suggests that there is a little more information about the expressions, on average, in the static arrays, but nowhere near as much as when the same frames are shown as a movie.

The results also show the same three groups of expressions found in Experiment 1 (see Figure 2). Agree and disagree are strongly impacted by the loss of dynamic information. Happy and thinking show a static advantage. It is interesting to note that these two expressions also tend to show a superiority of the dynamic 16. For happy, there is still an advantage of the static condition over the dynamic 16 condition. For thinking, performance is similar in the static and dynamic 16 conditions. The remaining expressions show a moderate effect.

Several expressions (disagreement, thinking, and happiness) show a slight increase in accuracy when only the final frames are shown (dynamic 16 is higher than dynamic). Others (clueless, sadness, disgust, and surprise) show a decrease. With the exception of clueless, these differences are small. This shows that there is expression-relevant information in the initial frames, as expected from Edward's results. This information is helpful in some cases (i.e., adding the initial frames increases recognition rates) and hurtful in others. For thinking, for example, the initial frames show some similarity to those from confusion, as discussed in Cunningham et al. (2005). For all expressions, the two array conditions were at least as good as the static condition. In some cases, they were better. While disgust shows a trend in the other direction, this difference is not statistically significant. Regardless,

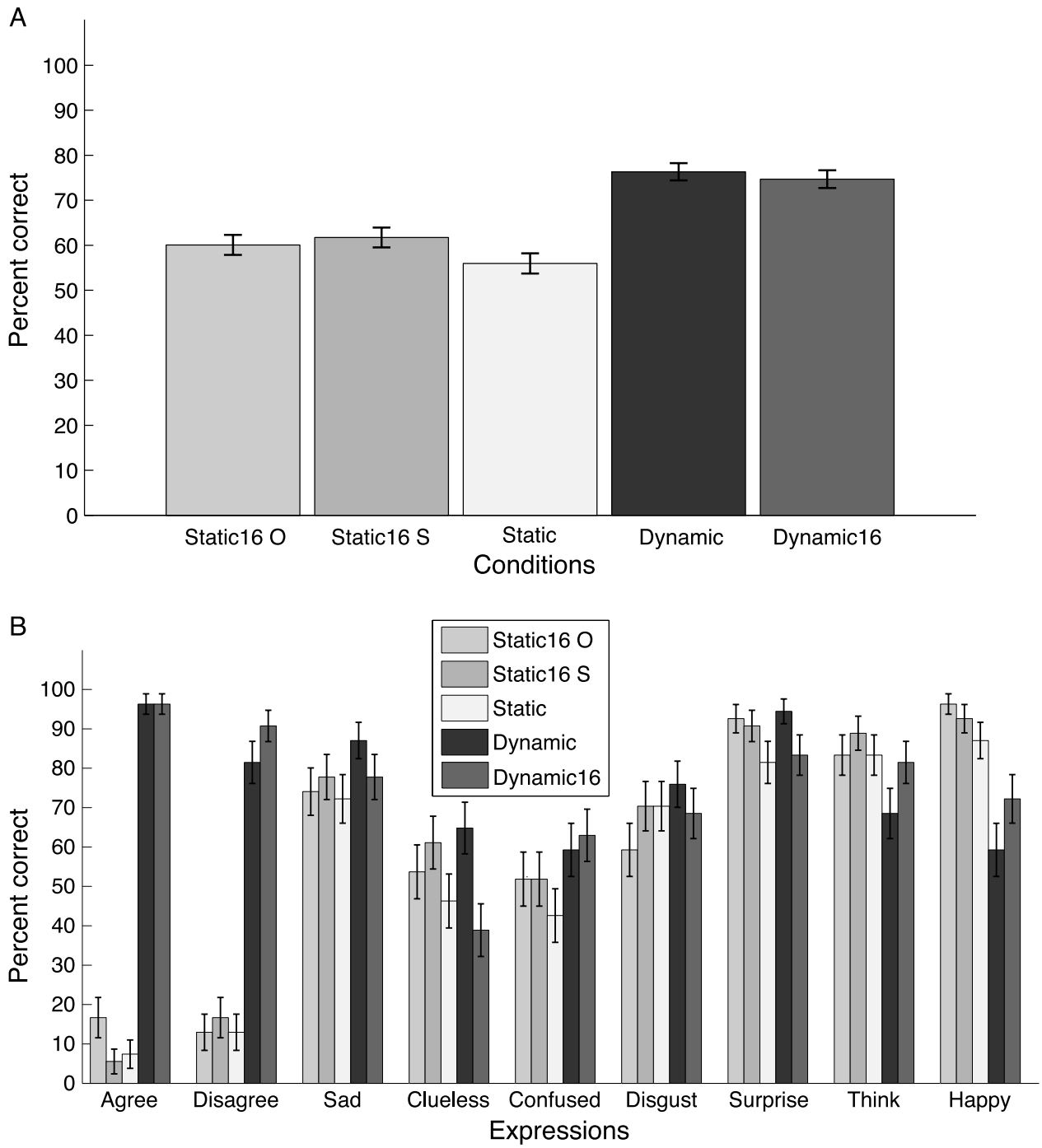


Figure 2. Recognition accuracy for Experiment 2. (Top) Overall accuracy. (Bottom) Accuracy as a function of expression for the different presentation conditions.

the higher of the two dynamic conditions is almost always better than the highest of the static conditions.

In sum, for most expressions people are not scanning the photographs in the dynamic conditions as if they were a series of still photographs. People do not seem to be "picking and choosing" the best parts or even simply looking for a better still shot of the face. It is, of course, possible that the grid nature of the array conditions somehow prevented the participants from extracting information, thus mitigating any advantage for having multiple images. This seems, however, unlikely given that performance was as good as, if not better than, the static condition—which would imply that they can extract exactly as much information from the grid of images (many of which show a near neutral expression) as when there is only one image. Moreover, the present results are similar to those of Ambadar et al. (2005), who presented the first 3 to 6 frames one after another with a mask

between them. Thus, it seems that there is some form of dynamic information that is not available in the individual frames. For most expressions, this information is also present in the final 16 frames. Interestingly, the large difference between the two static array conditions and the dynamic conditions suggests that, at least in terms of recognition accuracy, the participants were not able to mentally convert the spatial order of the frames into the proper temporal sequence and then infer the critical dynamic information.

Experiment 3: Scrambled

Experiment 2 showed that the dynamic advantage is not due to the presence of multiple images, but that some form of dynamic information is being used. It is possible that the mere *presence* of motion is what is important, and not its content. In other words, the change that occurs between two photos gives rise to a motion signal that not only provides some information about *how* the part changed but also *where* the change occurred. Perhaps it is this help in finding the changed area that is important.

In Experiment 3, we tested this hypothesis by presenting the frames as a movie but in a random order. Performance in this scrambled condition is compared to performance in the dynamic condition. In both conditions, there are motion signals, the same frames are seen, the amount of time that any given frame is presented is the same, and any potential static masking effects are the same. If neither the presence of multiple images (albeit ones that are presented very rapidly) nor the mere presence of motion signals at the correct locations are the cause of the dynamic advantage, then the dynamic condition should yield higher recognition rates than the scrambled condition.

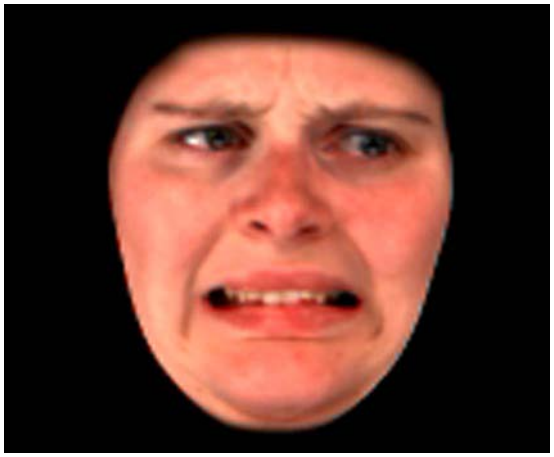
Methods

Ten new individuals participated in the experiment. The dynamic condition was the same as in the previous experiments. In the "scrambled" condition, all frames of the video sequence were shown sequentially, just as in the dynamic condition, but the order of the frames was randomized (see Movie 2). The experimental procedure was the same as in Experiment 1, with the exception that only three repetitions were used. Crossing six actors, nine expressions, two conditions, and three repetitions yielded 324 trials, which were shown in random order.

Results and discussion

Overall, the dynamic sequence produced higher recognition rates than the scrambled sequences (76% versus 56%, respectively). A two-way ANOVA was performed with condition and expression as within-participants factors. All effects were significant (all F's > 8.7, all p's < 0.001). While all expressions showed an intact advantage (i.e., recognition rates were higher in the dynamic condition than in the scrambled condition), the significant condition by expression interaction suggests that the effect is not the same for all expressions. As can be seen in Figure 3, the grouping of expressions has changed some. First, agree and disagree still show some of the largest effects. Unlike in Experiments 1 and 2, however, surprise also shows a large effect. Second, happy and thinking now show effects that are about the same size as the other expressions. These results suggest that while dynamic information may not be central for happy, thinking, or surprise, recognition of these expressions is nevertheless sensitive to incorrect dynamic information. In contrast, sadness shows a much smaller effect of scrambling than was caused by removing dynamic information. This means that sadness is impaired by the loss of dynamic information (as was shown in Experiments 1 and 2) but is not strongly affected by the scrambling of it.

One possible explanation for this pattern lies in the speed of normal motion. Sadness, which does not show an intact advantage, has a rather slow motion, while surprise and happy tend to have faster motions. Thus the artificial speeding up of the motion signals caused by the scrambling might push the motion from surprise and happy out of normal scales, but sadness might remain within a normal range. This explanation is undermined by the fact that researchers have shown that speeding up sadness (for example, by playing the movie faster) reduces recognition rates (it tends to look more like anger; see, e.g., Hill, Troje, & Johnston, 2005). Nonetheless, future research should examine the interaction between scrambled motion signals, the velocity of those signals, and normal motion perception thresholds.



Movie 2. The scrambled sequence of one actress's disgust expression.

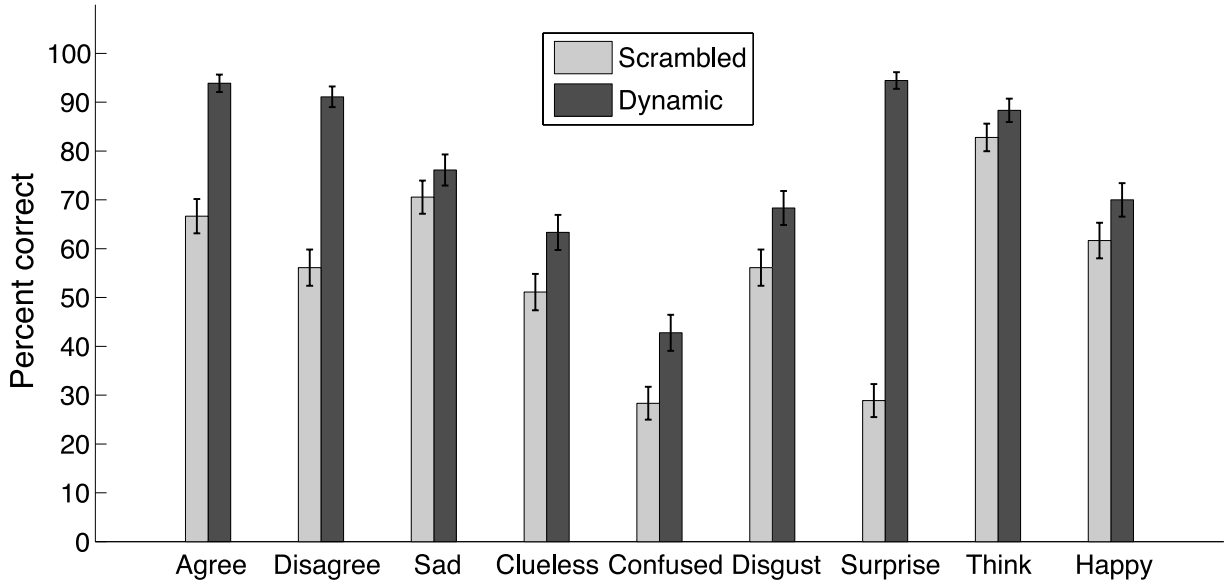


Figure 3. The accuracy ratings for Experiment 3 as a function of expression for the different presentation conditions.

Interestingly, the recognition rate in the scrambled sequences is very similar to the various static conditions used in the first two experiments. It seems, then, that dynamic advantage is neither due to the presence of multiple images nor to the mere presence of face-related motion signals.

Experiment 4: Temporal reversal

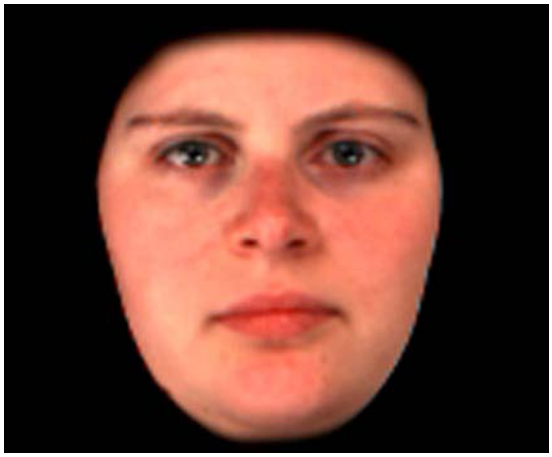
Experiments 1–3 have shown that there is some information about the specific expressions shown in a video that is available only over time. Experiment 3 showed that this information is not only specific to the individual expressions but is also sensitive to temporal ordering (i.e., scrambling). These observations combined with the fact that time, unlike space, has a characteristic direction raises the question of whether facial expressions are temporally reversible. That is, some events, such as the evaporation of a liquid or the disintegration of an object, are clearly one directional. When a film of such an event is viewed backward (i.e., starting at the last frame and working progressively toward the first frame), the event tends to look extremely odd (Gibson, 1979). Other events are equally plausible in both temporal directions (e.g., a sphere rolling across a flat, horizontal table). Work on both recognition of rigid (Stone, 1998) and non-rigid, novel objects (Chuang, Vuong, Thornton, & Bülthoff, 2006) has shown that temporal order plays an important role for correct object identification—time reversed sequences were recognized with less accuracy in both studies, which demonstrates that the visual system is acutely tuned toward and explicitly uses natural temporal statistics. The following experiment is designed to determine if the dynamic information present in facial expressions is sensitive to temporal direction.

Note that this can also be considered a rather subtle test of the hypothesis that the dynamic advantage is due to information available only over time. A video sequence played backward has the same number frames, the same overall temporal duration, and the same (relative) ordering of frames as a sequence played forward. Indeed, the motion signals in two conditions are identical (albeit backward in one condition). Thus, any difference in recognition performance between forward and backward sequences would strongly suggest that the dynamic advantage is due to some characteristic dynamic information present in the expressions.

Since the sequences used in the previous experiments started at a neutral expression and ended at the peak expression, one might find a difference between the forward and backward conditions based trivially upon the placement of the peak expression (although such a primacy or recency effect is unlikely, as was discussed in Experiment 1). Nonetheless, to help avoid this potential confound, the present experiment also included the "full" sequences (which went from neutral through peak and back to neutral). Thus, the first and last frames were identical in both the forward and the backward conditions, while the peak was a considerable number of frames away from either end.

Methods

Ten new participants took part in the experiment. The procedure and video sequences were the same as in the previous experiments. In addition to the dynamic condition, the present experiment contained a backward condition in which the frames in the video sequence were



Movie 3. The backward clipped sequence of one actress's disgust expression.

in the reverse temporal order (i.e., starting at the last frame and ending with the first frame; see Movie 3). Each video sequence was shown in its clipped form (from neutral to peak) and its full form (from neutral to peak and back to neutral).

Results and discussion

A three-way ANOVA was performed with condition, length (full versus clipped), and expression as within-participants factors. The significant main effect of condition ($F(1,9) = 23.19$, $p < 0.001$) means that the forward sequences were recognized significantly more often than the backward sequences (79% and 72%, respectively; see Figure 4). This suggests that the dynamic advantage is, at least in part, due to some form of characteristic dynamic information. The main effect of expression ($F(8,72) = 21.8$, $p < 0.001$) is similar to the previous experiments. The main effect for length was not significant, nor were any of the interactions with length (all F's < 1.93, all p's > 0.068). This suggests that the clipped and the full sequences yielded similar accuracies, confirming earlier results (Cunningham et al., 2005). Critically, the lack of an interaction between condition and expression ($F(8,72) = 1.54$, $p > 0.15$) means that the effect of temporal direction is similar for all expressions.

Experiment 5: The window of temporal integration

Experiments 2–4 showed that the individual images are somehow integrated over time to extract some form of information that is not present in any static image. Many perceptual processes that rely on dynamic information can only take advantage of information presented within a specific period of time: Increasing the length of a sequence increases the effectiveness of dynamic information up to the limits of the integration window. This occurs for low-level sensory processes (such as energy summation in retinal pigments as shown by Bloch's

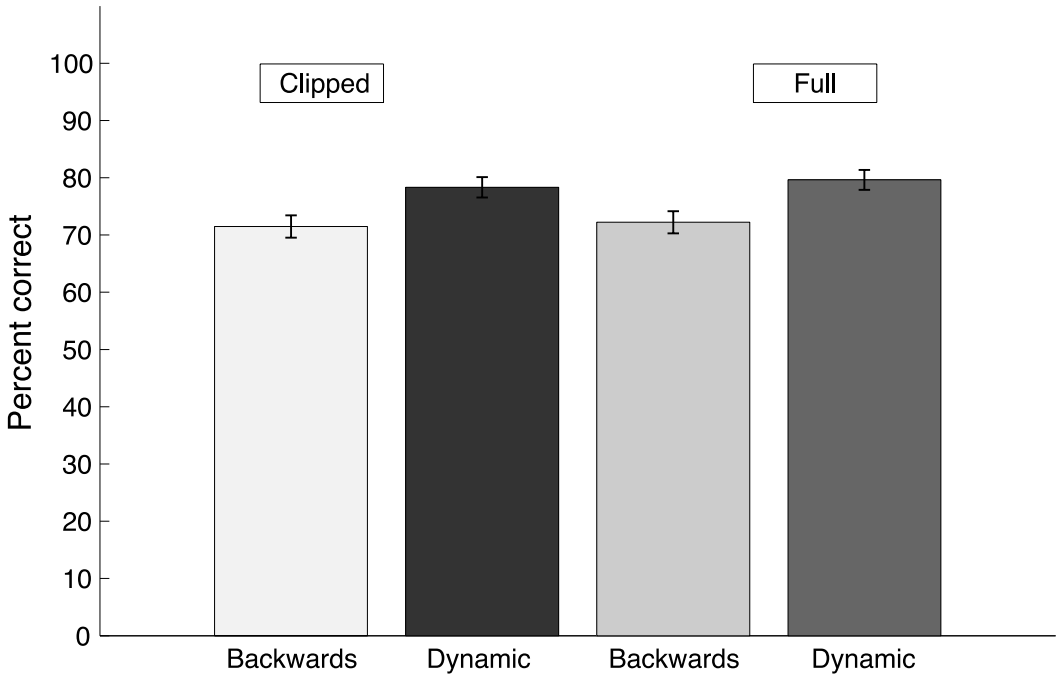
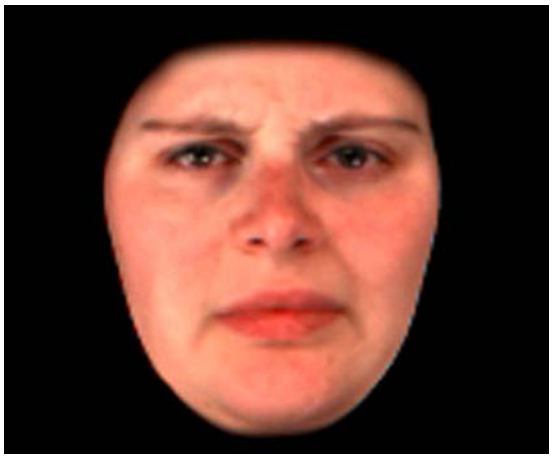
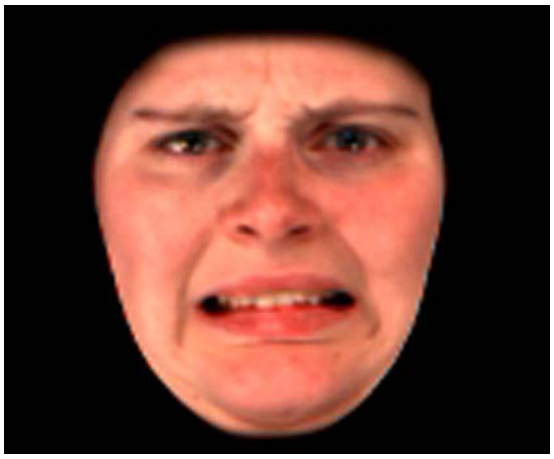


Figure 4. Overall recognition accuracy for Experiment 4.



Movie 4. The scrambled 2 sequence of one actress's disgust expression.



Movie 5. The scrambled 4 sequence of one actress's disgust expression.

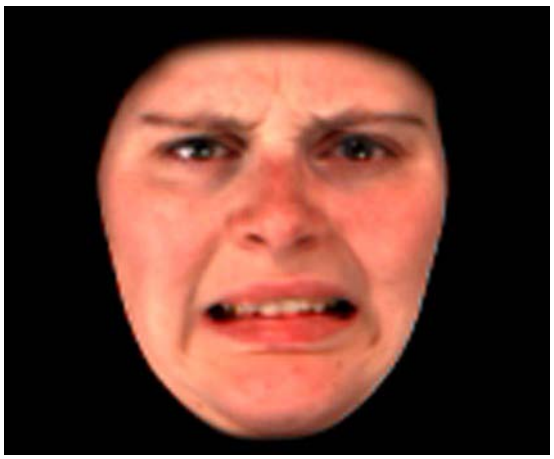
(1885) law), as well as mid-level perceptual processes (such as Spatiotemporal Boundary Formation; Shipley & Cunningham, 2001). For Spatiotemporal Boundary Formation, for example, the dynamic pattern of texture appearances and disappearances on a distant surface caused by an object moving in front of it can be integrated to define the shape, location, velocity, translucency, and relative depth of the moving object. Any increase in the length of the sequence up to 150 ms produces an increase in the perceptual strength and clarity of each of these dimensions. Increasing the length of the sequence beyond 150 ms, however, produces no further perceptual increases (Shipley & Kellman, 1994). Experiment 5 uses the same technique presented by Shipley and Kellman (1994) to determine the size of the temporal integration window for facial expressions: How many images must be present in the proper order for temporal integration to occur?

Methods

Ten new participants took part in the experiment. The procedure and video sequences were the same as in the previous experiments. The dynamic condition was the same as in the previous experiments. The four scrambled conditions were based on the scrambled condition in Experiment 3. For each of these conditions, a "preserved unit" was defined. For the scrambled 1 condition, the preserved unit was a single frame. This is identical to the scrambled condition in Experiment 3. In scrambled 2, the preserved unit was subsequent image pairs (i.e., frames 1 and 2, 3 and 4, 5 and 6, etc.). The order of the image pairs was randomized. In scrambled 4, the preserved unit was sequences of four subsequent frames (e.g., 1, 2, 3, and 4 were one unit; 5, 6, 7, and 8 another). Finally, in scrambled 6, sequences of 6 subsequent frames were kept together. For examples, see Movies 4, 5, and 6. Note that since not all image sequences are evenly divided by 2, 3, 4, and 6, some image sequences had one unit that was less than the appropriate preservation unit. For example, a 17-frame sequence would have 4 units of 4 in the scrambled 4 condition, and one unit of 1. Crossing six actors, nine expressions, and five conditions yielded 270 trials, which were shown in random order.

Results and discussion

Overall, recognition performance increased as the number of frames that were kept intact increased (see Figure 5). A two-way ANOVA was performed with condition and expression as within-participants factors. All effects were significant (all F's > 2.48, all p's < 0.001). The results for the dynamic and the scrambled 1



Movie 6. The scrambled 6 sequence of one actress's disgust expression.

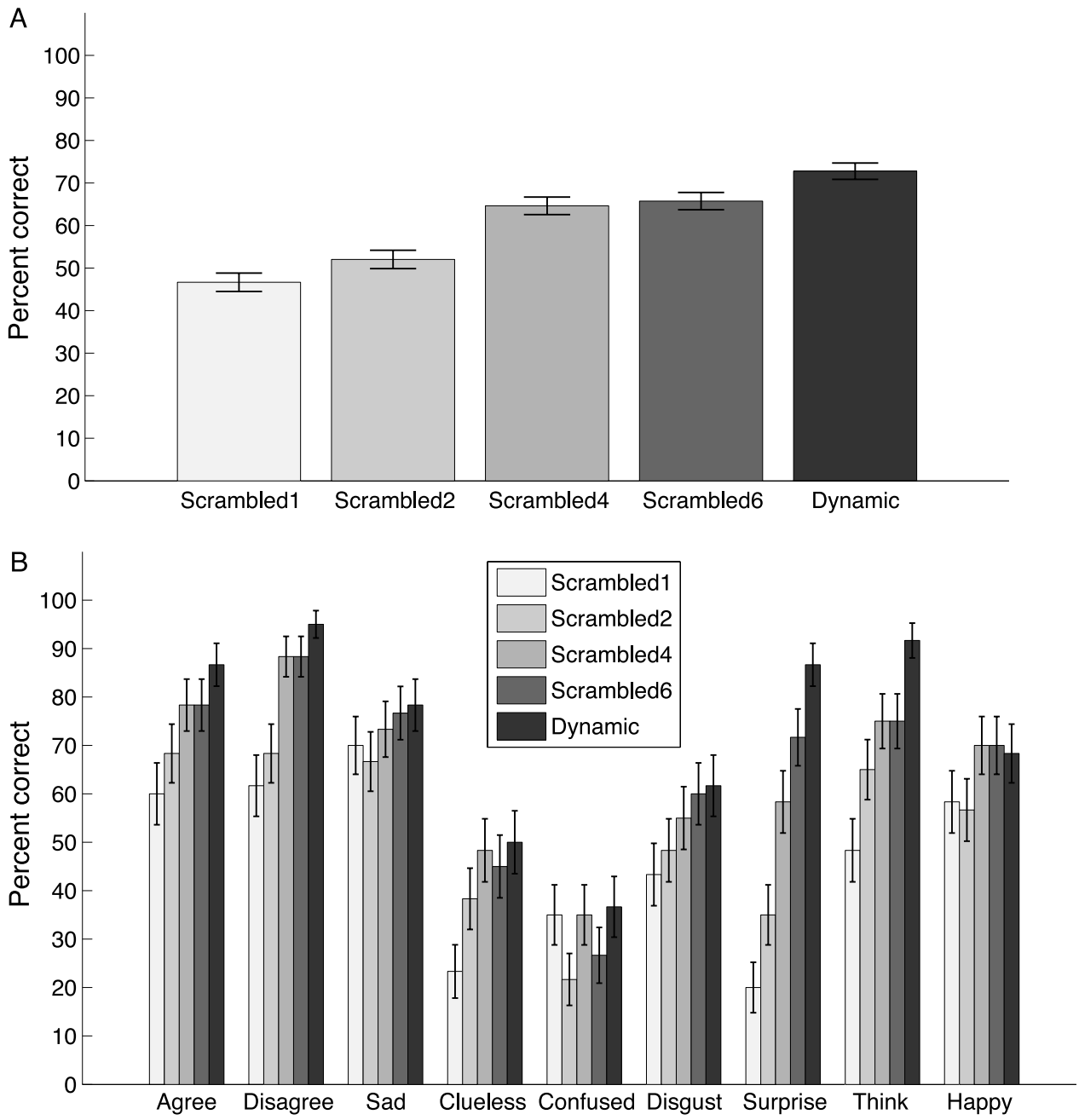


Figure 5. Recognition accuracy for Experiment 5. (Left) Overall accuracy. (Right) Accuracy as a function of expression for the different presentation conditions. Scr1, Scr2, Scr4, and Scr6 refer to the scrambled 1, 2, 4, and 6 conditions, respectively.

conditions were nearly the same as found in Experiment 3 (Overall, 73% and 47% here, as compared to 76% and 56% in Experiment 3).

A glance at Figure 5 shows little change in performance from scrambled 4 to scrambled 6, but a larger change between scrambled 6 and the dynamic condition (this latter change is significant: $t(485) = 3.11$, $p < 0.01$). There are many potential explanations for this. While additional work is necessary to precisely determine the upper bound to the integration window, it is clear that the window is at least 4 frames (100 ms) long.

Conclusion

Experiment 1 showed that the dynamic advantage can be found for video recordings and full peak static images using normal intensity, conversational expressions from real individuals. Experiments 2 and 3 showed that the results are not due to the mere presence of multiple images, or to a poorly chosen peak frame. Experiment 3 also showed that the results are not due to the mere presence of dynamic information that among other things

indicated the location of important change. Experiment 4 showed that for each of the nine expressions, there was some characteristic dynamic information for the expressions, and that recognition of the expressions is sensitive to temporal direction (i.e., being played backward). Finally, Experiment 5 showed that the more frames are kept together, the easier it is to recognize these expressions. In sum, dynamic expressions are recognized more easily and more accurately than static expressions, this effect is fairly robust, and the effect cannot be explained by simple, static-based explanations. In other words, there is some form of information that is available only over time that is being used in the dynamic sequences. The window of temporal integration for these facial expressions is at least 100 ms.

Having demonstrated that there is dynamic information that is characteristic for the specific expressions, and that the perception of expressions is sensitive to this information, it would be interesting to see what the specific features of that information are. One method to do this would be to alter the dynamic information, such as by altering the speed, acceleration, or the distance traveled by the different facial regions. Indeed, there is some evidence already that increasing the distance traveled by an area while holding the timing constant (which also alters the speed and acceleration of the part) can exaggerate the emotional content of sentence (Hill et al., 2005). Such research should, however, be careful in its choice of expressions, since some expressions are more reliant on dynamic information than others. Agreement and disagreement, for example, rely very heavily on dynamic information—they are not recognizable in static displays. Other expressions, such as happiness, do not require dynamic information. While dynamic information is not necessary for all expressions, all expressions were sensitive to its alteration, again to different degrees. Our research has shown that using a wider spectrum of expressions will help ensure the generality of the results.

The dynamic information present in the different expressions is located in different areas (Nusseck et al., 2008). Some expressions rely on changes in head orientation, others on the rotation of the eyes, still others on complex, non-linear, non-rigid deformation of facial parts. The fact that the dynamic advantage holds for so many different expression types, change locations, and change types strongly suggests that there will also be a dynamic advantage for gestures and body motion in general. Indeed, it has long been known that there is critical dynamic information in body motion. Research using point-light displays has shown that motion information is sufficient to recognize the action carried out by a moving person, as well as their identity, sex, and emotional state (Cutting & Kozlowski, 1977; Dittrich, 1993; Dittrich, Troscianko, Lea, & Morgan, 1996; Kozlowski & Cutting, 1977; Pollick, Paterson, Bruderlin, & Sanford, 2001). This phenomena, referred to as biological motion, shares many characteristics with the dynamic information in faces, including the fact that point-light displays are sufficient to recognize expressions (Bassili, 1978, 1979) and that facial motion can help determine the identity of someone (Knappmeyer, Thornton, & Bülthoff, 2003; Lander, 2005). Additionally, the present work has shown that dynamic perception of facial expressions is robust, reliable, and sensitive to temporal reversal. These are also features of biological motion in general.

One major difference between body motion and facial motion is that body motion is constrained to be rigid, while facial motion is generally non-rigid and non-linear. Early models of biological motion had a rigidity assumption, which greatly eases the task of recovering body shape and motion from the moving points (e.g., Johansson, 1973; Marr & Vaina, 1982; O'Rourke & Badler, 1980; Rohr, 1994). While such models could explain the perception of some expressions (such as agreement and disagreement), they would not work for expressions that contained non-rigid facial motion (such as surprise or disgust). Most models of biological motion use some form of integration of static poses. Some models, such as the ones by Beintema and Lappe (2002), Lange and Lappe (2006), and Webb and Aggarwal (1982) explicitly avoid any motion information. In such models, the effect of time is represented by differentially weighing neighboring poses. Assuming one could define "facial poses," and assuming that a linear weighting of them will produce meaningful results, such a model might be able to explain many of the present results. It could even explain the intact advantage found in Experiment 3. It is unclear if it can explain the reversal effect found in Experiment 4.

Some models of biological motion also explicitly include motion information, particularly optic flow (Casile & Giese, 2005; Giese & Poggio, 2003; Mather & Murdoch, 1994; Mather, Radford, & West, 1992; Troje, 2002). Such models would seem to provide better fits for the present data. Since the explicit reliance on optic flow allows one to easily extract complex, non-rigid motion patterns, these models might work well for the perception of facial expressions in general. Future work could test how specific predictions of this class of model, which were largely developed for body motion, are also applicable for dynamic facial expressions keeping in mind that facial expressions rarely are generated by oscillatory motions, for example. Additionally, such work should also compare not only the explicit prediction of these models but also computer vision models designed explicitly for facial expressions (see, e.g., Cohn & Kanade, 2007; Zeng, Pantic, Roisman, & Huang, 2009).

Regardless of how facial information is being processed, it is clear that there is some form of information that is only available over time. Any attempt to understand how humans use their eyes, face, and head to communicate that only uses static photographs—whether it is a single photograph or a series of photographs seen one after another—will never be able to explain the perception

of expressions. Likewise, any system designed to describe facial expressions that does not explicitly allow for the description of dynamic information will prove ultimately to be inadequate.

Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) Project "Perceptual Graphics." We would like to thank Mario Kleiner for his help with the MPI Videolab.

Commercial relationships: none.
Corresponding author: Douglas W. Cunningham.
Email: douglas.cunningham@tuebingen.mpg.de.
Address: Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tübingen, Germany.

Footnotes

[1]By facial information, we refer not only to deformations of the facial surface and eye gaze but also to head orientation.

[2]We will refer to information that is present only over time as dynamic information or spatiotemporal information.

[3]All post-hoc *t*-tests are two-tailed, dependent measure tests.

References

Adolphs, R., Tranel, D., & Damasio, A. R. (2003). Dissociable neural systems for recognizing emotions. *Brain and Cognition, 52,* 61–69. [PubMed]

Ambadar, Z., Schooler, J. W., & Cohn, J. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science, 16,* 403–410. [PubMed]

Anaki, D., Boyd, J., & Moscovitch, M. (2007). Temporal integration in face perception: Evidence of configural processing of temporally separated face parts. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1–19. [PubMed]

Averbach, E., & Coriell, A. S. (1961). Short-term memory in vision. *Bell System Technical Journal, 40,* 309–328.

Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a "language of the eyes"? Evidence from normal adults and adults with autism or Asperger syndrome. *Visual Cognition, 4,* 311–331.

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,* 2049–2059.

Bavelas, J. B., Black, A., Lemery, C. R., & 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., & Johnson, T. (2000). Listeners as conarrators. *Journal of Personality and Social Psychology, 79,* 941–952.

Becker, M., & Pashler, H. (2002). Volatile visual representations: Failing to detect changes in recently processed information. *Psychonomic Bulletin and Review, 9,* 744–750.

Beintema, J., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 5661–5663.

Bloch, A. M. (1885). Experiences sur la vision. *C. R. Seances Society of Biology Paris, 37,* 493–495.

Boyle, E., Anderson, A., & Newlands, A. (1994). The effects of visibility on dialogue in a cooperative problem solving task. *Language and Speech, 37,* 1–20.

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

Carrera-Levillain, P., & Fernandez-Dols, J. (1994). Neutral faces in context: Their emotional meaning and their function. *Journal of Nonverbal Behavior, 18,* 281–299.

Casile, A., & Giese, M. (2005). Critical features for the recognition of biological motion. *Journal of Vision, 5*(4):6, 348–360, http://journalofvision.org/5/4/6/, doi:10.1167/5.4.6. [PubMed] [Article]

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

Cassell, J., & 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.

Chuang, L., Vuong, Q. C., Thornton, I. M., & Bülthoff, H. H. (2006). Recognising novel deforming objects. *Visual Cognition, 14,* 85–88.

Cohn, J. F., & Kanade, T. (2007). Automated facial image analysis for measurement of emotion expression. In J. A. C. J. B. Allen (Ed.), *The handbook of emotion*

elicitation and assessment (pp. 222–238). New York: Oxford University Press Series in Affective Science.

Cunningham, D. W., Breidt, M., Kleiner, M., Wallraven, C., & Bülthoff, H. H. (2003a). How believable are real faces?: Towards a perceptual basis for conversational animation. In D. Metaxas, N. Magnenat-Thalmann, & H.-S. Ko (Eds.), *Computer animation and social agents 2003* (pp. 23–29). Washington, DC: IEEE Computer Society.

Cunningham, D. W., Breidt, M., Kleiner, M., Wallraven, C., & Bülthoff, H. H. (2003b). The inaccuracy and insincerity of real faces. In M. H. Hamza (Ed.), *Proceedings of visualization, imaging, and image processing 2003* (pp. 7–12). Calgary: ACTA Press.

Cunningham, D. W., Kleiner, M., Wallraven, C., & Bülthoff, H. H. (2005). Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception, 2,* 251–269.

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

Cunningham, D. W., & Wallraven, C. (2009). The interaction between motion and form in expression recognition. In B. Bodenheimer & C. O'Sullivan (Eds.), *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization (APGV2009)* (pp. 41–44). New York, NY: ACM.

Cutting, J., & Kozlowski, L. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9,* 353–356.

Dittrich, W. (1993). Action categories and the perception of biological motion. *Perception, 22,* 15–22. [PubMed]

Dittrich, W., Troscianko, T., Lea, S., & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception, 25,* 727–738. [PubMed]

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

Ehrlich, S. M., Schiano, D. J., & Scheridan, K. (2000). Communicating facial affect: It's not the realism, it's the motion. In G. Szwillus & T. Turner (Eds.), *Proceedings of the ACM CHI 2000 Conference on Human Factors in Computing Systems* (pp. 252–253). New York, NY: ACM.

Ekman, P. (1972). Universal and cultural differences in facial expressions of emotion. In J. R. Cole (Ed.), *Nebraska symposium on motivation 1971* (pp. 207–283). Lincoln, NE: University of Nebraska Press.

Fernandez-Dols, J., Wallbott, H., & Sanchez, F. (1991). Emotion category accessibility and the decoding of emotion from facial expression and context. *Journal of Nonverbal Behavior, 15,* 107–124.

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

Gibson, J. J. (1979). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum.

Giese, M., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements and action. *Nature Reviews Neuroscience, 4,* 179–192.

Harwood, N., Hall, L., & Shinkfield, A. (1999). Recognition of facial emotional expressions from moving and static displays by individuals with mental retardation. *American Journal on Mental Retardation, 104,* 270–278. [PubMed]

Hill, H., Troje, N., & Johnston, A. (2005). Range- and domain-specific exaggeration of facial speech. *Journal of Vision, 5*(10):4, 793–807, http://journalofvision.org/5/10/4/, doi:10.1167/5.10.4. [PubMed] [Article]

Humphreys, G., Donnelly, N., & 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.

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology, 23,* 420–456. [PubMed]

Isaacs, E., & Tang, J. (1993). What video can and can't do for collaboration: A case study. In J. J. Garcia-Luna & P. Venkat Rangan (Eds.), *MULTIMEDIA '93: Proceedings of the first ACM international conference on Multimedia* (pp. 496–503). New York: ACM.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14,* 201–211.

Kaetsyri, J., Klucharev, V., Frydrych, M., & Sams, M. (2003). Identification of synthetic and natural emotional facial expressions. In J.-L. Schwartz, F. Berthommier, M.-A. Cathiard, & D. Sodoyer (Eds.), *International Conference on Audio-Visual Speech Processing (AVSP 2003)* (pp. 239–243).

Kahneman, D. (1968). Method, findings, and theory in studies of visual masking. *Psychological Bulletin, 70,* 404–425. [PubMed]

Kleiner, M., Wallraven, C., & Bülthoff, H. H. (2004). *The MPI Videolab—A system for high quality synchronous recording of video and audio from multiple viewpoints* (Tech. Rep. No. 123). Tübingen, Germany: Max Planck Institute for Biological Cybernetics.

Knappmeyer, B., Thornton, I., & Bülthoff, H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43,* 1921–1936. [PubMed]

Kozlowski, L., & Cutting, J. (1977). Recognizing the sex of a walker from a dynamic point light display. *Perception and Psychophysics, 21,* 575–580.

LaBar, K., Crupain, M. J., Voyvodic, J. T., & McCarthy, G. (2003). Dynamic perception of facial affect and identity in the human brain. *Cerebral Cortex, 13,* 1023–1033. [PubMed]

Lander, K. (2005). Why are moving faces easier to recognize? *Visual Cognition, 3,* 429–442.

Lange, J., & Lappe, M. (2006). A model of biological motion perception from configural form cues. *Journal of Neuroscience, 26,* 2894–2906. [PubMed] [Article]

Marr, D., & Vaina, L. M. V. (1982). Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B, 214,* 501–524. [PubMed]

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London B, 258,* 273–279.

Mather, G., Radford, K., & West, S. (1992). Low-level visual processing of biological motion. *Proceedings of the Royal Society of London B, 249,* 149–155. [PubMed]

Maurer, D., Grand, R. L., & Mondloch, C. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6,* 255–260. [PubMed]

Mehrabian, A., & Ferris, S. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31,* 248–252. [PubMed]

Mondloch, C., & Maurer, D. (2008). The effect of face orientation on holistic processing. *Perception, 37,* 1175–1186. [PubMed]

Nusseck, M., Cunningham, D. W., De Ruiter, J. P., & Wallraven, C. (under review). Perception of emphasis intensity in audiovisual speech. *Speech Communication, 11.*

Nusseck, M., Cunningham, D. W., Wallraven, C., & Bülthoff, H. H. (2008). The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision, 8*(8):1, 1–23, http://journalofvision.org/8/8/1/, doi:10.1167/8.8.1. [PubMed] [Article]

O'Rourke, J., & Badler, N. (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2,* 522–536.

Pelachaud, C., & Poggi, I. (2002). Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation, 13,* 301–312.

Poggi, I., & Pelachaud, C. (2000). Perfomative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 115–188). Cambridge, MA: MIT Press.

Pollick, F., Paterson, H., Bruderlin, A., & Sanford, A. (2001). Perceiving affect from arm movement. *Cognition, 82,* B51–B61.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368–373.

Rohr, K. (1994). Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics, and Image Processing: Image Understanding, 59,* 94–115.

Schultz, J., & Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research, 194,* 465–475.

Schwaninger, A., Wallraven, C., Cunningham, D. W., & Chiller-Glaus, S. D. (2006). Processing of facial identity and expression: A psychophysical, physiological and computational perspective. *Progress in Brain Research, 156,* 325–348.

Shipley, T. F., & Cunningham, D. W. (2001). Perception of occluding and occluded objects over time: Spatiotemporal segmentation and unit formation. In T. F. Shipley & P. J. Kellman (Eds.), *From fragments to objects: Segmentation and grouping in vision* (pp. 557–585). Oxford, UK: Elsevier Science.

Shipley, T. F., & Kellman, P. J. (1994). Spatiotemporal boundary formation: Boundary, form, and motion perception from transformations of surface elements. *Journal of Experimental Psychology: General, 123,* 3–20. [PubMed]

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review, 5,* 644–649.

Stephenson, G., Ayling, K., & Rutter, D. (1976). The role of visual communication in social exchange. *British Journal of Social and Clinical Psychology, 15,* 113–120.

Stigler, R. (1910). Chronophotische Studien über den Umgebungskontrast [Effects of exposure duration and luminance on the contrast of the surround]. *Pflägers Archiv für die Gesamte Physiologie des Menschen und der Tiere, 134,* 365–435.

Stone, J. (1998). Object recognition using spatio-temporal signatures. *Vision Research, 38,* 947–951.

Tanaka, J., & Farah, M. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology A, 46,* 225–245.

Troje, N. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision, 2*(5):2, 371–387, http://journalofvision.org/2/5/2/, doi:10.1167/2.5.2. [PubMed] [Article]

Vertegaal, R. (1997). Conversational awareness in multiparty vmc. In S. Pemberton (Ed.), *Extended abstracts of CHI'97* (pp. 496–503). Atlanta, GA: ACM.

Wallraven, C., Breidt, M., Cunningham, D. W., & Bülthoff, H. H. (2008). Evaluating the perceptual realism of animated facial expressions. *ACM Transactions on Applied Perception, 4,* 1–20.

Webb, J., & Aggarwal, J. (1982). Structure from motion of rigid and jointed objects. *Artificial Intelligence, 19,* 107–130.

Wehrle, T., Kaiser, S., Schmidt, S., & Schere, K. R. (2000). Studying the dynamics of emotional expressions using synthesized facial muscle movements. *Journal of Personality and Social Psychology, 78,* 105–119. [PubMed]

Weyers, P., Mühlberger, A., Hefele, C., & Pauli, P. (2006). Electromyographic responses to static and dynamic avatar emotional facial expressions. *Psychophysiology, 43,* 450–453. [PubMed]

Williams, M., Breitmeyer, B., Lovegrove, W., & Gutierrez, L. (1991). Metacontrast with masks varying in spatial frequency and wavelength. *Vision Research, 31,* 2017–2023. [PubMed]

Yngve, V. H. (1970). On getting a word in edgewise. In S. Pemberton (Ed.), *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society* (pp. 567–578). Chicago: Chicago Linguistic Society.

Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 59,* 39–58.