The inaccuracy and insincerity of real faces

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

ABSTRACT

An avatar who's behavior is unbelievable or easily misinterpreted will be an inefficient and possibly counterproductive conversational partner. Here, we psychophysically determined how recognizable and believable several real expressions were. In general, there is systematic confusion between particular expressions. Critically, even these real facial expressions were not always understood or believed. The results also provide the ground work necessary for a fine-grained analysis of the core components of these expressions. Some initial results from a model-based manipulation of the image sequences shows that such a detailed analysis can be an invaluable aid in the synthesis of unambiguous and believable Avatars.

KEY WORDS
Facial Animation, Human/Computer Interaction, Perception, Psychophysics

1 Introduction

Facial motions, which can play an important role in conversations, come in an almost bewildering variety. While lip motion may be one of the most recognized types of conversational facial motion, it is by no means the only type. Indeed, many of the various contortions that a face undergoes during a conversation have little or nothing to do with the production of words. For example, it is well documented that facial motions serve to modify the meaning of what is being said [1, 2, 3, 4, 5]. A statement of appreciation takes on quite a different meaning when accompanied by a look of displeasure. Likewise, when vocally emphasizing a word in a sentence, the face moves to reflect this emphasis. This intimate connection between spoken meaning and facial motion has prompted some to suggest that the two signals together form the basic unit of meaning in speech, rather than providing independent contributions [2].

Facial motion can also help to control the flow of a conversation [6, 7, 8, 9, 10]. In the simplest version of this, head and eye gaze direction can be used to indicate to which conversational partner a request is directed. Improper or absent eye gaze information is an oft cited problem with most video-conferencing technology [11, 12]. Facial control of conversational flow can also be quite subtle, using such techniques as "back-channel" responses [13, 14]. For example, a nod from a listener will encourage a speaker to continue, while a look of confusion will probably prompt the speaker to stop and clear up the confusion.

The differences between an identifiable expression and an unrecognizable expression can be quite subtle. Even if a physically accurate Virtual Human perfectly duplicates all spatial and temporal aspects of facial motion, and is driven in real-time from a real human face, there is still no guarantee that the resulting expressions will be understood. Moreover, even if an expression is synthesized well enough that it is easy to identify, it may still be considered insincere. While sincerity is not currently a big concern, believability will undoubtedly become a central issue for interface agents. Who would buy anything from a virtual salesperson if the sales agent seems dishonest and insincere?

A systematic description of how real faces move in real expressions would be of great help. Experimentally validating the correlation between individual facial motions and the ambiguity and believability of the intended expression would greatly advance the state of the art. Before one can do this, however, one must first find and rate real expressions for their distinguishability and believability. Here, we lay this critical groundwork for eight core conversational facial expressions.

2 Recording Equipment

The expressions were recorded using a custom designed, distributed recording setup [16]. The system is made up of six recording units, each of which consists of a digital video camera, a frame grabber and a computer. Each unit can record up to 60 frames/sec of *fully synchronized* non-interlaced, uncompressed video in PAL resolution (768 x 576 pixels). For the present recordings, the six cameras were arranged in a semi-circle around the subject at a distance of approximately 1.5 m. The individuals were filmed at 25 frames/s, with an exposure time of 3 ms. To help avoid artifacts and unintended information in the recorded sequences, care was taken to light the actors' faces as flatly as possible. Special effort was devoted to the avoidance of directional lighting effects (cast shadows, highlights).

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 eight expressions (rows). The diagonal (bold) shows the percent correct.

		Participants' Responses								
		Disgust	Agree	Disagree	Happy	Clueless	Thinking	Confusion	Surprise	Other
Actual Expression	Disgust	**65%**	0%	2%	0%	4%	0%	7%	0%	22%
	Agree	4%	**94%**	0%	0%	0%	0%	0%	0%	2%
	Disagree	5%	7%	**67%**	0%	13%	0%	6%	0%	2%
	Happy	0%	0%	2%	**70%**	0%	0%	0%	15%	13%
	Clueless	0%	0%	2%	0%	**78%**	0%	11%	0%	9%
	Thinking	2%	0%	6%	0%	7%	**67%**	9%	2%	7%
	Confused	8%	0%	7%	0%	13%	0%	**59%**	4%	9%
	Surprise	2%	0%	0%	4%	2%	0%	0%	**85%**	7%

3 Recording Methodology

Eight expressions were recorded from six different people. Five of the individuals were amateur actors, one was a professional actor. The expressions were elicited using a protocol based on method acting. Specifically, a situation was described in detail to the actors. They were asked to imagine that they were in that situation and to react appropriately. For each reaction, the actors were told that they could use any motion they felt necessary, but were asked to try to refrain from talking and moving their hands in front of their faces unless they felt that they had to. Each expression was recorded at least three times, with the actor relaxing into a neutral expression before and after each repetition. One of the three repetitions was chosen and then edited so that the video sequence began on the frame after the face began to move away from the neutral expression and ended after reaching the peak of the expression. The resulting 48 video sequences varied considerably in length; The shortest sequence was 17 frames long (0.68 second) and the longest lasted 194 frames (7.76 seconds). No straight-forward correlation between expression and duration was apparent.

4 Experimental Methodology

The video sequences were shown to nine individuals (hereafter referred to as *participants*) in a psychophysical experiment. The image size was reduced for purposes of the experiment to 256 by 192 pixels (10 by 7.5 degrees of visual angle). The order in which the 48 expressions were presented was completely randomized for each participant, with each video sequence being shown repeatedly until the participant indicated that they were ready to respond. A 200 ms blank screen was inserted between repetitions of the video clip. When participants were ready to respond, the video sequence was removed from the computer screen, and the participants were asked to perform three tasks.

The first task was to identify the expression by selecting the name of an expression from a list that was displayed on the side of the screen. The participant could choose one of the eight expressions, or "none of the above" to indicate that the expression was not on the list. Since not all of the participants were native German speakers, the experiment was conducted in both German and English. In English, the first six expressions were: agreement, disagreement, disgust, pleased / happy, thinking, and pleasantly surprised. The remaining two consisted of a label and a brief description. One was referred to as clueless, and was accompanied by the statement that the expression represented the case were the actor "does not know" the answer. The final expression was referred to as confusion, and was to represent the situation where the actor "does not understand" what was just said. In German, the labels were: zustimmen, nicht zustimmen, angewidert, glücklich / zufrieden, nachdenklich, angenehm überrascht, unwissend ("weiss nicht"), and verwirrt ("versteht nicht").

Previous research in other labs has shown that responses on this procedure (a nine alternative, non-forced-choice task) is highly correlated with other identification procedures (e.g., free description of the expressions), at least for the "universal expressions" (according to Paul Ekman, these are happiness, sadness, anger, fear, surprise, disgust and possibly contempt [17]). The present methodology has several significant advantage over other techniques. First, the inclusion of a "none of the above" option helps to avoid some of artifacts (including the inflation of recognition rates). Second, the categorization of responses is more objective than free-response methods. See Frank and Stennett [18] for more information on this type of task.

The second task was to indicate how intense the expression was. This rating took the form of a 5 point scale, with a rating of 1 indicating a weak expression, and 5 indicating a very strong or intense expression.

Finally, the third task was to indicate how believable the expression was using a similar 5 point scale. The participants were to indicate if the actor was merely pretending (a rating of 1) or looked like they really meant the underlying expression (a rating of 5).

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

Expression \ Actor	Actor 1	Actor 2	Actor 3	Actor 4	Actor 5	Actor 6
Disgust	0%	78%	89%	89%	89%	44%
Agree	89%	89%	100%	100%	100%	89%
Disagree	78%	67%	56%	67%	56%	79%
Happy	89%	67%	56%	67%	78%	67%
Clueless	78%	78%	78%	44%	100%	89%
Thinking	89%	33%	56%	67%	67%	89%
Confusion	67%	33%	22%	56%	89%	89%
Surprise	67%	78%	100%	100%	78%	89%

5 Results and Discussion

In general, the participants were able to successfully identify the expressions. Overall, the pattern of confusions is similar to the pattern that Cunningham et al. [15] found using different actors and a smaller subset of expressions. This stability across studies strongly suggests that the confusions reflect inherent characteristics of either the production or the perception of these expressions, or both.

5.1 Confusions

Specific insights into the perception of these expressions can be gained by examining the confusions they generated. Some of the confusions are not entirely surprising. For example, 15% of the time, pleased expressions were considered to be pleasantly surprised expressions. As the labeling of the expressions indicates, the latter expression might be considered to be the former expression combined with surprise. Consistent with this explanation, the reverse type of confusion rarely seems to occur.

There is a complementary pattern of errors for the clueless expression and the confusion expression. Since the difference between "I don't know" and "I don't understand" is somewhat subtle, this confusion is to be expected. Indeed, it is perhaps more surprising that subjects did so well at separating these expressions. While the underlying messages are undeniably similar, they do indeed seem to represent separate ideas. Further studies should show exactly which facial motions differentiate the two expressions. Such knowledge could be an invaluable asset in synthesizing clear versions of these expressions.

Thinking is often (17% of the time) mistaken to be either cluelessness or confusion. These results are strikingly similar to Cunningham et al.'s results [15] (22% of the time thinking was mistaken to be confusion in Cunningham et al.'s study). It seems that an expression of thoughtfulness and one of a lack of knowledge are indeed quite related. Despite this natural entanglement, mistaking thinking for either confusion or ignorance could lead to severe difficulties, particularly in the realm of human-machine interfaces [19].

It is potentially more interesting that neither the clueless nor the confusion expressions were ever mistaken for thinking. This, along with the pattern of confusions between pleased and pleasantly surprised, points to a rather prominent asymmetry in the underlying facial expression space. One potential source for this asymmetry might lie in the rules for how various expressions are combined to produce compound expressions. The pleasantly surprised expressions is a good example of a compound expression, combining pleased or happiness with surprise. Further study of the components of individual expressions should help to elucidate these rules, and ease the synthetic traversal of expression space.

The pattern of confusions in Table 1 might give the impression that some expressions are simply more ambiguous than others, particularly in the absence of the appropriate conversational context. While there may be some truth to this impression, the story is actually much more complex. Table 2 depicts the success of the different actors at producing identifiable expressions. The first thing that becomes apparent from a glance at this table is the wide degree of variation in identification scores. It seems that actors can be quite skilled at one expression, but rather incompetent at others. For example, Actor 1's disgust expression was never correctly identified, while three of his other expressions were identified correctly by 8 of the 9 participants. While the addition of a conversational context, and the concomitant expectations, should improve the ability of participants to identify these expressions, Table 2 clearly shows that all of the expressions are potentially unambiguous without a context: Each expression had a high identification rate from at least one actor.

In addition to reinforcing previous warnings about using ambiguous facial expressions [19, 20], the pattern of confusions clearly demonstrates that even the perfect duplication of real expressions would not produce an unambiguous interface agent. The results also suggest that the production of a synthetic agent capable of producing unambiguous expressions might be well aided by using the clearest expressions from different actors. Contrariwise, to

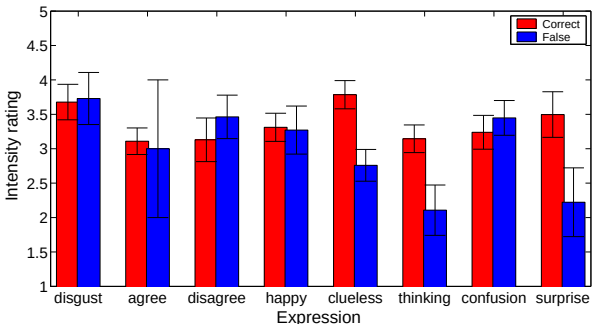


Figure 1. The intensity ratings. The ratings were on a five-point scale, with a value of one representing a weak expression and five representing a strong expression. The red bars depict ratings for expressions that were correctly identified, and the blue bars for incorrectly identified expressions. The error bars represent the standard error of the mean.

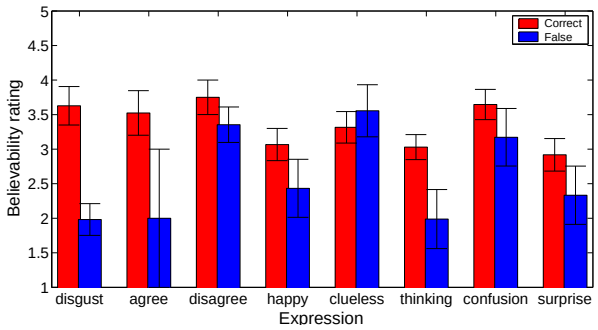


Figure 2. The believability ratings.

produce a highly individualistic agent, one may have to sacrifice clarity.

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 intentionally chosen as the specific words and phrases used in a conversation.

5.2 Perception of Intensity and Believability

In Figure 1, the intensity ratings are shown as a factor of whether the expression was correctly identified or not. For example, when a participant saw an expression of cluelessness and mistakenly labeled it as one of confusion (which happened 11% of the time), he or she tended to rate the expression as being less intense than had he or she correctly identified it. In general, the ratings hover around the middle of the scale (a rating of 3). This may be, in part, due to the lack of a context or to the posed nature of the expressions. It definitely reflects the well-known fact that participants rarely use the extreme values on a scale. Overall, the accuracy of a response does not seem to modulate the perceived intensity of the expressions much, suggesting that the physical type and extent of motion is what determines intensity (i.e., the intensity of an expression may be partially independent of what the expression is supposed to signify).

Finally, the expressions were considered rather believable, but not completely convincing (see Figure 2). In contrast to the intensity ratings, the participants found an expression 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 find the expression to be somewhat unconvincing or contrived. While this pattern is nearly identical to that found by Cunningham et al. [15], the absolute values are surprisingly lower here than in the previous study. One might well expect that the improved elicitation protocol and the usage of trained actors would yield expressions that were more believable and easier to identify. Instead, it seems that the untrained individuals in the previous study were able to produce more believable (and possibly clearer) expressions. Future research directly comparing recordings taken from trained and untrained individuals using identical recording situations are needed to clarify this issue. At the vary least, however, the results give doubt to the aphorism that actors will produce better expressions than untrained individuals.

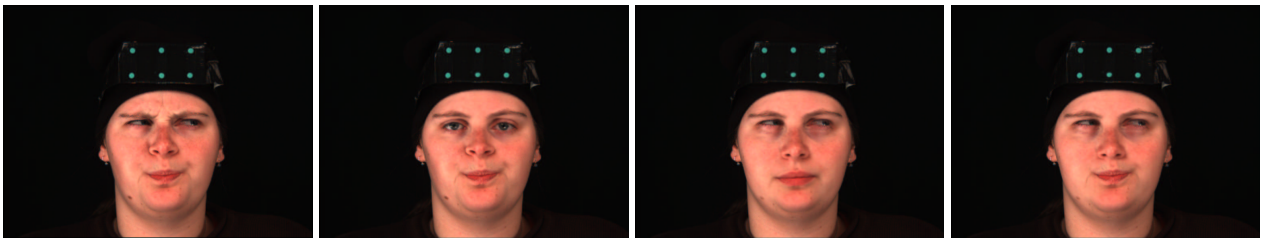
6 Components of Motion

As can be seen in Table 2, some instances of an expression were clearer than others. Which facial motions lead to clearer expressions? Likewise, some expressions were more believable than others. Why? Now that at least the beginning of a corpus of expressions has been collected and those expressions have been rated for their clarity and believability, a more detailed analysis of the necessary and sufficient facial motions can begin. One method to accomplish this is to take the recordings and manipulate them so that only certain areas of the face move while the rest is held still in a neutral position. In this manner, one can begin to determine which areas of the face need to move and when they need to move in order for different expressions to be understood and believed.

6.1 Image manipulation technique

In order to replace facial motion of parts of the face with static snapshots (therefore ”freezing” parts of the face), the following model-based image manipulation procedure was applied to the recorded video footage.

Before the recordings were taken, a Cyberware 3D laser range scanner was used to acquire a detailed three dimensional model of the shape and texture of the actor’s



(a) (b) (c) (d)

Figure 3. Components of thought. One of actor 3's thinking expressions was manipulated. a) A snapshot from the original recording. b) All of the face except for the mouth region is held in a neutral expression. c) All of the face except for the eyes and eyebrows are held still. d) The eyes, eyebrows, and mouth region were all allowed to move, while the rest of the face was frozen.

head. The resulting model consists of a 3D polygon mesh of approximately 150,000 triangles, defined by 75,972 vertices with a spatial resolution of approximately 0.1 mm. A texture map (512 x 512 texels) accompanies the mesh.

All models of all actors were brought into correspondence with each other by a post-processing procedure, described in detail in Blanz and Vetter[21]. This procedure guarantees that corresponding components of the shape vectors and texture maps of different models always define the same facial region. For example, vertex 22,345 in each head mesh always defines the position of the tip of the nose, while the texel at position (256,238) always represents the color value of the tip of the nose. This makes it possible to define region-based manipulations of head geometry or texture once on a reference model and then apply them automatically to an arbitrary number of models.

During the video recordings of expressions, each of the actors wore a black hat with a tracking target. The tracking target consists of a black rectangular plate with six green markers on it (see, e.g., Figure 3). After the recording, a custom, image-based, three-dimensional motion-tracking algorithm was applied to the video footage. First, the algorithm employs color segmentation to find the 2D image positions of the six green markers in each of the stereo pair images. It then uses the tracked image positions of corresponding markers in both images to recover the 3D spatial position of the markers via stereo triangulation. Finally, the algorithm fits a geometric model of the tracking target to the 3D point cloud of markers, thereby recovering position and orientation of the tracking target in space. As there is a fixed spatial relationship between the rigid motion of the actor's head and the motion of the tracking target (N.B., the relationship is set up by manual interactive initialization on the first video frame of each recorded sequence), the recovered position and orientation of the target is used to position and orient the 3D shape model of the actor's head accordingly, thereby establishing a point-to-point correspondence between texels in the texture map of the model and image pixels in the video footage. This correspondence is used to perform texture extraction on suitable frames of the video sequence e.g., frames where the eyes or mouth are in a neutral position.

To freeze parts of the face, the head model is superimposed onto the video footage by rendering it with standard OpenGL graphics, using alpha-blending to smooth out transitions between the rendered mesh and the video frame. In face regions where we do not want to remove facial movement, the corresponding parts of the model mesh are rendered with an alpha value of zero (fully transparent and therefore invisible). In regions where we want to freeze the face, the model is rendered opaque with one of the previously extracted texture maps applied (e.g., a static texture of the eyes for freezing the eye region). Due to the head model's correspondence properties, we are able to define the facial regions which are to be frozen with a single texture mask and then apply these manipulations to all recordings of all actors, greatly reducing the amount of manual setup work involved in manipulation of a large number of sequences.

6.2 Initial Observations

The conversion of the entire corpus into a variety of manipulated sequences is currently underway. From those recordings that have been converted, some initial results are already apparent. In some cases, it is clear that simple rigid motion of the head is sufficient for an expression. The expression of agreement from one of the actors, for example, contains little facial motion and the immobilization of the entire face does not seem to alter the appearance of the expression. On the opposite end of the spectrum lie expressions which contain a variety of facial motion types.

Figure 3 shows the manipulation of one type of thinking. Figure 3a shows a snapshot from the original recording. Freezing all of the face except for the mouth region in a neutral expression (see 3b) does not produce a recognizable thinking expression (indeed, it looks more like an expression of displeasure). Keeping all of the face except the eyes still (see 3c) resembles thinking somewhat more, but is still not unambiguous. Allowing both the eyes and the mouth region to move (see 3d) does produce an expression recognizable as thinking, but it still lacks something from the original. To some degree, this version looks less intense or less believable, although more systematic exper-

imentation is required to be certain of what is missing. At the very least, the motion of the nose, cheeks and forehead seem to add some depth to the expression.

7 Conclusion

In general, the eight conversational expressions used in this study are identifiable, even in the complete absence of conversational context. There were, however, some noteworthy patterns of confusion, which seem to be consistent across recording style, actor, and observer. This clearly shows that even real expressions are not always clear and believable. Thus, even if a three dimensional model of a head perfectly duplicated the physical structure of human heads as well as all aspects of a real human's facial motion, there is still a good chance that the resulting animations will be misunderstood. Realism is not the same thing as clarity.

There were also significant variations in expressiveness across actors: Some people seem to be better at certain expressions than others. Preliminary work with image manipulation has begun to detail these variations. For some expressions, rigid head motion seems to be sufficient, for others a full complement of head, eye, eyebrow, and mouth motions does not appear to be sufficient. As can already be seen, producing a systematic description of what needs to move when in order to produce clear and believable expressions is not likely to be a simple undertaking. Nonetheless, such a detailed exploration of conversational expressions promises not only to be a great aid in the synthesis of clear and believable expressions, but also will help Virtual Humans to be more expressive, varied, and individualistic.

8 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 (www.hcrc.ed.ac.uk/comic/). We would like to thank Jan Peter de Ruiter for fruitful discussions.

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] 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.

[7] P. Bull. State of the art: Nonverbal communication. *The Psychologist*, 14:644 – 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*, 13:519 – 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*, 14:22 – 64, 2001.

[10] 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.

[11] E. Isaacs and J. Tang. What video can and can't do for collaboration: a case study. In *"ACM Multimedia '93"*, pages 496 – 503. ACM, New York, 1993.

[12] R. Vertegaal. Conversational awareness in multiparty vmc. In *"Extended Abstracts of CHI'97"*, pages 496 – 503. ACM, Atlanta, 1997.

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

[14] 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.

[15] 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. submitted.

[16] Mario Kleiner, Christian Wallraven, and Heinrich H. Bülthoff. The MPI VideoLab. Technical Report 104, Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany, 2003.

[17] 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.

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

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

[20] 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.

[21] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99 Conference Proceedings*, pages 187 – 194, 1999.