Behavior Research Methods, Instruments, & Computers
2000, 32 (4), 505-514

Measuring and modeling facial affect

DIANE J. SCHIANO, SHERYL EHRLICH, KRISNAWAN RAHARDJA, and KYLE SHERIDAN
Interval Research Corporation, Palo Alto, California

In recent years, researchers in computer science and human–computer interaction have become increasingly interested in characterizing perception of facial affect. Ironically, this applied interest comes at a time when the classic findings on perception of human facial affect are being challenged in the psychological research literature, largely on methodological grounds. This paper first describes two experiments that empirically address Russell's methodological criticisms of the classic work on measuring "basic emotions," as well as his alternative approach toward modeling "facial affect space." Finally, a user study on affect in a prototype model of a robot face is reported; these results are compared with the human findings from Experiment 1. This work provides new data on measuring facial affect, while also demonstrating how basic and more applied research can mutually inform one another.

Emotion (or "affect") is central to human experience, and facial displays are our primary means of communicating emotion. Long studied by psychologists, facial affect has become increasingly of interest to computer scientists, especially in the areas of artificial intelligence and human–computer interaction. Indeed, the emerging field of "affective computing" centers on computational modeling of human perception and display of emotion and on the design of affect-based computer interfaces (Lisetti & Schiano, 2000; Picard, 1997). Ironically, this growing applied interest is coming at a time when long-accepted data on human facial affect are being challenged in the psychological research literature.

The classic work on facial expression of emotion was performed in large part by Paul Ekman and colleagues, beginning in the 1960s (reviewed in Ekman, Friesen, & Ellsworth, 1972). An extensive body of evidence was gathered on the recognition of a small number of "basic" emotions: anger, disgust, fear, happiness, sadness, and surprise (contempt was tentatively added only recently). In Ekman's theory, these basic emotions are the elemental building blocks of more complex feeling states. Ekman's data showed that each basic emotion was recognized cross-culturally with high agreement across study participants (Ekman et al., 1972; Russell, 1994). Ekman and Friesen (1978) developed the "facial action coding system" (FACS), a method for quantifying visible facial movements in terms of component muscle actions. The FACS is a highly complex coding system requiring extensive training to use appropriately. Recently automated (Bartlett, Hager, Ekman, & Sejnowski, 1999), it is the single most comprehensive and commonly accepted method for measuring emotion from the visual observation of human faces.

This work was originally presented at the annual meeting of the Society for Computers in Psychology in Los Angeles, November 1999. Correspondence should be addressed to D. J. Schiano, Psychology Department, Jordan Hall, Stanford University, Stanford, CA 94304 (e-mail: diane@psych.stanford.edu).

In recent years, James Russell and colleagues (e.g., Russell, 1994; Russell & Fernandez-Dols, 1997) have contested Ekman and colleagues' classic data on facial affect, largely on methodological grounds. Russell argues that affect in general—and facial affect in specific—is best characterized in terms of a multidimensional "affect space" rather than as discreet emotion categories (such as "fear" or "happiness"). In particular, Russell claims that two dimensions—"pleasure" and "arousal"—are sufficient to characterize facial affect space (Russell, 1980; Russell & Fernandez-Dols, 1997). He calls for new research on perception of facial affect using improved methods and multidimensional analyses (Russell & Fernandez-Dols, 1997). The results of such a research program could have profound implications for both basic and applied research, since they will determine the appropriate bases—and baselines—for measuring facial affect.

This paper presents three experiments on perception of facial affect. Experiment 1 was a replication of the classic work on recognition of basic emotions using improved methods and multidimensional scaling (MDS) analyses, as suggested by Russell. Results are compared against predictions based on both Ekman's and Russell's approaches. In Experiment 2, we directly assessed Russell's model characterizing facial affect in terms of a "pleasure versus arousal" space. Finally, Experiment 3 was a user study on affect in a prototype robot face; the results are compared with the human data from Experiment 1. Taken together, this work provides (1) new data on the measurement of perceived human facial affect under improved experimental conditions, (2) a comparison of alternative approaches toward characterizing facial affect recognition, and (3) a "real-world" demonstration of how basic and more applied studies can be mutually informative.

EXPERIMENT 1

Russell (e.g., Russell, 1994; Russell & Fernandez-Dols, 1997) attacks the "standard method" used in the classic

studies by Ekman and others on several grounds. Much of the data was generated using only a single corpus of fairly unnatural stimuli, primarily black-and-white photographs of a few highly trained actors moving specific sets of facial muscles. Certain experimental design flaws (e.g., failure to properly randomize stimuli, small numbers of trials) are cited, and the frequent use of "within-subjects" designs is challenged. However, Russell's primary criticism of the classic research concerns response format. The standard method relied almost exclusively on the "forced-choice" response format, in which the participant was given a list of labels for the basic emotions and the task was to choose the label that seemed to best match each stimulus image. Russell's critique of the forced-choice format is that (1) it tends to be implemented in an all-or-none fashion that is insensitive to perceived differences in emotional intensity, and (2) if participants were free to pick multiple responses—or, better, to describe the emotions in their own terms—the results might bear little resemblance to the classic findings.

Experiment 1 was designed to respond to Russell's criticisms of the classic studies by providing a direct replication using improved methodology and further analyses. A new, more naturalistic and very high-quality stimulus set was constructed for this research. Several experimental design flaws common to the classic studies were eliminated; for example, a large number of trials and appropriate techniques for stimulus randomization and presentation were used. A rating scale was added to the standard (forced-choice) response format to indicate degree of perceived emotional intensity. Finally, response format was varied in two "between-subjects" comparison conditions ("multiple-choice" and "open-ended"), permitting direct comparison of our results to the classic data as well as additional independent analyses.

The results of Experiment 1 were assessed in several ways. First, we compared performance in this experiment with the classic results, using the standard dependent measure, "correct recognition." Second, the effect of response format was addressed by observing the data derived from using alternative response formats in the comparison conditions. Third, MDS analyses were performed on confusion errors in the recognition data. If Russell's model is correct, a two-dimensional (2-D) MDS solution—with axes of pleasure and arousal—should be sufficient to characterize these data. Finally, the MDS results from Experiment 1 were also compared against MDS analyses of FACS-based predictions of confusion errors (derived from the degree of FACS unit overlap between each emotion pair). While predictions from Russell's model concern subjective states (pleasure and arousal), the FACS-based predictions derive from purely structural similarities between the facial expressions for each basic emotion.

Method

Participants. Eighteen Stanford University students between the ages of 18 and 35 years participated in Experiment 1.

Materials. Four drama students (2 female, 2 male) produced the facial expression stimuli for this experiment. The actors were briefly shown some of Ekman's standard images for each emotion during an initial orientation session. To promote naturalness, the actors were then instructed to simply imagine a time when they strongly felt each emotion (after Ekman & Friesen, 1975). Each actor provided a total of 14 different front-view exemplars of the six basic emotions (anger, disgust, fear, happiness, sadness, surprise). Fourteen "neutral" (no expression) images were also collected from each actor, but these were considered context rather than test images in this experiment. Figure 1 provides an exemplar of each emotion and each actor's face. Each participant viewed all of the stimuli created by two (randomly chosen) actors. The high-resolution, uncompressed digital images were shown full screen on a 14-in. Panasonic color TV monitor (640 × 480 pixels) connected to a PowerMac computer and were viewed at a distance of about 30 in.

Procedure. The participants viewed stimulus images depicting emotional expressions and responded with emotion label(s) for each image. The images were viewed in random order. In each trial, the participant was shown the stimulus image on the TV monitor; the computer monitor was used to present a response screen.

Forced-choice response format condition. The forced-choice (plus ratings) response format was implemented in the following manner: An alphabetized list of labels for the basic emotions was displayed on the computer monitor. The participants chose the one label that best corresponded to the depicted emotion in each image and then rated the degree to which that emotion was present in the image on a scale of 0 (*not at all*) to 6 (*extremely high*). The rating scale appeared as radial buttons adjacent to the selected emotion label.

Comparison conditions. Two comparison conditions ("multiple-choice" and "open-ended") were also run, to explore the effects of response format. An additional 18 participants from the same pool as Experiment 1 were used in each comparison condition. The comparison conditions differed from Experiment 1 (forced-choice condition) and from each other solely in terms of response format. The multiple-choice (plus ratings) response format procedure was identical to that for the forced-choice format condition, with two exceptions. First, the participant could respond with more than one emotion label, if desired. Second, an additional response alternative, "other," was provided at the end of the alphabetized list of emotion labels. When choosing this response, a text window appeared and the participant could freely type in any response. Rating scales were provided for each of the emotion labels (including "other"). The open-ended response format consisted solely of an open text window; the participant simply typed in responses as felt to be appropriate. No rating scale was used. Results from the open-ended condition were assessed by asking a set of 48 participants from the same participant pool to assign each response obtained in this condition to one or more categories, which included the six basic emotion labels and "other." Face validity of the findings was informally assessed against entries in several dictionaries and a thesaurus.

In each condition, 10 initial (nonfeedback) practice trials used randomly selected images from actors not viewed during the test trials. The experimental protocol was implemented in HyperCard on a PowerMac. The participants proceeded at their own pace; the entire procedure lasted under 1 h.

Results and Discussion

Figure 2 presents the correct recognition scores (i.e., the mean proportion of trials in which the participants responded with the expression portrayed by the actor) for each emotional expression in this experiment. Correct recognition scores are the standard dependent measure in the classic studies. The (forced-choice) findings for Experiment 1 are shown in the context of the data ob-

Figure 1. Sample stimulus images for each emotion (anger, disgust, fear, happiness, sadness, surprise) in Experiment 1 (forced-choice response format).

tained in the (multiple-choice and open-ended) comparison conditions. For this graph, the highest rated response was used as the index of correct recognition in each of the comparison conditions (results of further analyses available on request). As the figure shows, correct recognition in Experiment 1 was highest for happiness (M = 99%) and lowest for fear (M = 82%). Performance was uniformly high. These findings closely replicate Ekman and colleagues' classic results in terms of both relative pattern and absolute levels of performance. The similarity of results is especially impressive considering the differences not only in methods but also in the stimuli used in this research. The classic stimuli were created in a painstaking fashion by highly trained actors moving specific muscle groups ("facial action units"); ours were made in a much more naturalistic way.

The rating scale data served as a manipulation check in this experiment, ensuring that the recognized emotion was in fact seen as present in the stimulus image to at least a moderate degree. If Russell's critique of the forced-choice method is correct and the recognition scores are inflated due to constrained response options, extremely low ratings might be expected for at least some expressions. However, the mean rating for the emotional expressions was 3.93 (out of 6) overall; the mean ratings did not fall below the moderate level for any of the ex-

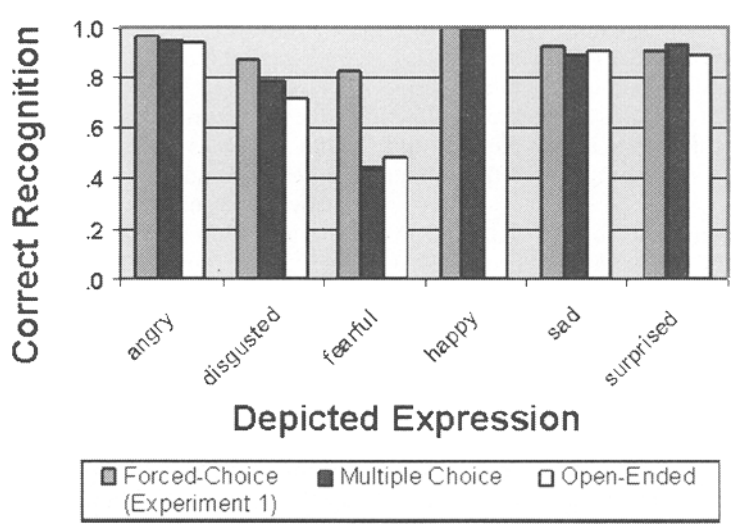

Figure 2. Recognition scores for each emotional expression by response format.

pressions. This suggests that, in general, the participants did see the depicted emotion in the images to at least a moderate degree.

Comparing findings across forced-choice (Experiment 1), multiple-choice, and open-ended response formats is informative. Contrary to Russell's predictions, the results for the three response formats show a strikingly similar pattern. Correct recognition was generally quite high, highest for happiness and lowest for fear. The similarity of the recognition scores across response formats is broken only in the case of fear. When the alternative response formats were used, fear was often "misrecognized" as surprise (M = 25%) or sadness (M = 17%). This confusion pattern is confirmed by various further analyses of the dataset (available on request) and is consistent with similarities in the FACS codes for these emotions. Why fear alone should show such a performance decrement with response format is not clear. Fear appears to be the least compelling emotion under posed conditions, and because it is one of the more ambiguous expressions in terms of FACS unit overlap, when observers are encouraged to give multiple responses, they may tend to do so more for fear. Still, the pattern of incorrect responses to fear was highly systematic, not simply showing greater variability. Taken together, the results do generally support the classic findings.

One further qualification of this conclusion is in order. Additional analyses showed that more than one (M = 2) response was given for 36.7% of the images in the multiple-choice condition. However, the ratings of the additional responses tended to be low, and they followed a pattern of confusion errors predictable from FACS unit overlap. Moreover, when the participants were given the opportunity to provide their own responses in the open-ended condition, the percentage of multiple responses dropped to a mere 1.2%. This suggests that the large number of multiple responses in the multiple-choice condition may reflect a demand characteristic of that format.

MDS analyses were performed on the forced-choice data. The similarity space was derived from a confusion matrix generated from the number of times each basic emotion was mistakenly recognized as any other. The solutions were compared with Russell's 2-D model of affect space, a "circumplex" about the two axes of pleasure and arousal (Russell, 1980). A schematization of this model is shown in Figure 3.

The results of the MDS analyses showed that a 2-D solution accounted for 85% of data variance (stress = 0.12). Figure 4 shows this 2-D pattern, rotated to optimal fit with Russell's model. At first glance, this 2-D space looks similar to Russell's predictions. The datapoints do show a roughly circular arrangement. However, the ordering of emotions on the circle does not quite match; in particular, the relative ordering of anger and sadness is reversed, and the question of where to put disgust is problematic. The relative ordering of the datapoints is the primary result of the 2-D MDS solution and determines the interpretation of the orthogonal dimensions. The 2-D space

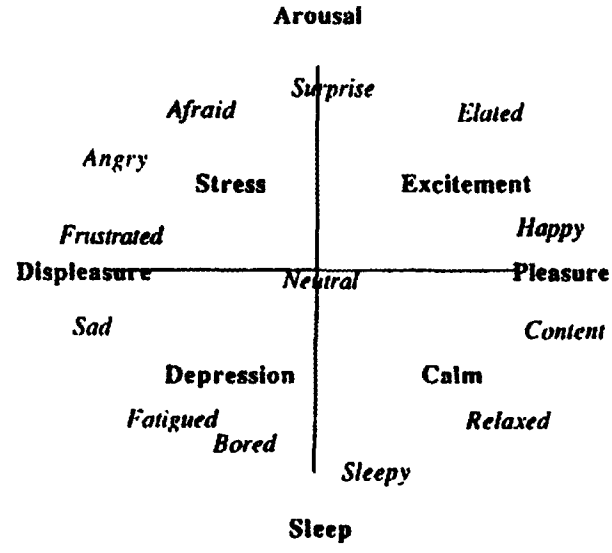

Figure 3. Russell's circumplex model of affect space.

derived from the results of Experiment 1 could perhaps be interpreted as showing an axis corresponding to pleasure, but identifying the second dimension as arousal is less plausible.

Further analyses of this dataset were performed. A 3-D MDS solution accounted for 96% of the variance (stress = 0.04), a substantial improvement over the 2-D approach. These findings clearly do not fit with Russell's model. Figure 5 presents this 3-D solution, together with the results of another, independent (FACS-based) analysis (described below). Note that solutions for each dataset shown in Figure 5 were computed independently; the plots were then rotated for best fit.

For the FACS analysis, a trained FACS coder created a FACS-based confusability index for the basic emotions, derived from the degree of overlap of FACS units (corresponding to muscle movements) between all pairs of emotional expressions (number of overlapping units over total number of units for the two emotions). While this is a fairly crude measure of structural similarity (e.g., it does not permit differential weighting of coded units), it does serve as a benchmark for comparison. Only 76% of the FACS index dataset (stress = 0.17) was accounted for by a 2-D solution (not shown here). However, a 3-D solution accounted for 90% of the variance (stress = 0.09). As Figure 5 demonstrates, the pattern derived from the FACS solution was strikingly similar to that derived from the forced-choice recognition data. (MDS analyses based on data from the multiple-choice and open-ended conditions also show generally similar results.) Again, Russell's 2-D, pleasure × arousal, approach cannot account for the pattern observed here.

Since FACS coding is based on physical facial features and muscle movements, the close similarity of the human recognition and FACS index results may be taken to suggest that the dimensions of facial affect space may

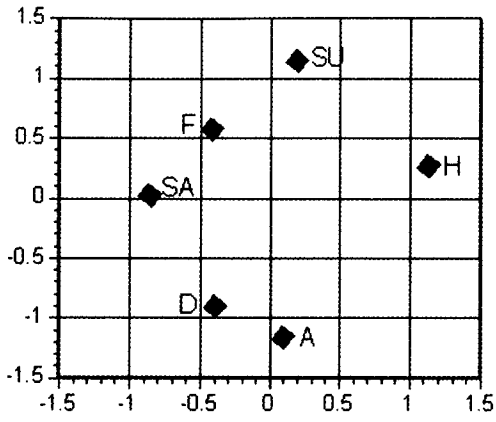

Figure 4. 2-D MDS solution for human (Experiment 1) dataset.

correspond more to physical or image parameters than to feeling states per se (or at least to salient physical parameters associated with feeling states). Still, FACS coding is complex, and the index we used was fairly crude; further research on this topic is clearly needed. First, however, we performed an experiment to more directly assess Russell's 2-D model of facial affect space, using the same stimuli as those in Experiment 1 and the response method that is commonly used in Russell's research.

EXPERIMENT 2

Russell's model of facial affect space (Russell, 1980) was presented above in conjunction with the results of the MDS analyses of the data from Experiment 1. If facial affect space is truly robustly characterizable in terms of two dimensions, pleasure and arousal, then these dimensions should have clearly emerged in the 2-D MDS solution for that dataset. Moreover, the 3-D solution should have contributed only incrementally toward accounting for the variability of the dataset. That this was not the case suggests that Russell's model (and similar approaches used by several other researchers; see Russell & Fernandez-Dols, 1997) may be in error. Experiment 2 was performed to more directly assess Russell's model, using the same stimuli as those created for Experiment 1, and the response method that is characteristic of Russell's research. The participants were asked to explicitly rate each stimulus image on two scales, pleasure and arousal.

Method

Participants. Eighteen Stanford University students between the ages of 18 and 35 years participated in this experiment.

Materials. The materials for this experiment were identical to those for Experiment 1 and were presented in the same way, using the same equipment and software. As in Experiment 1, each participant viewed all stimuli created by two (randomly chosen) actors.

Procedure. The participants viewed stimulus images depicting emotional expressions and rated each expression on Russell's two bipolar rating scales: *displeasure–pleasure* and *sleepiness–arousal*. The scales ranged from −3 to +3. The participants were instructed that the scales were designed with "0" as a neutral point, with −3 and +3 indicating the extremes, and −1 and +1 indicating slight amounts of the given dimension. The stimuli were viewed in random order at a distance of about 30 in. In each trial, the participant was shown the stimulus image on the TV monitor and simultaneously viewed a response screen on the computer monitor. The rating scale appeared as radial buttons adjacent to the each scale label. Ten ini-

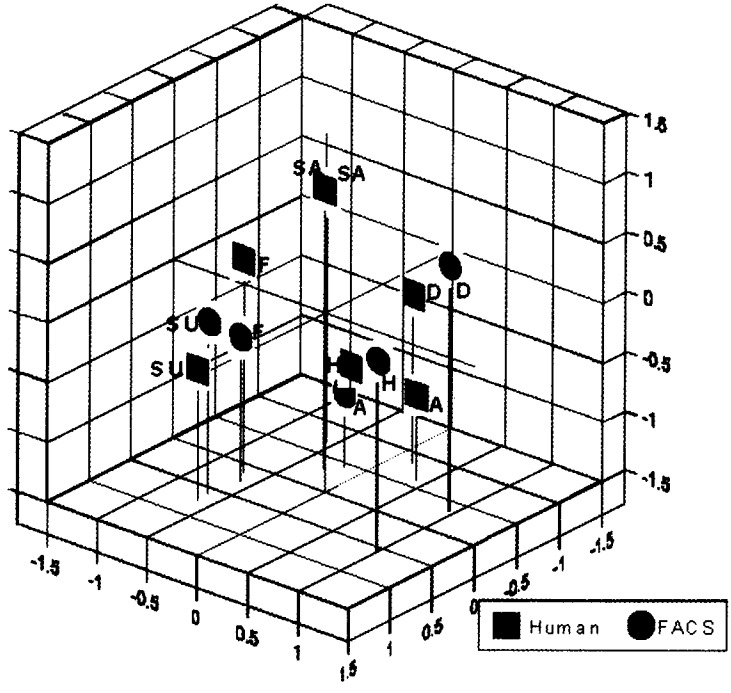

Figure 5. 3-D MDS solution for human (Experiment 1) and FACS datasets.

tial (nonfeedback) practice trials used randomly selected images from actors not viewed during the test trials. The experimental protocol was implemented in HyperCard on a PowerMac. The participants proceeded at their own pace; the entire procedure lasted under 1 h.

Results and Discussion

The mean ratings for pleasure and arousal were plotted for each emotion on orthogonal axes; the resulting pattern is shown in Figure 6. A circular pattern is found, and the sequence of emotions about the circle does correspond roughly to Russell's predictions. However, the close clustering of anger, fear, and disgust is problematic for Russell's 2-D approach and suggests the need for an added dimension to disambiguate the positions of each member of the cluster. Pilot testing in our laboratory suggests that adding another dimension ("dominance," after Mehrabian & Russell, 1974) may be sufficient to disambiguate these data (using "approach–avoidance" instead was not so helpful). That three dimensions may be needed to account for the complexity of facial affect perception—and the difficulty of identifying these dimensions—has been discussed in the psychological literature at least since Scholsberg's (1954; see also Picard, 1997; Russell & Fernandez-Dols, 1997; Schiano, Ehrlich, Sheridan, Beck, & Pinto, 1999) time. Comparing the results of this experiment with those of Experiment 1—which used the same stimuli—lends emphasis to the conclusion that Russell's model is insufficient to capture the complexity of human facial affect perception.

Taken together, the results of Experiments 1 and 2 argue against Russell's methodological criticisms of the classic studies on facial affect recognition and against his model of facial affect space. This has obvious importance for the psychological literature, but it also has direct implications for more applied work on facial affect. First, the use of physiological indices of arousal to validate emotions inferred from facial expressions (degree of pleasure either is assumed to be known or is derived off line from independent subjective reports) is becoming increasingly common in affective computing research (see Picard, 1997). The present findings suggest that the results of those studies may be misleading; at the least, they require close scrutiny. Second, the classic data on recognition of facial expressions were generally supported. This is good news for artificial intelligence researchers, who tend to use the FACS and Ekman's classic data to assess performance of models of facial affect recognition (see Lisetti & Schiano, 2000).

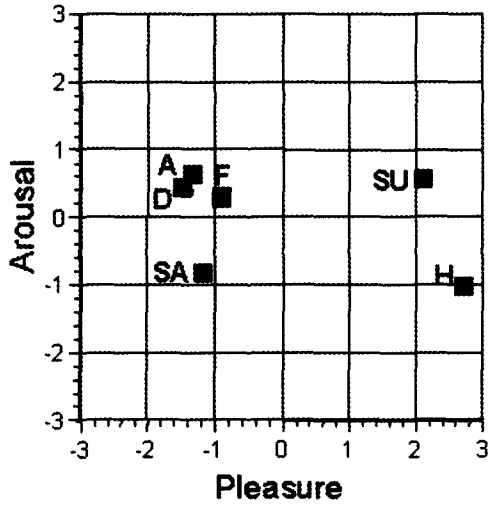

Figure 6. 2-D pleasure by arousal plots for each emotion in Experiment 2.

The final section of this paper describes a "user study" performed to inform a research project on robot facial affect at Interval Research Corporation. Such pragmatic studies, designed to test whether users perceive or deploy an artifact as intended, are standard practice in the field of human–computer interaction. This study also allowed us to begin to explore structural aspects of facial affect under simplified conditions. We include it here because it provides a real-world example of how basic and more applied work can prove mutually informative.

EXPERIMENT 3
User Study

As discussed above, the observation of a close correspondence between the MDS solutions for the data from Experiment 1 and from our index of FACS code overlap suggests that the underlying dimensions of facial affect space may in fact be physical or image-based. That certain structural characteristics of the face may explain judgments of emotion categories better than feeling states per se has been suggested previously. Katsikitis (1997) focused on the relative dominance of different parts of the face; Yamada and colleagues (Yamada, Matsuda, Watari, & Suenaga, 1993) emphasized the presence of curved or slanted lines. Further research on this topic is clearly needed. One approach would be to explore structural aspects of facial affect under simplified conditions, which could shed some light on the most salient cues for emotional expression. This was the approach taken in the user study described here.

Simplified or schematic faces are used to express emotion in many computer interfaces, from icons to virtual pets. Recent work has shown that cartoon-like facial icons are sufficient to serve as an affective interface for interactions with avatars and autonomous computer agents (Kurlander, Skelly, & Salesin, 1996). Experiment 3 was performed in the context of user-testing an early design prototype of a mechanical robot face. While made of metal, the face was intended to be cartoon-like in nature. Created by an independent group of researchers at Interval Research Corporation (see Tow, 1998), the robot was designed largely through intuition inspired by cartoon animation principles (e.g., Hamm, 1982) and some reading of Darwin (1965/1872) on emotions. This initial

model was constructed largely as a "proof of concept" to demonstrate that a very simplistic robot face could effectively express a range of emotions to varying degrees. A similar face would be incorporated into a later, more complete model, capable of whole-body movement and expression.

The initial prototype consisted of a box-like face containing eyes with moveable lids, tilting eyebrows, and upper and lower lips that could be independently raised or lowered from the center. Figure 7 shows the face displaying the six basic emotional expressions. The robot facial features are extremely sparse, and their motion is highly constrained relative to the subtle detail and mobility of the human face, as seen in Figure 1. The robot face has no skin, so the telltale folds, lines, and wrinkles specifying many FACS codes are simply not available. The motion of the features (especially the eyebrows and lips) is only schematically related to human facial muscle movements.

Experiment 3 was designed to inform the design team and was conducted in close collaboration with them. They created the stimuli and helped implement the study. The team was primarily interested in achieving some assurance that users felt satisfied with the robot's ability to express a range of emotions. Another goal was to obtain feature settings for various emotional displays, to be stored as templates so that the prototype could be programmed to quickly display emotional responses as desired. We were especially interested in comparing the robot affect space with that derived from our human data (Experiment 1), after verifying that the displays were indeed correctly recognized by an independent group of observers.

Method

Participants. Eighteen participants between the ages of 18 and 35 years were involved in this study. Nine Interval Research Corporation employees participated in the first condition of the study, and 9 Stanford University students participated in the second condition.

Materials. The robot face consisted of a 12 × 14 cm mechanical metal (primarily aluminum) face with independently moveable eyelids, eyebrows, and upper and lower lips (see Figure 7). The eyelids were small metal sheets that could move up or down. The eyebrows were metal bars, placed with a pivot point toward each side of the face, to allow rotations between horizontal and vertical positions. Each lip consisted of a spring fixed at both ends and with a tie in the center that could be pulled up or down (stretching the spring on both sides). Each feature was controlled by a computerized motor with 255 possible positions. For all features (except the eyelids), the neutral position was in the center of the range of motion. The neutral position for the eyelids was fully open (or up).

Procedure. The procedure consisted of two conditions, feature-setting adjustments and recognition validation.

Feature-setting adjustments condition. In each of two blocks of trials, instructions on the computer monitor asked the participant to set the features of the robot face to express each of the six emotions (angry, disgusted, fearful, happy, sad, or surprised) at each of three degrees of intensity (slightly, moderately, or very), twice, in random order. The participants adjusted feature positions by pressing keys on a computer keyboard; "up" and "down" keys were labeled on the keyboard for each of the four features. On completion of each expression, the participants rated their overall satisfaction with the expression on a scale of 1 (*least*) to 5 (*most*). Each trial began with the features in the neutral position, except for the eyelids, which were closed. Each testing session began

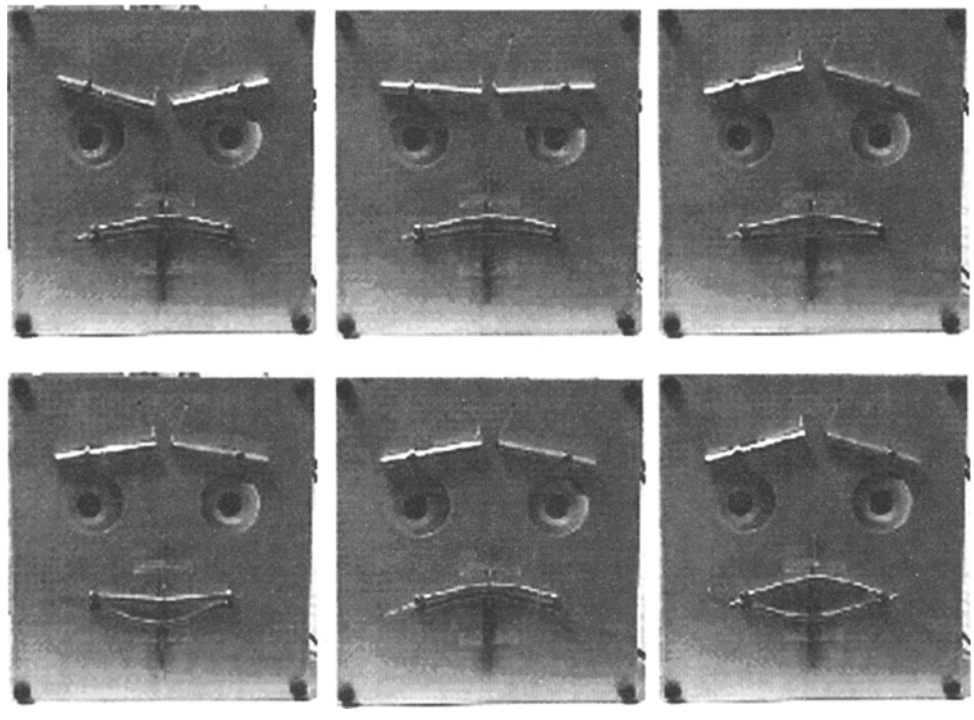

Figure 7. Average robot face settings for each for each emotion (anger, disgust, fear, happiness, sadness, surprise) in Experiment 3.

with 10 randomly chosen (nonfeedback) practice trials. The participants proceeded with this task at their own pace, and the entire procedure took less than 1 h. The robot face was attached to a Toshiba laptop PC, which implemented the experimental protocol.

Recognition validation condition. The participants viewed robot facial expressions as obtained from the feature-setting adjustments condition. The expressions were given by the mean feature settings for each emotion (angry, disgusted, fearful, happy, sad, or surprised) at each of the three degrees of intensity (slightly, moderately, or very), plus the "average" setting. The average setting for each emotion is depicted in Figure 7. In all, four exemplars of each emotion were shown, three times, in random order. For each trial, the participants used the forced-choice (plus ratings) response method of Experiment 1, choosing the one term from an alphabetized list of basic emotion labels that best described the expression of the robot face. They then rated the degree to which the emotion was present in the robot face on a scale of 0 (*not at all*) to 6 (*extremely high*). In this condition, the robot facial expressions were controlled by a Toshiba laptop PC while the rest of the experimental protocol was implemented in HyperCard on a PowerMac computer. The participants proceeded at their own pace; the entire procedure took about 30 min.

Results and Discussion

The results for the feature-setting adjustments condition are given in terms of numerical setting values for each feature, which are difficult to summarize succinctly except pictorially. Figure 7 shows the "average" display for each emotion, derived from the mean feature settings for each participant across the three degrees of emotional intensity. The figure does suggest that the interface was capable of expressing various emotions, although perhaps not all equally well. While the participants' satisfaction ratings were fairly high overall (M = 3.7 out of 5, for all degrees of intensity), satisfaction with disgust, in particular, was fairly low (M = 2.9 out of 5). The FACS codes characterize disgust by the drawing up of the nasal-labial muscles, producing striking patterns of wrinkles and folds around the mouth and nose. However, the robot face has neither nose nor skin. That disgust was found to be especially difficult to express was not especially surprising.

The mean correct recognition scores for each of the emotions (averaged over degree of intensity) are shown in Figure 8, in comparison with the human data from Experiment 1. The scores for the robot are generally somewhat lower than those for human faces (especially for disgust), but this is not very surprising. First, the schematic nature of the robot face (as described above) should have made it more difficult to express emotion, relative to human faces. Second, these scores were averaged over stimuli intended to depict emotion at varying intensity levels; the human actors presumably at least intended to create stimuli that showed each emotion to a high degree of intensity. Third, due to time constraints, our sample size was small, and so the dataset is fairly variable. That the human and robot results were nonetheless so close is noteworthy.

Intensity ratings served as a manipulation check in the recognition validation condition of this study, as in Experiment 1. The mean ratings over all emotions were moderately high (M = 4.34 out of 6). Further analyses (available on request) found that the ratings did generally vary with the intensity of the depicted emotion. And, as expected, recognition of the emotions tended to increase with rated intensity.

MDS analyses were performed on the robot display data. The mean direction and amount of movement of each of the four facial features (taken from the average of all expressions) for each emotion were used to generate 4-D vectors; the distance between each vector gave the (dis-)similarity matrix for the MDS analysis of the robot dataset. Ninety-seven percent of the variance of the robot data was accounted for by a 2-D MDS solution (stress = 0.06). Figure 9 presents the 2-D solution for the robot dataset plotted with the 2-D human recognition pattern obtained in Experiment 1. The close similarity of the patterns is immediately obvious, with ordering of emotions identical for the two datasets.

Ninety-nine percent of the variance of the robot dataset was accounted for by a 3-D MDS solution (stress = 0.02). Figure 10 shows the 3-D solutions for the robot dataset plotted with the 3-D pattern from Experiment 1. The similarity of the patterns is remarkable when con-

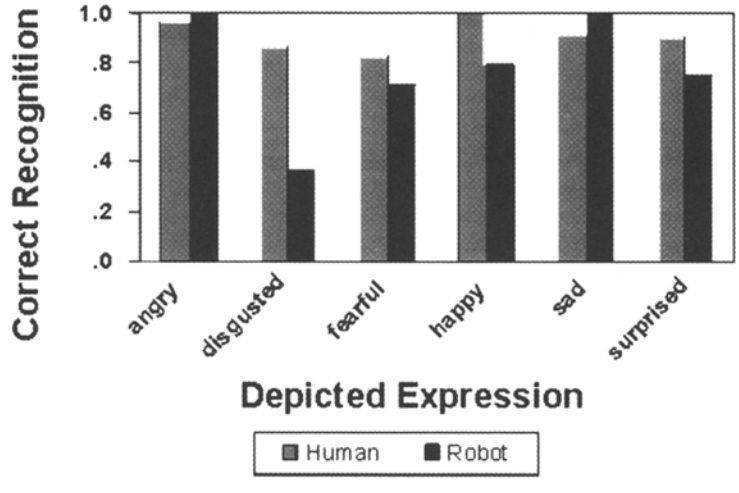

Figure 8. Correct recognition scores for each emotional expression for human (Exeperiment 1) and robot (Experiment 3) data sets.

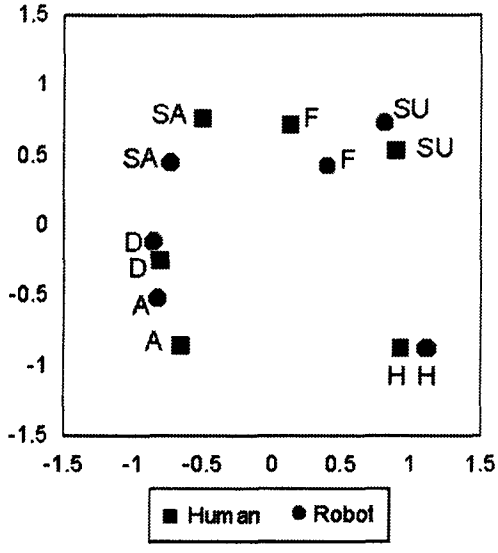

Figure 9. 2-D MDS solutions for human (Experiment 1) and robot (Experiment 3) datasets.

sidering the disparity of the stimuli and the fact that the robot data were plotted directly from the feature-setting parameters. As in the case of the human data, the robot results map easily onto the FACS index pattern (see Figure 4) but not so easily onto Russell's model (see Figure 3). Moreover, the similarity of findings across the human, FACS, and robot datasets further supports the notion that the dimensions of the facial affect space might correspond most closely to physical or image parameters—indeed, to very simple ones. Our initial speculation is that the primary axis may correspond to concavity/convexity of the lips and that the second may correspond to the upward/downward tilt of the eyebrows. The third dimension is less clear but may be related to the set of the mouth, perhaps its degree of openness (note that many of our disgust stimuli had open mouths). Further research is clearly needed, but these results do suggest implications for many applications in which the complexity of the face is constrained or compressed. We are currently looking at human facial affect under a variety of compressed image conditions, to see whether a similar affect space is found.

GENERAL DISCUSSION

Experiment 1 provided new baseline data on human facial affect recognition, using improved experimental methods and somewhat more naturalistic stimuli than those of the classic studies. The pattern of results for the forced-choice response format closely replicated Ekman's classic findings, and (except for fear) this was generally true for the alternative response formats as well. Thus, on the whole, Russell's criticisms are not borne out by the data. Our MDS analyses suggest that 3 dimensions

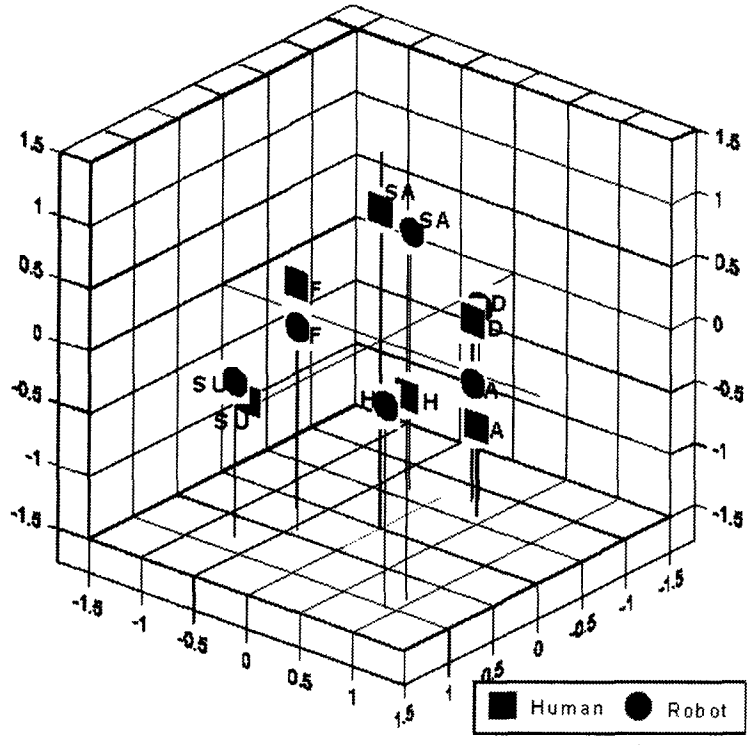

Figure 10. 3-D MDS solution for human (Experiment 1) and robot (Experiment 3) datasets.

are substantially better than two in specifying facial affect space; however, even the 2-D solution does not match Russell's model. Indeed, our data match the FACS-based solutions much more closely. We find this intriguing, suggesting that the dimensions of facial affect may be based more on physical or image parameters than on feeling states (such as pleasure and arousal) per se.

Experiment 2 was performed to directly test Russell's pleasure versus arousal model of facial affect space by using the same stimuli as those of Experiment 1 and the response method commonly used in Russell's research. When the participants were asked to rate each stimulus expression on pleasure and arousal dimensions, the pattern of results did not show clear support for Russell's model. Instead, the pattern of results suggests that a third dimension may be needed to disambiguate three of the six emotions.

Finally, we report a user test whose primary aim was to inform the designers of the affective robot face. We succeeded both in demonstrating that the robot face was sufficient to communicate various emotional expressions and in providing setting templates for specific emotions of varying intensities. The revised prototype of the affective robot incorporates a face very similar to the one we tested. The pattern of results for this study was strikingly similar to our human data despite extreme schematization of the robot face, far fewer participants, and various design differences between the studies. The similarity of the MDS solutions for robot, human, and FACS-based data underscore the notion that physical or image-based parameters—perhaps very simple ones—could be used to interpret the dimensions of facial affect space. Some speculations on what those parameters may be were provided above. Interestingly, 3-D models of affect have been suggested before (e.g., Schlosberg, 1954; see also Picard, 1997, and Russell & Fernandez-Dols, 1997), largely based on feeling states, but no consensus in axis interpretation was found in that earlier research.

This paper presents some initial findings from a large research effort on perception of facial affect. Some of the work described here was also presented in Schiano, Ehrlich, Rahardja, and Sheridan (2000) and Ehrlich et al. (1998). Related work includes a systematic study of the evidence for categorical perception of facial affect (Schiano et al., 1999) and a collaboration with another laboratory in training a neutral-net AI model on our stimuli, to see what features it picks up (see Lisetti & Schiano, 2000) and Ehrlich, Schiano, Sheridan, & Beck (1998). We have also explored the effects of various compression techniques on perceived facial affect in still and moving images (e.g., Ehrlich, Schiano, & Sheridan, 2000). The corpus of stimuli created for these and related studies (Ehrlich, Sheridan, & Schiano, 2000) is available on request.

This is an exciting time for research on facial affect in both humans and machines. We hope this paper helps demonstrate the importance of the systematic study of methods—and the mutual informativeness of basic and applied research—in this rapidly growing field.

REFERENCES

Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, **36**, 253-263.

Darwin, C. (1965). *The expression of emotion in man and animals*. Chicago: University of Chicago Press. (Original work published 1872)

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

Ehrlich, S. [M.], Schiano, D. [J.], Sheridan, K., & Beck, D. (1998, November). *Facing the issues: Methods matter*. Poster session presented at the annual meeting of the Psychonomic Society, Dallas.

Ehrlich, S. M., Sheridan, K., & Schiano, D. J. (2000). *Corpus of facial affect stimuli* (Interval Research Corporation External Tech. Rep.). Palo Alto, CA: Interval Research Corp.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P., & Friesen, W. V. (1978). *The facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face*. New York: Pergamon.

Hamm, J. (1982). *Cartooning the head and figure*. New York: Perigree.

Katsikitis, M. (1997). The classification of facial expressions of emotion: A multidimensional-scaling approach. *Perception*, **26**, 613-626.

Kurlander, D., Skelly, T., & Salesin, D. (1996). Comic chat. In *Proceedings of SIGGRAPH 96, Computer Graphics Proceedings, Annual Conference Series* (pp. 225-236). New York: ACM.

Lisetti, C. L., & Schiano, D. J. (2000). Automatic facial expression interpretation: Where human–computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics & Cognition*, **8**, 185-235.

Mehrabian, A., & Russell, J. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality & Social Psychology*, **39**, 961-1178.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? *Psychological Bulletin*, **95**, 102-141.

Russell, J. A., & Fernandez-Dols, J. M. (1997). *The psychology of facial expression*. New York: Cambridge University Press.

Schiano, D. J., Ehrlich, S. M., Rahardja, K., & Sheridan, K. (2000). Face to interface: Facial affect in (hu)man and machine. In *Proceedings of ACM CHI 2000 Conference on Human Factors in Computing Systems* (pp. 193-200). New York: ACM.

Schiano, D. J., Ehrlich, S. M., Sheridan, K., Beck, D., & Pinto, J. (1999, November). *Evidence for continuous perception of facial affect*. Paper presented at the annual meeting of the Psychonomic Society, Los Angeles.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, **69**, 81-88.

Tow, R. (1998, November). *Affect-based robot communication methods and systems* (U.S. Patent No. 5,832,189).

Yamada, H., Matsuda, T., Watari, C., & Suenaga, T. (1993). Dimensions of visual information for categorizing facial expressions of emotion. *Japanese Psychological Research*, **35**, 172-181.

(Manuscript received November 4, 1999;
revision accepted for publication March 3, 2000.)