How Believable Are Real Faces? Towards a Perceptual Basis for Conversational Animation

Douglas W. Cunningham, Martin Breidt, Mario Kleiner, Christian Wallraven, Heinrich H. Bülthoff
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany
{first.lastname}@tuebingen.mpg.de

Abstract

Regardless of whether the humans involved are virtual or real, well-developed conversational skills are a necessity. The synthesis of interface agents that are not only understandable but also believable can be greatly aided by knowledge of which facial motions are perceptually necessary and sufficient for clear and believable conversational facial expressions. Here, we recorded several core conversational expressions (agreement, disagreement, happiness, sadness, thinking, and confusion) from several individuals, and then psychophysically determined the perceptual ambiguity and believability of the expressions. The results show that people can identify these expressions quite well, although there are some systematic patterns of confusion. People were also very confident of their identifications and found the expressions to be rather believable. The specific pattern of confusions and confidence ratings have strong implications for conversational animation. Finally, the present results provide the information necessary to begin a more fine-grained analysis of the core components of these expressions.

Keywords: perceptual models, social and conversational agents, virtual humans and avatars, behavioral animation, vision techniques in animation

1 Introduction

An astonishing variety of face and hand motions occur during the course of a conversation. Many of these non-verbal behaviors are central to either the flow or the meaning of the conversation. For example, speech is often accompanied by a variety of facial motions which modify the meaning of what is said, e.g. [1, 2, 3, 4, 5]. Likewise, when producing certain forms of vocal emphasis (e.g., like one would for the word "tall" in the sentence: "No, I meant the **tall** duck") the face moves to reflect this emphasis. Indeed, it can be exceedingly difficult to produce the proper vocal stress patterns without producing the accompanying facial motion. These facial motions are often so tightly integrated with the spoken message that it has been argued that the visual and auditory signals should be treated as a unified whole within linguistics and not as separate entities [2]. Indeed, in many instances it is not clear if the visual and the auditory acts can be separated without significantly altering the intended meaning. For example, a spoken statement of extreme surprise has a rather different meaning when accompanied by a neutral facial expression than when accompanied by a surprised expression.

Non-verbal behaviors can also be used to control the course of a conversation [6, 7, 8, 9, 10]. Cassell and colleagues [8, 9], for example, have created agents that utilize head motion and eye gaze to help control the flow of the conversation (i.e., to help control turn-taking). Even more subtle control of the conversation is possible through the use of *back-channel* responses, e.g. [11, 12]. For example, when a listener nods in agreement, the speaker knows that they were understood and can continue. A look of confusion, on the other hand, signals that the speaker should probably stop and explain the last point in more detail. A look of disgust would signal that a change of topics might be warranted, and so on.

Knowing which expressions people use during a conversation is not, however, sufficient for fully accurate conversational facial animation. There are many ways to incorrectly produce an expression, and attempts to synthesize an expression without knowing exactly what components are perceptually required will lead to miscommunication. Of course, one could try to circumvent this problem by perfectly duplicating real facial expressions. This will not, however, fully alleviate the problem since facial expressions are not always unambiguous. In other words, even within the proper conversational context, we are not always 100% accurate in determining what the expressions of a conversational partner are supposed to mean. Moreover, in some instances a more abstract, cartoon-like, or even non-human embodied agent may be preferred [13], in which case full duplication of all the physical characteristics of human expressions may not be possible. In both cases, knowledge of how humans perceive conversational expressions would be very helpful. What facial motions or features distinguish

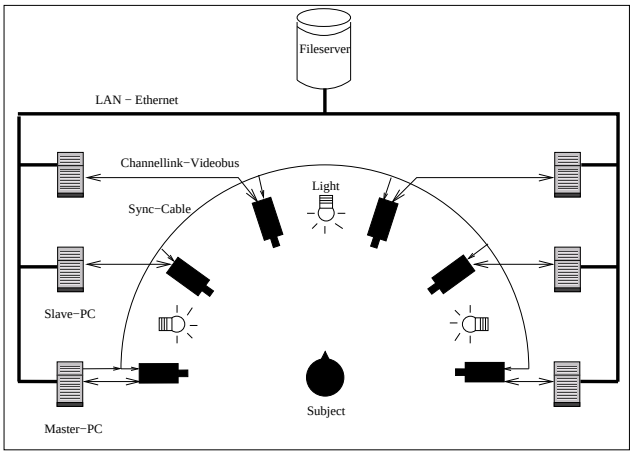

Figure 1: Sketch of the 6 camera layout.

one expression from another? What makes a given instance of an expression easier to identify than another instance of the same expression?

While a fair amount is known about the production and perception of the "universal expressions" (these are happiness, sadness, fear, anger, disgust, contempt, and surprise according to Ekman [14]), considerably less is known about the non-affective expressions which arise during a conversation. Several points regarding conversational expressions are, however, already clear. First, humans often produce expressions during the course of normal conversations that are misunderstood, leading to miscommunication. Second, it is possible to produce an expression that is correctly recognized, but is perceived as contrived or insincere. In other words, realism is not the same thing as clarity and believability. As interface agents become more capable and progress into more business critical operations (e.g., virtual sales agents), the believability of the agent will most likely become a very critical issue. Who would buy anything from an agent if it is obviously lying or insincere, regardless of how good or realistic it looks?

In order to maximize the clarity, believability, and efficiency of conversational agents, it would be strongly adventitious to know what the necessary and sufficient components of various facial expressions are. Here, we lay the groundwork for such a detailed examination by first determining how distinguishable and believable six core conversational facial expressions are (agreement, disagreement, happiness, sadness, thinking, and confusion). Using a recording setup consisting of 6 synchronized digital cameras (described in Section 2), we recorded 6 individuals performing these conversational expressions (described in Section 3). The results (described in more detail in Section 4) show that people can identify these expressions quite well, although there are some systematic patterns of confusion. The specific patterns of confusion indicate several potential problem areas for conversational animation. The results also show that people were very confident of their identifications, even when they misidentified an expression. Finally, the expressions were not all fully convincing; the believability of expressions varied considerably across individuals. Having identified expressions which differ considerably in believability and ambiguity, and having recorded the expressions from multiple viewpoints with tracking markers placed on the faces, additional experiments will be able to provide a more detailed description of exactly which types of facial motion led to the confusions, and which types of information led to the clearest comprehension of the intended message (see Section 5).

2 Recording Equipment

To record the facial expressions, a custom camera rig was built using a distributed recording system with six recording units each of which consisting of a digital video camera, a frame grabber and a host computer.

Each unit can record up to 60 frames/sec of *fully synchronized* non-interlaced, uncompressed video in PAL resolution (768 x 576), which is stored in raw CCD format.

The six cameras were arranged in a semi-circle around the subject (see Figure 1) at a distance of approximately 1.5m. The individuals were filmed with 30 frames/s and an exposure time of 3 ms in order to reduce motion blur.

To facilitate later processing of the images, care was taken to light the actors' faces as flat as possible to avoid directional lighting effects (cast shadows, highlights). For a more detailed description of the recording setup, see [15].

3 Methodology

Six expressions were recorded from six different people (three males, three females), yielding 36 video sequences. The six expressions were agreement, disagreement, thinking, confusion, pleased, and sadness [1] (see Figure 2). These particular expressions can play an important role in the structure and flow of a conversation, and knowledge of how humans produce and perceive believable versions of these expressions will be used in the construction of a conversational agent as part of the IST project COMIC (COnversational Multimodal Interaction with Computers).

For later processing of the recordings (e.g., stereo-reconstruction), black tracking markers were applied to the faces using a specific layout. After marker application, each

[1] Both the recordings and the experiment were conducted in German. The exact labels used for six expressions were Zustimmung, Ablehnung, Glücklich / Zufrieden, Traurigkeit, Nachdenken, and Verwirrung.

(a) (b) (c)

(d) (e) (f)

Figure 2: The six expressions. (a) Agreement; (b) Disagreement; (c) Happiness; (d) Sadness; (e) Thinking; (f) Confusion. It is worth noting that the expressions are more difficult to identify in the static than in the dynamic versions.

(a) (b)

(c) (d)

Figure 3: The many faces of thought. Despite the variability in facial contortion, all of these expressions are recognizable as thoughtful expressions.

actor was centered in front of all six cameras, and he or she was asked to imagine a situation in which the requested expression would occur. The "actor" was then filmed with a neutral face first, followed by the transition to the requested expression and the reversal to neutral face again. The actors were completely unconstrained in the amount and type of movements that they could produce, with the single exception that they were asked to not talk during their reaction unless they felt they absolutely had to. This procedure was repeated at least three times for each emotion. The best of each repetition for each expression from each person was selected and edited so that the video sequence started at the beginning of the expression and ended just as the expression started to shift back towards neutral. The length of the sequences varied considerably. The shortest was 27 frames long (approximately 0.9 seconds), and the longest was to 171 frames (approximately 5.7 seconds). No straightforward correlation between type of expression and length of expression was apparent.

Each of the resulting 36 video sequences was shown to 10 different people (hereafter referred to as *participants*) in a psychophysical experiment. The primary goal of psychophysics is to systematically examine the functional relationship between physical dimensions (e.g., light intensity), and psychological dimensions (e.g., brightness perception). The work presented here examines the functional relationship between rather high-level dimensions (i.e., patterns of facial motion and the perception of expressions), and thus might more be precisely referred to as mid- or high-level psychophysics.

The sequences were presented at 30 frames/s on a computer. The images subtended a visual angle of about 10 by 7.5 degrees. The order in which the 36 expressions were presented was completely randomized for each participant.

Participants were given three tasks. While viewing an expression (which was repeated until the participant responded, with 200 ms blank between repetitions), the participant was first supposed to identify it using a multiple choice procedure. More specifically, participants identified an expression by selecting the name of one of the 6 expressions from a list or by selecting "none-of-the-above" to indicate that the expression was not on the list. Previous research has shown that performance on this type of task (7 alternative, no forced-choice task) is highly correlated with other identification tasks (e.g., free naming – where participants choose any word they want to describe an expression or the emotion behind an expression, connecting an expression with a short story, etc.), at least for the "universal expressions". See [16] for more information. While these tasks are very well suited for elucidating the relationship between an expression (and, to some degree, the intention behind the expression) and the perception of that expression (i.e., which facial motions are correlated with clear and believable facial expressions) when the expression is shown in isolation (i.e., not enclosed within a series of expressions), it is less clear that these tasks can be used to determine the central theme or message in a complex sentence or scene. For this type of question, an approach similar to that used by Emiel Krahmer and colleagues (where participants determined in which of two sentences a given word was more prominent) may be more appropriate; see [17].

Of course, in order to determine what role a single expression plays within a concatenation of expressions, one must first determine what information that expression in and of itself carries. The same is true for examining the role of context and multimodal expression of meaning: in order to understand the interactive effects of context and multiple channels with facial expressions, one must first be certain that the facial expression used is a clear, and believable exemplar of the intended message. To that end, this present task should help to determine which facial motions are most closely correlated with a given intended message, allowing one to then investiget multimodal interactions.

Immediately after identifying an expression, the participants were asked to indicate on a 5 point scale exactly how confident they were about their response. They were told that a rating of 1 indicates that they are completely unconfident (i.e., merely guessing) and 5 means they were completely confident. Finally, the participants were asked to rate from 1 (completely fake) to 5 (extremely convincing) how believable or realistic the expressions was.

4 Results and Discussion

Overall, participants were quite successful at identifying the expressions. Table 1 shows a confusion matrix of the participants' responses [2]. The pattern of responses for the "thinking" expressions is particularly interesting: Participants thought this expression was actually "confusion" 20% of the time. In many respects this is not too surprising, as people will often stop and think when they are confused. As such, thinking and confusion are naturally somewhat intertwined. Regardless, such a mistake in a conversation with an interface agent could well lead to miscommunication and other problems, as well as decrease the overall efficiency of the system [18]. For example, a thinking expression might be used as an activity index. If a user were to mistake the agent's thinking expression (indicating that the agent is busy) for one of confusion (indicating that the agent is waiting for more information), the user might well attempt to clear up the perceived confusion - and speak to an already busy system.

[2] For each expression, the responses of all ten participants were collapsed across each of the six "actors". The resulting frequency histogram of responses was converted into a percentage.

Actual Expression \ Participants' Responses	Agreement	Disagreement	Happiness	Sadness	Thinking	Confusion	Other
Agreement	**95%**	0%	2%	0%	0%	0%	3%
Disagreement	0%	**85%**	2%	7%	0%	3%	3%
Happiness	7%	2%	**73%**	3%	0%	5%	10%
Sadness	0%	3%	0%	**82%**	5%	2%	8%
Thinking	0%	3%	0%	2%	**73%**	20%	2%
Confusion	0%	18%	0%	2%	5%	**73%**	2%

Table 1: Confusion Matrix of the identification responses. The percentage of the time a given response was chosen (columns) is shown for each of the six expressions (rows).

Actual Expression \ Actor	Actor 1	Actor 2	Actor 3	Actor 4	Actor 5	Actor 6
Agreement	100%	100%	100%	80%	100%	90%
Disagreement	90%	80%	100%	100%	90%	50%
Happiness	80%	50%	100%	60%	60%	90%
Sadness	40%	80%	90%	80%	100%	100%
Thinking	60%	90%	80%	70%	60%	80%
Confusion	90%	70%	50%	40%	90%	100%

Table 2: Actor accuracy Matrix. The percentage of the time a given expression was correctly identified is shown for each of the six actors.

Slightly more surprising is the fact that "confusion" was often mistaken for "disagreement". Such a misidentification in the interaction with an interface agent would also most likely decrease efficiency (e.g., the user might chose to defend his or her position, rather than clarify it as the system expects).

In addition to reinforcing previous warnings about using ambiguous facial expressions [18, 13], the pattern of confusions clearly demonstrates that even the perfect duplication of real expressions would not produce an unambiguous interface agent. Duplication of a confusing template will only lead to additional confusion.

A simpler explanation for the pattern of confusions in Table 1 would be to claim that the since the "actors" were not trained, they might not be producing the right expressions. While this is a research topic in and of itself, it should be kept in mind that humans often "pretend", producing an expression that is appropriate to the given context regardless of whether they really feel the proper emotion. Regardless, Table 2, which depicts the success of the different actors at producing correctly identifiable expressions, begins to disentangle "bad acting" from real confusions. The first thing that becomes apparent from a glance at this table is the wide degree of variation in identification scores, both within and across expressions. Sadness is a good example of the latter: Some actors were only correctly identified 40% of the time, while others were correctly recognized 100% of the time. Clearly some individuals were producing the wrong (or at least ambiguous) expression, but one cannot say that every actor was producing the wrong expression. Variation across expressions is well exemplified by the "Thinking" and "Agreement" expressions. All of the "Agreement" expressions were identified 80% or more of the time, whereas only one actor produced a "Thinking" expression that was recognized more than 80% of the time. It seems, then, that "Thinking" can be produced in a recognizable fashion, but often is not. The interesting question here is what differs in the image sequences that allows one to be well identified but the others not. Having identified instances of expressions that vary in terms of accuracy, future research can begin to provide a more detailed description of what image differences cause the perceptual differences.

While the addition of a conversational context, and the concomitant expectations, would no doubt improve the ability of participants to identify these expressions, Table 2 clearly shows that all of the expressions are potentially unambiguous even without a context: Each expression was recognized 100% of the time from at least one actor, with the exception of "thinking" which was recognized 90% of the time at best. Furthermore, having found the degree to which these expressions can, by themselves, convey a given meaning, one can begin to systematically examine exactly how context can modulate that signal.

Participants were generally quite confident in their deci-

Actual \ Response	Correct	False
Agreement	4.67	4.0
Disagreement	4.51	3.43
Happiness	4.44	4.23
Sadness	4.08	4.19
Thinking	4.30	3.96
Confusion	4.14	3.75

Table 3: Confidence ratings. The average confidence of the participants in their responses is listed as a function of whether they correctly identified the expression or misidentified it. Confidence was rated on a 5 point scale from completely unconfident (a value of 1) to completely confident (a value of 5).

Actual \ Response	Correct	False
Agreement	3.81	3.25
Disagreement	3.96	2.79
Happiness	3.54	3.28
Sadness	3.26	3.69
Thinking	3.91	3.33
Confusion	3.67	3.73

Table 4: Believability ratings. The average believability ratings are listed as a function of whether the expression was correctly or incorrectly identified it. Believability was judged on a 5 point scale from completely unbelievable (1) to completely convincing (5).

sions (see Table 3). Although they were clearly less confident when they made a mistake, they were still relatively certain that they had correctly identified the expression. That is, even when they made a mistake, people were relatively certain that they were **not** making a mistake. This confidence in one's mistakes can have strong implications for the design of conversational flow in general, and for the design of an interface agent's confusion handling and persuasion routines in specific (see, e.g., [19, 20, 21]).

In general, the expressions were considered rather believable (see Table 4), but were not considered completely convincing. The participants found the expressions to be less believable when they had incorrectly identified it. That is, if an expression was really one of "thinking", but a participant thought it was "confusion", they would be relatively certain that the expression was "confusion" (i.e., confidence ratings), but would find the expression to be somewhat unconvincing or contrived.

5 Conclusion and Outlook

In general, these six conversational expressions are easy to identify, even in the complete absence of conversational context. There were, however, some noteworthy patterns of confusion, most notably that "thinking" and "confusion" were often misinterpreted. People were also generally confident that they had correctly identified an expression, even if they were, in fact, wrong. The specific pattern of confusions and this confidence in the face of errors not only have implications for the animation of conversational expressions, but also for the design of an interface agent's conversational flow capabilities.

The assymetry of the identification confusions (Table 1) hints at a potential assymetry in the underlying perceptual space of facial expressions. This is important to know when analyzing and synthesizing expressions, particularly when dealing with variability within this expression space. Variability can arise for several reasons, including the presence of sub-categories of expressions. For example, when thinking, one can be pensive, contemplative, calculating, etc (see, e.g., Figure 3). All of these are recognizable as "thinking", but the distribution of them in expression space may not be symmetrical, so one needs to be careful when traversing this subregion. A second source of variability is the fact that humans are largely incapable of *exactly* duplicating a behavior. While the resulting minor variations in the motions accompanying an expression may or may not carry any communicative meaning, their absence may well be important (as the mechanically perfect repetition would almost certainly be recognized as unnatural).

It is, of course, possible that some of the confusion arose from the fact that the expressions were intentionally generated (i.e., were posed). There is considerable evidence, however, that during normal conversation humans not only intentionally generate various facial expressions, but do so in synchrony with the auditory portion of a conversation [2]. That is, normal conversational expressions may be, at least in part, just as "posed" as the specific words and phrases used in a conversation. Moreover, people generally found the present expressions to be believable, even when they misidentified the expressions. That is, even when people were confused about an expression, they still found the expression to be rather believable and not contrived. This is a rather critical point, for two reasons. First, it serves to reinforce the idea that real does not equate with believable. The creation of a computerized 3D model of a head that is a perfect physical duplicate of a real human head does not automatically mean that the expressions generated with such a head will be unambiguous or convincing. Second, and perhaps more important, the simulation of proper conversational behaviors must include socially determined behaviors and expressions as well as any "truly genuine" ex-

pression of emotion. In other words, regardless of the underlying reason for why some individuals produced clearer and more believable expressions than others, it is perhaps more interesting to ask what portions of the image sequence lead people to be confused, and which components enhance proper recognition. Likewise, a detailed knowledge of the components that enhance the perceived believability of the expression would be of great help. To that end, the fact that the present expressions were recorded from multiple viewpoints with tracking markers placed on the faces allows us to begin to manipulate the video sequences, and additional experiments will help elucidate which components are necessary and sufficient for unambiguous, believable conversational expressions. Such a knowledge would allow us to answer in an informed fashion what we need to animate for the end result not only to be well understood, but also believed.

6 Acknowledgments

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 (http://www.hcrc.ed.ac.uk/comic/). We would like to thank Dorothee Neukirchen for help in recording the sequences, and Jan Peter de Ruiter and Adrian Schwaninger for fruitful discussions.

References

[1] R. E. Bull and G. Connelly, "Body movement and emphasis in speech," *Journal of Nonverbal Behaviour*, vol. 9, pp. 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*, vol. 19, pp. 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*, vol. 143, pp. 338 – 347, 1966.

[4] M. T. Motley, "Facial affect and verbal context in conversation - facial expression as interjection," *Human Communication Research*, vol. 20, pp. 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*, 2002, pp. 11 – 16.

[6] 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*, vol. 59, pp. 322 – 329, 1986.

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

[8] 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*, vol. 13, pp. 519 – 538, 1999.

[9] 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*, vol. 14, pp. 22 – 64, 2001.

[10] I. Poggi and C. Pelachaud, "Perfomative facial expressions in animated faces," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. Cambridge, MA: MIT Press, 2000, pp. 115 – 188.

[11] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as conarrators," *Journal of Personality and Social Psychology*, vol. 79, pp. 941 – 952, 2000.

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

[13] M. Wilson, "Metaphor to personality: The role of animation in intelligent interface agents," in *Proceedings of the IJCAI-97 Workshop on Animated Interface Agents: Making them Intelligent*, Nagoya, Japan, 1997.

[14] P. Ekman, "Universal and cultural differences in facial expressions of emotion," in *Nebraska Symposium on Motoivation 1971*, J. R. Cole, Ed. Lincoln, NE: University of Nebraska Press, 1972, pp. 207 – 283.

[15] M. Kleiner, C. Wallraven, and H. H. Bülthoff, "The MPI VideoLab," Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany, Tech. Rep. 104, 2003.

[16] M. G. Frank and J. Stennett, "The forced-choice paradigm and the perception of facial expressions of emotion." *Journal of Personality and Social Psychology*, vol. 80, pp. 75 – 85, 2001.

[17] E. Krahmer, Z. Ruttkay, M. Swerts, and W. Wesselink, "Audiovisual cues to prominence," in *Proceedings ICSLP*, 2002, pp. 1933 – 1936.

[18] D. M. Dehn and S. van Mulken, "The impact of animated interface agents: a review of empirical research," *International Journal of Human-Computer Studies*, vol. 52, pp. 1 – 22, 2000.

[19] J. Jaccard, "Toward theories of persuasion and belief change," *Journal of Personality and Social Psychology*, vol. 40, pp. 260 – 269, 1981.

[20] J. J. Jiang, G. Klein, and R. G. Vedder, "Persuasive expert systems: the influence of confidence and discrepancy," *Computers in Human Behavior*, vol. 16, pp. 99 – 109, 2000.

[21] R. E. Petty, P. Brinol, and Z. L. Tormala, "Thought confidence as a determinant of persuasion: The self-validation hypothesis," *Journal of Personality and Social Psychology*, vol. 85, pp. 722 – 741, 2002.