The role of image size in the recognition of conversational facial expressions

Douglas W. Cunningham and **Manfred Nusseck** and **Christian Wallraven** and **Heinrich H. Bülthoff**
Max Planck Institute for Biological Cybernetics
Phone: +44 7071 601 725
Email: {firstname.lastname}@tuebingen.mpg.de

Abstract

Facial expressions can be used to direct the flow of a conversation as well as to improve the clarity of communication. The critical physical differences between expressions can, however, be small and subtle. Clear presentation of facial expressions in applied settings, then, would seem to require a large conversational agent. Given that visual displays are generally limited in size, the usage of a large conversational agent would reduce the amount of space available for the display of other information. Here, we examine the role of image size in the recognition of facial expressions. The results show that conversational facial expressions can be easily recognized at surprisingly small image sizes.

Keywords: computer graphics, conversational agents, facial expressions, image size

Introduction

Facial motion plays a complex and important role in communication. It can be used to modify the meaning of what is said[1, 2, 3, 4, 5], or to express meaning by itself[6, 7, 8]. Facial motion can also be used to help direct the flow of a conversation[9, 10, 11, 12, 13]. Thus, it should not be surprising that proper usage of facial communication can be advantageous for Human-Computer Interfaces (HCI). For example, Cassell and colleagues [11, 12] have created agents that utilize head motion and eye gaze to help control turn-taking. It should even be possible to provide more subtle conversational aid through the use of facial expressions as *back-channel* information[14, 15]. For example, when a listener nods in agreement, the speaker knows that they were understood and can continue. A look of confusion, disgust, or boredom on the part of the listener, however, will prompt very different behaviour on the part of the speaker. The synthesis of a conversational agent that is capable of properly employing the subtleties of human facial communication could improve the efficiency and usefulness of multimodal HCIs.

The differences that allow us to distinguish one expression from another can be quite small. To ensure that these subtle differences are visible to the user of an HCI, one might wish use as large a conversational agent as possible. Of course, in most applied settings, the visual display is not infinitely large. Thus, the larger the interface agent is, the less room there will be for visual presentation of other information. What, then, is the optimal size for a conversational agent?

Here, we present a psychophysical experiment designed to determine exactly how small a conversational agent can be without sacrificing clarity and understandability. To ensure that the expressions were as recognizable as possible, we used video sequences of real individuals in this initial experiment. We recorded nine conversational expressions (agree, disagree, pleased/happy, sad, thinking, confused, clueless, pleasant surprise, and disgust) from six different individuals. The image sequences were present one at a time in the center of a computer screen, and a number of viewers were asked to identify the expression. The results were somewhat surprising: Even when the images were a mere 64 by 48 pixels (the face subtended about 2 degrees of visual angle), the expressions were as easy to recognize as they were at very large image sizes (512 by 384 pixels).

Experimental Methods

A series of expressions were recorded from six different individuals (one professional and five amateur actors and actresses) using the Max Planck Institute for Biological Cybernetics's VideoLab (for detailed information on the VideoLab and the recordings, see [7, 16]). Each expression was recorded three times, with a relaxed, neutral expression preceding and following each repetition. The present experiment used nine of the recorded expressions: agree, disagree, pleased/happy, sad, confused (as if the actor did not understand what was just said), thinking, clueless (as if the actor did not know the answer to a question), pleasant surprise, and disgust. None of the recordings contained speech or visible hand motion.

In a pilot experiment, the most recognizable and believable repetition was found for each expression for each actor/actress. Each of the 54 video sequences (nine expressions from six individuals) was rescaled from its original size of 768 x 576 pixels. Six different size conditions were used: 512 by 384 pixels (approximately 20 by 15 degrees of visual angle), 256 by 192, 128 by 96, 64 by 48, 32 by 24, and 16 by 12 pixels. The face did not fill entire image. For example, at an image size of 16 by 12, the face covered a mere 8 by 6 pixels.

Crossing nine expressions from six individuals with six image sizes yielded 324 trials. The 324 video sequences were shown in a randomized order to 10 individuals (hereafter referred to as *participants*) in a psychophysical experiment. None of the participants had previously seen any of the recordings. Each participant was seated in front of a computer monitor, with his/her head positioned exactly 50 cm from the

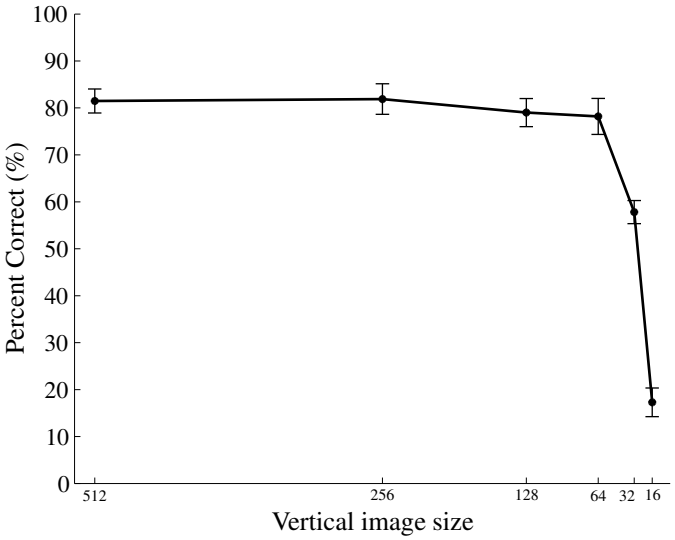

Figure 1: Overall recognition accuracy. The percentage of the time that the participants correctly identified the expressions is shown for the six image size conditions. The error bars represent the standard error of the mean.

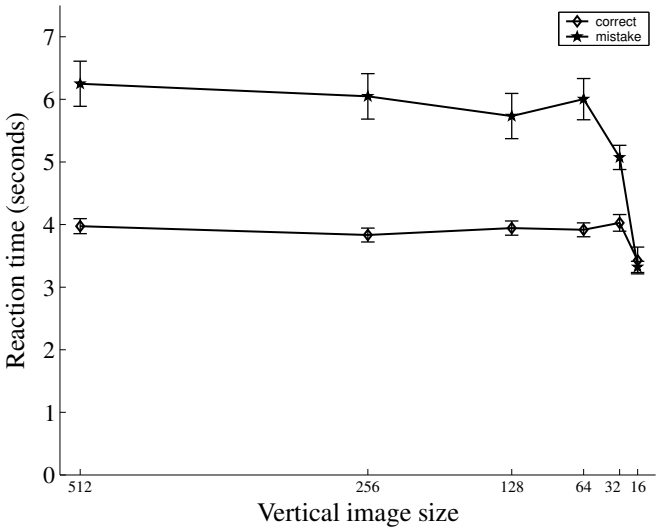

Figure 2: Overall reaction times. The amount of time it took the participants to respond is shown for the six image size conditions. The error bars represent the standard error of the mean.

screen (the distance from the monitor was fixed using a chin rest). At the beginning of each trial, the participant pressed a button and a video sequence was shown in the exact center of the screen. After the video sequence ended, it was removed from the screen. The participants were asked to identify the expression in the video sequence by selecting an entry from a list that was displayed on the side of the screen at all times. The list contained the names of the nine expressions as well as a 'none of the above" option. The participants could enter their response at any time.

Results and Discussion

Figure 1 shows the mean recognition accuracies for the six image sizes. Overall, the expressions were easy to identify, despite the absence of a conversational context. This is consistent with previous work with conversational expressions[6, 7]. Recognition accuracy was surprisingly insensitive to changes in image size: Participants could recognize the expressions just as well at an image size of 64 by 48 pixels (where the face covered a mere 768 pixels) as at an image size of 512 by 384 pixels (where the face covered over 49,000 pixels). Equally striking is that performance suddenly dropped when the images were reduced below a size of 64 x 48 pixels. It is perhaps interesting to note that at 64 x 48 pixels, the face subtended approximately 2 degrees of visual angle, which is the size of the human fovea (i.e., the area of the retina that has the highest spatial resolution).

Although recognition accuracy was just as good in the 64 pixel condition as it was for larger image sizes, this does not necessarily mean that the task was just as easy. It is possible, for example, that decreases in image size increases the difficulty in extracting relevant facial information. Moreover, some facial information (e.g., direction of eye gaze) may be undetectable or ambiguous at smaller image sizes, and therefore one might expect decisions about the expressions to take longer. In other words, it might take participants longer to react to smaller images than larger images. As can be seen in Figure 2, reaction times (defined as the amount of time that elapsed between the onset of the video sequence and point when a participant entered a response) do not change as the image size decreases from the 512 pixel condition to the 64 pixel condition. Thus, it seems that these nine expressions are as easy to identify with an image size of 64 by 48 pixels as with an image size of 512 by 384 pixels (at least with the task used here).

At first glance, it is somewhat surprising that the participants were actually *faster* at the two smallest image sizes (where they had difficulty recognizing the expressions) than at the higher image sizes. This decrease in reaction time is particularly large for those trials where the expression was not correctly identified. One potential explanation for this is that, at the smallest image sizes, the participants had a tendency to simply give up when they did not know what the expression was. Spontaneous reports by the participants support this explanation.

Tables 1 through 6 show the confusion matrixes for the six image sizes. In general, the pattern of confusions at the larger image sizes is similar to those found in previous work with conversational expressions[6, 7]. Consistent with the patterns seen for the overall means (Figure 1), the pattern of confusions shows little change between the 512 and the 64 pixel conditions. In the 32 pixel condition, there is a massive drop in accuracy for all expressions except disagreement. Unlike the other expressions, disagreement showed an *increase* in recognition accuracy as the image size was reduced. Why is disagree getting better while everything else is getting worse? The obvious explanation is that this increase represents a response bias: In the 32 and 16 pixel conditions, participants responded disproportionately often with "disagree". Note, however, that the participants had the option of saying the expression was "none of the above" and that they often used this option. This suggests that the high level of accuracy for

	Agree	Disagree	Happy	Sad	Clueless	Thinking	Confused	Disgust	Surprise	Other
Agree	**98**	2	0	0	0	0	0	0	0	0
Disagree	2	**74**	0	0	13	0	2	2	2	6
Happy	0	2	**87**	2	0	2	0	0	7	0
Sad	0	0	0	**89**	4	2	4	0	0	2
Clueless	0	4	0	4	**72**	6	7	0	2	6
Thinking	0	0	0	2	6	**91**	2	0	0	0
Confused	2	6	0	0	13	0	**63**	7	4	6
Disgust	0	0	0	6	2	0	17	**70**	0	6
Surprise	0	0	7	0	0	0	0	0	**89**	4

Table 1: Confusion Matrix for images that were 512 by 384 pixels. The percentage of the time a given response was chosen (columns) is shown for each of the nine expressions (rows).

	Agree	Disagree	Happy	Sad	Clueless	Thinking	Confused	Disgust	Surprise	Other
Agree	**98**	2	0	0	0	0	0	0	0	0
Disagree	0	**80**	0	0	11	0	6	4	0	0
Happy	0	0	**83**	0	0	4	0	0	11	2
Sad	0	0	0	**91**	2	4	4	0	0	0
Clueless	2	4	0	2	**80**	4	4	0	2	4
Thinking	0	0	0	0	9	**87**	2	0	0	2
Confused	0	4	0	0	22	6	**54**	2	4	9
Disgust	0	0	0	7	4	2	7	**74**	0	6
Surprise	0	0	4	0	0	0	0	2	**91**	4

Table 2: Confusion Matrix for images that were 256 by 192 pixels.

	Agree	Disagree	Happy	Sad	Clueless	Thinking	Confused	Disgust	Surprise	Other
Agree	**89**	11	0	0	0	0	0	0	0	0
Disagree	0	**81**	2	0	7	0	0	4	0	6
Happy	6	0	**83**	2	0	2	0	0	4	4
Sad	0	0	0	**85**	0	4	6	4	0	2
Clueless	0	4	2	2	**72**	6	4	0	2	9
Thinking	0	0	0	0	6	**89**	6	0	0	0
Confused	0	7	0	0	13	7	**56**	6	2	9
Disgust	0	2	2	4	0	4	9	**76**	0	4
Surprise	0	0	11	0	0	0	4	0	**80**	6

Table 3: Confusion Matrix for images that were 128 by 96 pixels.

disagreement is not merely a response bias, but that there is some information in the smaller image sequences that lead the participants to suspect that the expression was disagreement. Expressions of disagreement consists primarily of large (horizontal) head rotations, which should be relatively well preserved at small image sizes. In this respect, it is interesting to note that cluelessness and confusion also often contain large horizontal head rotations. Unlike disagreement, however, cluelessness and confusion rely heavily on internal face motion to convey their meaning[17]. Thus, if the participants were using horizontal motion energy to detect expressions of disagreement, then cluelessness and confusion should be mistaken for disagreement at small image sizes. This is exactly what is seen in Tables 5 and 6. The high recognition accuracies for disagreement seen the smallest of images, then, may not represent a response bias, but instead the strategic use of rigid head motion information to identify expressions. This suggests an effective method for presenting at least some expressions at extremely small image sizes.

Conclusions

People can recognize the conversational expressions of strangers even in the absence of conversational context – The motion in an image sequence is sufficient to determine what is being communicated. The results here suggest that an image size of about 64 by 48 pixels is the smallest image size where conversational expressions can be accurately recognized. Moreover, increasing the image size above 64 by 48 pixels does not seem to provide any advantage for the **recognition** of expressions, at least for the task used here (a 10 alternative, non-forced choice task). Spontaneous reports from the participants makes it clear that although they could identify the expressions just as easily at an image size of 64 by 48 pixels as at 512 by 384 pixels, they found it subjectively easier to recognize the expressions in the larger images. This suggests that the participants required more attentional and cognitive resources in order to correctly recognize expressions as the images size was reduced. Since expression recognition was the only task the participants were supposed to perform during the experiment, they were free to devote as much at-

	Agree	Disagree	Happy	Sad	Clueless	Thinking	Confused	Disgust	Surprise	Other
Agree	**91**	4	2	0	2	2	0	0	0	0
Disagree	0	**80**	2	0	9	0	4	6	0	0
Happy	4	0	**81**	2	0	4	0	0	6	4
Sad	0	0	4	**81**	6	7	2	0	0	0
Clueless	0	7	0	2	**69**	9	4	0	2	7
Thinking	0	0	4	2	4	**87**	4	0	0	0
Confused	0	6	0	4	19	6	**54**	6	4	4
Disgust	0	0	4	0	4	6	6	**76**	0	6
Surprise	4	0	7	0	0	0	2	0	**85**	2

Table 4: Confusion Matrix for images that were 64 by 48 pixels.

	Agree	Disagree	Happy	Sad	Clueless	Thinking	Confused	Disgust	Surprise	Other
Agree	**78**	2	7	2	4	0	2	0	2	4
Disagree	0	**91**	0	0	6	0	2	2	0	0
Happy	4	2	**72**	2	2	11	0	0	2	6
Sad	4	0	2	**48**	4	24	9	0	0	9
Clueless	4	9	6	7	**48**	6	6	2	0	13
Thinking	0	2	2	9	2	**65**	2	0	2	17
Confused	4	15	2	6	11	6	**30**	4	6	19
Disgust	7	4	7	2	7	13	7	**35**	2	15
Surprise	11	2	19	0	4	0	4	0	**54**	7

Table 5: Confusion Matrix for images that were 32 by 24 pixels.

	Agree	Disagree	Happy	Sad	Clueless	Thinking	Confused	Disgust	Surprise	Other
Agree	**28**	7	0	4	6	4	4	0	2	46
Disagree	2	**74**	2	2	2	2	0	2	0	15
Happy	6	6	**4**	2	6	2	2	4	6	65
Sad	7	2	0	**7**	2	11	0	2	2	67
Clueless	6	22	0	4	**7**	2	2	0	0	57
Thinking	7	2	0	13	2	**11**	4	0	0	61
Confused	9	11	0	4	2	6	**2**	2	6	59
Disgust	7	7	2	4	9	4	2	**4**	4	57
Surprise	6	6	2	4	2	7	2	2	**19**	52

Table 6: Confusion Matrix for images that were 16 by 12 pixels.

tention to the recognition of the expressions as they felt necessary. In applied settings, however, users are generally doing more than simply staring at the avatar. Future work is needed to examine the role of attention in the recognition of conversational expressions. Nonetheless, it is clear that in some situations, very small images are sufficient for the accurate recognition of conversational expressions.

Acknowledgements

This research was supported by the IST project COMIC (COnversational Multi-modal Interaction with Computers), IST-2002-32311. For more information about COMIC, please visit the web page (www.hcrc.ed.ac.uk/comic/).

References

[1] R. E. Bull and G. Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behaviour*, 9:169 – 187, 1986.

[2] J. B. Bavelas and N. Chovil. Visible acts of meaning - an integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology*, 19:163 – 194, 2000.

[3] W. S. Condon and W. D. Ogston. Sound film analysis of normal and pathological behaviour patterns. *Journal of Nervous and Mental Disease*, 143:338 – 347, 1966.

[4] M. T. Motley. Facial affect and verbal context in conversation - facial expression as interjection. *Human Communication Research*, 20:3 – 40, 1993.

[5] D. DeCarlo, C. Revilla, and M. Stone. Making discourse visible: Coding and animating conversational facial displays. In *Proceedings of the Computer Animation 2002*, pages 11 – 16, 2002.

[6] D. W. Cunningham, M. Breidt, M. Kleiner, C. Wallraven, and H. H. Bülthoff. How believable are real faces?: Towards a perceptual basis for conversational animation. In *Computer Animation and Social Agents 2003*, pages 23 – 29, 2003.

[7] D.W. Cunningham, M. Breidt, M. Kleiner, C. Wallraven, and H.H. Bülthoff. The inaccuracy and insincer-

ity of real faces. In *Proceedings of Visualization, Imaging, and Image Processing 2003*, 2003.

[8] P. Ekman. Universal and cultural differences in facial expressions of emotion. In J. R. Cole, editor, *Nebraska Symposium on Motoivation 1971*, pages 207 – 283. University of Nebraska Press, Lincoln, NE, 1972.

[9] J. B. Bavelas, A. Black, C. R. Lemery, and J. Mullett. I show how you feel - motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, 59:322 – 329, 1986.

[10] P. Bull. State of the art: Nonverbal communication. *The Psychologist*, 14:644 – 647, 2001.

[11] J. Cassell and K. R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13:519 – 538, 1999.

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

[13] I. Poggi and C. Pelachaud. Perfomative facial expressions in animated faces. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 115 – 188. MIT Press, Cambridge, MA, 2000.

[14] J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79:941 – 952, 2000.

[15] V. H. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pages 567 – 578. Chicago Linguistic Society, Chicago, 1970.

[16] Mario Kleiner, Christian Wallraven, and Heinrich H. Bülthoff. The MPI VideoLab. Technical report, Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany, 2004.

[17] D.W. Cunningham, M. Kleiner, C. Wallraven, and H.H. Bülthoff. The components of conversational facial expressions. Under Review.