Temporal Context and the Recognition of Emotion from Facial Expression

Rana El Kaliouby[1]*, Peter Robinson*[1]*, Simeon Keates*[2]

[1]Computer Laboratory
University of Cambridge
Cambridge CB3 0FD, U.K.
{rana.el-kaliouby, peter.robinson}@cl.cam.ac.uk

[2]Engineering Design Centre
University of Cambridge
Cambridge CB2 1PZ, U.K.
lsk12@cam.ac.uk

Abstract

Facial displays are an important channel for the expression of emotions and are often thought of as projections or "read out" of a person's mental state. While it is generally believed that emotion recognition from facial expression improves with context, there is little literature available quantifying this improvement. This paper describes an experiment in which these effects are measured in a way that is directly applicable to the design of affective user interfaces. These results are being used to inform the design of *emotion spectacles*, an affective user interface based on the analysis of facial expressions.

1 Motivation

Facial displays are an important channel for the expression of emotions and are often thought of as projections or "read out" of a person's mental state (Baron-Cohen et al., 2001). It is therefore not surprising that an increasing number of researchers are working on endowing computing technologies with the ability to make use of this readily available modality in inferring the users' mental states (Colmenarez et al., 1999; Picard and Wexelblat, 2002; Schiano, 2000).

To design the *emotion spectacles*, an emotionally intelligent user interface, we examined the process of emotion recognition in humans. Whereas most existing automated facial expression recognition systems rely solely on the analysis of facial actions, it is theorized that humans make additional use of contextual cues to perform this recognition (Bruce and Young, 1998; Ratner, 1989; Wallbott, 1988). Existing literature, however, falls short of quantifying the type or amount of context needed to improve emotion perception in a form applicable to computer interface design. In this paper, we investigate the effect of temporal facial-expression context on emotion recognition.

2 The Emotion Spectacles and Related Work

The *emotion spectacles* are designed to identify and respond to a user's affective state in a natural human-computing environment. This could be integrated into a wide range of applications ranging from ones that respond to user frustration, ones that gauge the level of user engagement during an online learning task, and virtual salesmen that learn an online shopper's preferences from his/her reaction to the offered goods.

The process with which the *emotion spectacles* infer a user's mental state is two-fold and involves facial action analysis followed by emotion recognition. For the analysis stage, we use dynamic facial action analysis, which identifies and analyzes facial motion in video sequences using feature point tracking. The emotion recognition stage however, has received little attention to date with the exception of Calder et al. (2001) and Colmenarez et al. (1999) who classify facial actions into a number of basic emotions. Our goal in *emotion spectacles* is to automatically infer a wider range of emotions. We propose combining the analyzed facial motion with other contextual information in order to do so.

3 Putting Facial Expressions in Context

A number of studies indicate how humans make considerable use of the contexts in which expressions occur to assist interpretation (Bruce and Young, 1998). Ratner (1989) shows that the reliability of facial expressions as indicators of emotion is significantly improved when they are perceived in relation to contextual events instead of in isolation. Wallbot (1988) found that the same expression perceived in varying contexts was judged to indicate different emotions.

In order to quantify the effect of context, we follow Edwards' (1998) description of an expression of emotion as a sequence of concurrent sets of facial actions, or micro expressions. In addition, we define temporal facial-expression context as the relationship between consecutive micro expressions. It is temporal, because it represents the transition, in time, between two consecutive micro expressions. It is context, because micro expressions are judged to indicate different emotions when perceived with respect to preceding ones, versus in isolation (Edwards, 1998).

In this paper, we examine the effect of temporal facial-expression context on the recognition accuracy of both basic and complex emotions. Basic emotions include happy, sad, angry, afraid, disgusted, surprised and contempt. Emotions such interest, boredom, and confusion, on the other hand, involve attribution of a belief or intention–a cognitive mental state–to the person, and are hence referred to as complex (Baron-Cohen et al., 2001).

4 Experiment

The goal of this experiment was to investigate the effect of temporal facial-expression cues on the recognition accuracy of both basic and complex emotions. We specifically addressed the following questions:

1. To what extent does temporal context have an impact (if any) on recognition of emotions from facial expressions?
2. Is temporal facial-expression context equally effective in improving (if at all) the recognition accuracy of both basic and complex emotions?
3. What relationship exists between the amount of context and degree of improvement, whether this relationship has critical inflection points and whether it tapers off with additional context becoming irrelevant?

4.1 Video Material

Material used throughout the experiment was developed using videos from "Mind Reading", a computer-based interactive guide to emotions (Human Emotions Ltd., 2002). The resource has a total of 412 emotions organized into 24 groups. Six video clips are provided for each emotion showing natural performances by a wide range of people. We picked 16 videos representing 13 complex emotions (such as confused, enthusiastic, and undecided), and 8 videos representing 6 basic ones (such as afraid, angry, and disgusted), for use throughout the experiment. The duration of the videos vary between 3 to 7 seconds (mean= 5.24, SD= 0.45).

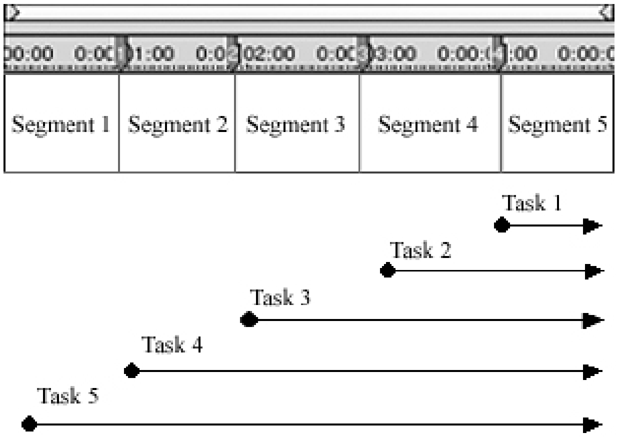

Figure 1 The process of clip construction showing how segments of a video are joined to form the clips used during each of the five experimental tasks.

Each video was divided into five separate segments, such that each segment is composed of essentially different micro expressions. The segments were then joined to construct the clips used during each of the 5 experimental tasks. The process of clip construction is illustrated in Figure 1. The last segment of every video (segment 5) is shown during the 1st task. Clips for the 2nd task are constructed by combining the 4th and 5th segments of a video and so on.

4.2 Experimental Variables

Two independent variables were defined. Clip span (5 conditions) defines the span of a video clip, and emotion category (2 conditions) could either be basic or complex. Accuracy of recognition measured in percentage of correctly identified emotions was the dependent variable.

4.3 Participants

36 participants between the ages of 19 and 65 (mean= 31, SD= 11) took part in the experiment, 21 males and 15 females. Participants were either company employees who covered a wide range of occupations or university research members (mostly computer science). Participants were of varied nationalities, but all had substantial exposure to western culture. All participated on a voluntary basis.

4.4 Experimental Tasks and Procedure

We designed a total of six tasks, five experimental and one control. The experimental tasks were designed to test the effect on recognition accuracy of the last segment of each video, when gradually adding earlier segments. The control task tested that for each emotion, none of the five segments played a greater role in "giving away" an emotion.

During each of the tasks, participants viewed 24 video clips of varying length and were asked to identify the underlying emotion. A forced-choice procedure was adopted, where three foil words were generated for each emotion (for a total of 4 choices on each question). All participants were asked to carry out all five tasks making this a within-subject repeated measures study. This set-up minimizes differences in the task responses that can be attributed to varying emotion-reading abilities of the participants. A typical test took 40 minutes on average to complete, and generated 120 trials per participant. Tasks were carried out in increasing order of clip length to prevent any memory effect. The order in which the videos were viewed within each task was randomized. During the control task, participants were randomly shown segments of each of the 24 videos.

5 Results

Twenty-eight participants, each performing 5 experimental tasks for 24 videos produced a total of 3360 trials in total. Nine hundred and sixty trials were conducted for the control task.

A pair-wise analysis of the complex emotion samples show a statistically significant ($p<0.0001$) improvement of 25.8% in accuracy, moving from clip span 5 to clip span 45. A smaller but statistically significant ($p<0.017$) improvement of 9.4% is observed between clip span 45 and clip span 345. We were surprised to find that the percentage improvement between clip span 345 and 2345 is almost negligible (0.5%) and is not statistically significant ($p<0.9$). Similarly, the improvement seen in moving from clip 2345 to 12345 is also negligible (0.6%) and statistically insignificant ($p<0.75$). The responses of basic emotions show negligible improvement between clip spans (mean improvement =2.8%, SD=0.8) and the differences were not statistically significant. Analysis of the results from the control task showed no statistically significant difference ($p<0.2$) between the recognition accuracy of any of the 5 segments of both the basic and complex emotions.

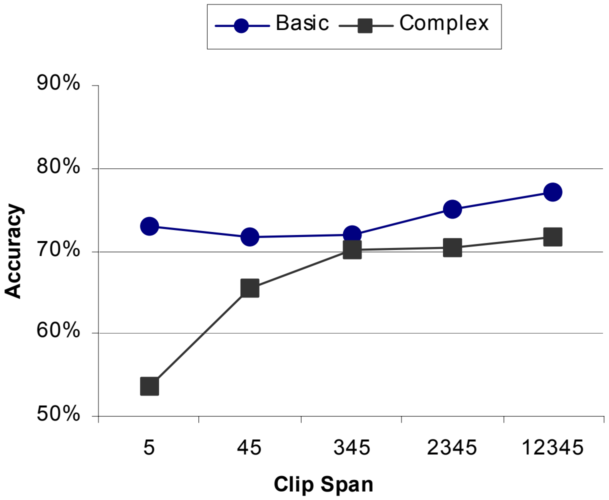

Figure 2: Effect of clip span on recognition accuracy in the case of basic and complex emotions.

The pattern of improvement in accuracy is summarized in Figure 2. In the case of basic emotions, the recognition is nearly constant. In the case of complex emotions, a small amount of context yields a pronounced improvement in recognition accuracy. This correlation however is not linear: as more temporal facial-expression context is added, the percentage of improvement tapers off.

Although we predicted a pronounced improvement in the case of complex emotions, we had also anticipated some impact, even if less obvious, in the case of basic ones. We were somewhat surprised that there was no significant change in accuracy to report during the experiment.

6 Implications on the Emotion Spectacles

Our experimental results have significant implications for the design of affective user interfaces, especially those based on facial affect. To start with, the pronounced improvement of accuracy associated with the addition of temporal facial-expression context suggests integrating context in the design of emotionally intelligent interfaces. Whereas prevalent classification methodologies treat every micro expression in isolation of preceding ones, we suggest that classification should use the result of prior inferences in recognizing the current expression. Our findings in this experiment suggest that the two immediately preceding micro expressions are responsible for all of the statistically significant improvement. This has favourable implications on the complexity from a design point of view.

7 Conclusion

Our work presents a multidisciplinary study on the effect of context on the recognition accuracy of both basic and complex emotions. We show that a relatively small amount of temporal facial-expression context has a pronounced effect on recognition accuracy in the case of complex emotions, but no corresponding effect is seen in the case of basic emotions. The results are used to inform the design of *emotion spectacles*, a facial-expression based affective user interface currently under development, but can also be utilized in the design of embodied conversational agents, avatars, and in computer-mediated communication.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001) The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry,* 42 (2), 241-251.

Bruce, V. and Young, A. (1998) In the Eye of the Beholder: The Science of Face Perception. Oxford University Press.

Calder, A.J., Burton, A. M., Miller, P., Young, A.W., Akamatsu, S. (2001) A principal component analysis of facial expressions. *Vision Research,* 41, 1179-1208.

Colmenarez, A., Frey, B., and Huang, T. (1999) Embedded Face and Facial Expression Recognition, *International Conference on Image Processing*.

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

Human Emotions Ltd. (2002) Mind Reading: Interactive Guide to Emotion. http://www.human-emotions.com

Picard, R.W., & Wexelblat, A. (2002) Future Interfaces: Social and Emotional. *Extended Abstracts of The CHI 2002 Conference on Human Factors in Computing Systems*, ACM Press, 698-699.

Ratner, C. (1989) A Social Constructionist Critique of Naturalistic Theories of Emotion. *Journal of Mind and Behavior*, 10, 211-230.

Schiano, D.J., Ehrlich, S., Rahardja, K., and Sheridan, K. (2000) Face to Interface: facial affect in (hu) man machine interaction. *Proceedings of CHI 2000 Conference on Human Factors in Computing Systems,* ACM Press, 193-200.

Wallbott, H.G. (1988) In and out of context: Influences of facial expression and context information on emotion attributions. *British Journal of Social Psychology*, 27, 357-369.