The 18th IEEE International Symposium on
Robot and Human Interactive Communication
Toyama, Japan, Sept. 27-Oct. 2, 2009

Feedback Interpretation based on Facial Expressions in Human–Robot Interaction

Christian Lang[1,3], Marc Hanheide[1,3], Manja Lohse[1,3], Heiko Wersing[2], and Gerhard Sagerer[1,3]

Abstract— **In everyday conversation besides speech people also communicate by means of nonverbal cues. Facial expressions are one important cue, as they can provide useful information about the conversation, for instance, whether the interlocutor seems to understand or appears to be puzzled. Similarly, in human–robot interaction facial expressions also give feedback about the interaction situation.**

We present a Wizard of Oz user study in an object–teaching scenario where subjects showed several objects to a robot and taught the objects' names. Afterward, the robot should term the objects correctly. In a first evaluation, we let other people watch short video sequences of this study. They decided by looking at the face of the human whether the answer of the robot was correct (unproblematic situation) or incorrect (problematic situation). We conducted the experiments under specific conditions by varying the amount of temporal and visual context information and compare the results with related experiments described in the literature.

I. Introduction

Human–robot interaction has received much attention in recent years. One important and crucial part of this research is to achieve a fairly natural communication between human and robot. To communicate successfully in dialog situations, people align at different levels during conversations [1]. People also adapt their discursive behavior when interacting with a robot depending on their beliefs about the knowledge and abilities of the robot that they acquired during the interaction [2]. On the contrary, the abilities of present robots to adapt their own behavior based on the behavior of their human interaction partners are rather limited.

In order to provide robots with sufficient communication skills for natural conversations with humans, besides the understanding of speech, also the recognition and interpretation of nonverbal cues is important, as these cues can provide useful information. Gestures might be used to support or complement speech [3]. Furthermore, the recognition and interpretation of facial expressions can yield important information about the interaction situation, for instance, whether the interlocutor seems to understand or appears to be puzzled.

The six emotional facial expressions happiness, anger, disgust, fear, surprise, and sadness according to Ekman [4] are not the most important ones in this context. According to our experiences, most of these emotional expressions occur much less frequently in human–robot interaction than facial expressions that carry some communicative semantics. Examples of this kind of "communicative" facial expressions are looking puzzled or disappointed, appearing to agree or disagree with the interlocutor, or seeming satisfied with or frustrated by the situation. In this context, we think about facial expressions in a broader sense which also includes head poses and head gestures, because they often carry a communicative meaning as well. However, emotional and communicative facial expressions are not disjunct. A repetitive failure of the robot in performing some task might cause anger or the behavior of the robot in a particular situation could be surprising, for instance.

[1] Research Institute for Cognition and Robotics, Bielefeld University
[2] Honda Research Institute Europe, Offenbach
[3] Applied Informatics, Faculty of Technology, Bielefeld University
E–Mail Contact: clang@cor-lab.uni-bielefeld.de

A. The Contribution of this Paper

In this paper, we report a Wizard of Oz user study in which we tried to provoke communicative facial expressions by letting people interact with a remote controlled robot in an object–teaching scenario. The users were instructed to teach the names of several objects to the robot, which was expected to term them correctly afterwards. The goal of this user study was to create a video corpus of people giving substantive nonverbal feedback by means of authentic, communicative facial expressions while interacting with a robot. This video corpus shall serve as test data for automatic feedback interpretation methods and other investigations in future work. To our knowledge, no such corpus is publicly available in the scientific community.

As a first evaluation of this corpus, we present a user study were we showed short video sequences to subjects who were instructed to interprete the shown facial expressions by distinguishing problematic (the robot misnamed the object) from unproblematic (the robot termed the object correctly) situations. This "problem detection" approach is an important, special case of feedback interpretation. The recognition performance of the subjects can serve as a baseline for the development of automatic recognition approaches. We compare our results with a study of Barkhuysen et al. [5]. The authors showed video fragments of people interacting with a spoken dialog system to subjects, who had to decide whether there was a communication problem in a particular interaction.

One motivation for our choice of this experiment for a first corpus evaluation was to avoid a common problem with experiments involving authentic, spontaneous facial expressions: the acquisition of reliable ground truth data for a classification of the displayed expressions into categories, such as basic emotions. This can be addressed by asking the subjects about their feelings in specific situations after the experiment, but this is not unproblematic. When asked after the experiment is over, it might be very difficult for a subject

978-1-4244-5081-7/09/$26.00 ©2009 IEEE

to remember the feeling in a particular situation. On the other hand, interrupting immediately after every interesting situation is likey to disturbe the experiment or influence the subject in an undesired way.

The situation is easier when the ground truth data can be aquired objectively — independent from the feelings of the subject. This is the case for the detection of communication problems in this study, as one usually knows for sure whether the answer of the robot was correct or not. In a sense, this is an inverse approach: instead of trying to find the correct ground truth data for given facial expressions, one looks for facial expressions in a given situation with implicitly given ground truth data. However, here another problem can arise: There is no guarantee that the subject will show one of the expected facial expression or a prominent feedback signal at all. Fortunately, it seems likely that people often will show some striking facial expression in the presence of a communication problem. This assumption is confirmed by the experiments of Barkhuysen et al. [5] and also by preliminary studies we carried out.

B. Paper Structure

The remainder of this paper is organized as follows. The next section II briefly discusses some related work. The subsequent section III describes the object–teaching corpus in detail. A first evaluation in terms of a feedback interpretation user study and the results of this study are presented in section IV and V, respectively. Section VI compares these results with related experiments in the literature and, finally, section VII draws conclusions and discusses future work.

II. Related Work

Most work about the detection of communication problems considers speech. Krahmer et al. [6] showed that people can correctly classify disconfirmation fragments of dialogs as positive or negative communication signals and concluded that prosodic features such as duration, intonation, and pitch are relevant for communication. The automatic recognition of user corrections in spoken dialog systems has been investigated by Hirschberg et al. [7]. Zhou et al. [8] conducted user studies to find cues to error detection that could be used to improve the error correction capabilities of speech recognition systems.

As humans are capable of reasonably interpreting nonverbal feedback, one wants to achieve this also for technical systems to improve the communication between humans and robots. Many techniques have been developed for automatic facial expression recognition in general; Fasel & Luettin [9], and Pantic & Rothkrantz [10] presented surveys on this topic. Most work considers the classification into the six basic emotion categories according to Ekman [4] or the recognition of facial actions in terms of the facial action coding system proposed by Ekman & Friesen [11]. Buenaposada et al. [12] recently presented a real–time capable system that can classify basic emotions. Bartlett et al. [13] have developed a system that classifies 20 action units. The system's performance was tested on a database of spontaneous facial expressions, in contrast to databases of posed facial expressions that are usually used. Sebe et al. [14] also created a database of spontaneous, authentic facial expressions. They noted that there is a remarkable difference between authentic and posed emotional facial expressions in visual appearance, as the latter ones are not "felt" by the subjects displaying them and thus do not correspond to their true emotional state.

III. The Object–Teaching Corpus

A. Motivation

The overall goal of our research is to enable a robot to make use of facial expressions to get substantive nonverbal feedback from its human interaction partner. We think that this is one important step to make the interaction more natural, i.e. more human–like. For the development of appropriate feature extraction and recognition methods for automatic interpretation, a corpus containing videos of interaction situations where the subjects give nonverbal feedback by means of authentic, communicative facial expressions is essential. To create such a corpus was therefore the goal of this object–teaching user study.

By evaluating videos from a previous user study [2], we found that the object–teaching scenario seems to be well suited, in general, to "provoke" communicative facial expressions, thus we chose this scenario for the user study. (A new study was necessary because the videos from the previous study do not contain close up views of the faces of the subjects, which is required for further analysis.)

B. Scenario

The participants were instructed to show several manipulable objects to the robot "Biron"[1][15] and to teach the objects' names. Furthermore, they should validate that the robot had actually learned the objects. It was not specified how they termed the objects or how they presented them (pointing to them or lifting them up). We performed a Wizard of Oz study where Biron was remote controlled to determine exactly its behavior (when to recognize the object correctly, when to misunderstand the subject, what to say, and where to look). Of course, the subjects did not know this beforehand, but assumed that the robot would act autonomously. The robot did not move but reacted to the subjects using speech production and movements of its pan tilt camera, for instance, to focus on the objects or the face. The study was conducted with 11 subjects (five female and six male) ranging from 22 to 77 years in age, nine of them had never interacted with the robot before. Per person, two counterbalanced sessions were performed: a "good" one where Biron termed most of the objects correctly, and a "bad" one where Biron misclassified the majority of objects. A session lasted about ten minutes. Between the sessions, the objects were exchanged to make the subjects believe that the recognition performance of the robot on another object set was to be evaluated. For each session, videos were

[1] Bielefeld Robot Companion

recorded from three different perspectives as shown in Figure 1. One stationary camera in front of a table with nine objects recorded the whole scene, showing the robot Biron on the left and the subject on the right. Another stationary camera was placed behind Biron to record the face of the subject during the whole experiment. Additionally, the videos taken by Biron's pan tilt camera were stored.

C. Corpus Description

To support the latter evaluation of the corpus in terms of facial expressions, all videos recorded by the stationary face camera were annotated. In addition to the transcription of the speech of both subject and robot, all object–teaching scenes were annotated and subdivided into four phases:

1) *present:* The subject presented the object to Biron and said its name or asked for the name.
2) *waiting:* The subject waited for the answer of the robot (not mandatory).
3) *answer:* The robot answered the subject, for instance, by classifying the object or asking a question.
4) *react:* The subject reacted to the answer of the robot.

These scenes sometimes overlapped, as a part of the react phase of one scene might be part of the present phase of the next scene. The exact times of the phases were sometimes ambiguous (especially the end of react or present phases). To achieve consistency nevertheless, all scenes were annotated according to a predefined coding scheme. Each of the object–teaching scenes was classified into one of the following categories, depending on the answer of Biron (example answers in parenthesis):

- *success:* Biron said the correct object name. ("So, this is a book." after the subject taught the object name or "This is a book." after the subject asked for the object's name)
- *failure:* Biron said an incorrect object name (same answer structure as in the success case).
- *problem:* There was a communication problem, but Biron did not say any object name ("I don't know the object.", "I don't know the word.", "I don't know.").
- *vague:* Biron claimed to understand, but did not say any object name ("I have seen the object.", "This is interesting.", "I like it.").
- *clarification:* Biron asked a clarification question ("Pardon?", "I could merely understand you partially. Can you repeat this, please?", "Did you show me the object before?").
- *abort:* Biron did not answer in a reasonable period of time thus the subject aborted this scene and taught a new object.

There were only very few cases where a scene did not match any of these categories. Those scenes were omitted. In addition to the phases, the period of time the robot said an object name (in "success" and "failure" scenes) was annotated. This was used for the feedback interpretation user study as reported in the next section.

Table I gives an overview of the number of scenes in the database. A total number of 751 scenes were annotated,

subject	succ	fail	prob	vagu	clar	abor
01	15	18	12	6	26	0
02	17	11	6	1	16	0
03	32	21	14	1	16	0
04	20	17	4	2	16	0
05	16	16	4	1	14	0
06	15	13	2	0	10	3
07	25	31	4	6	23	3
08	32	26	23	5	22	1
09	13	24	5	0	19	0
10	12	12	0	1	12	1
11	24	35	2	1	21	1
total	221	226	76	24	195	9

TABLE I: Number of object–teaching scenes of different categories and subjects in the video corpus.

providing a large test data set for evaluations. We succeeded in creating a suitable video corpus for nonverbal feedback analysis by means of authentic, communicative facial expressions and present a first evaluation in the next section.

IV. The Feedback Interpretation User Study

A. Motivation

The goal of the feedback interpretation user study was to find out how good humans are in distinguishing problematic from unproblematic interaction situations in our object–teaching scenario, depending on the available context information. The results shall serve as a baseline for automatic recognition approaches. The special case of "problem detection" was chosen due to the availability of reliable ground truth data, as discussed in section I-A.

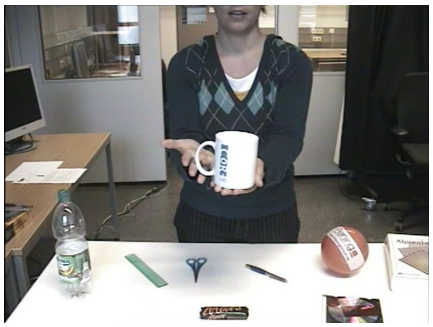
B. Material

We randomly selected 88 object–teaching scenes from the corpus: 44 "success" and 44 "failure" scenes (four success and four failure scenes for each of the 11 subjects). For each scene, we extracted a subpart of the associated video sequence from the stationary face camera, starting near the end of the answer phase, exactly when Biron started to say the object name, and ending at the end of the react phase. This starting point of the videos was chosen because it is the first moment from which the subject could know whether the answer of the robot was correct or not.

We presented these 88 video sequences to 44 subjects (15 female and 29 male, ranging from 16 to 70 years in age) who were not involved in the object–teaching user study for the corpus creation. The subjects decided whether the displayed situation was problematic (Biron named the object incorrectly) or unproblematic (Biron named the object correctly). To vary the amount of context information, we used four different variants of each video sequence: full length versus half length (starting from the beginning of the sequence in every case), each combined with showing the whole video in one case and only the face of the subject in

(a) Stationary scene camera

(b) Stationary face camera



(c) Biron's pan tilt camera

Fig. 1: The object teaching corpus contains videos from three perspectives showing (a) the whole scene, (b) the subject's face, and (c) the view of Biron's pan tilt camera.

the other case.[2] All videos were presented without sound. The video sequences were distributed over the 44 subjects such that the following conditions were met:

- Each subject saw each video sequence in one context variant only. To avoid the effect of priming, we decided not to show the same sequence twice (in different variants) to the same person.
- Each subject saw all 88 video sequenes (and thus four success and four failure scenes for each of the 11 subjects from the object–teaching study) in randomized order.
- Each subject saw exactly 22 videos in each of the four context variants: 11 "success" and 11 "failure" interaction situations (in randomized order).
- Summed up over all 44 subjects, each video was seen by 11 subjects in each of the four context variants.

V. Results

On average, the subjects of the feedback interpretation user study were able to classify the video segments with 79.1% recognition performance. We did not observe differences between female and male subjects, the classification rate was 79.1% for both. Figure 2 shows the recognition performance distributed over all 44 subjects for all videos, only "success" videos, and only "failure" videos, in each case for all four context variants:

- *all*: average over all context variants
- *fs-ft*: full scene and full time
- *fs-ht*: full scene and half time
- *of-ft*: only face and full time
- *of-ht*: only face and half time

The subjects are sorted by the recognition performance in each case. Table II lists the mean recognition performance and standard deviation.

There were big differences between the subjects, ranging from 89% to 59% on average. The visual context helped in the classification, as the performance was better for full scene videos compared to face–only videos, significantly for half time videos (t–test, $p < 0.01$) and very slightly only (not significantly) for full time videos ($p < 0.61$). The temporal context seemed to be even more important, as the performance was higher for full time videos compared to half time videos, and the difference was greater than for the visual context. This effect was significant for both full scene ($p < 0.03$) and face–only ($p < 0.001$) videos. On average, it was easier to classify failure videos than to classify success videos ($p < 0.011$). In both cases, the variance was higher than the total variance for all videos, because most subjects (26) were better in classifying failure videos than in classifying success videos, but for some subjects (12) the opposite was the case (six subjects performed equally well in either case).

sub-set	all videos		success videos		failure videos	
	mean	std	mean	std	mean	std
all	79.1	8.2	75.8	11.9	82.4	12.0
fs-ft	83.4	12.8	80.2	16.8	86.6	16.2
fs-ht	78.2	8.1	75.0	12.6	81.4	15.3
of-ft	82.0	11.1	78.1	16.3	86.0	13.5
of-ht	72.8	9.9	69.8	15.9	75.8	15.9

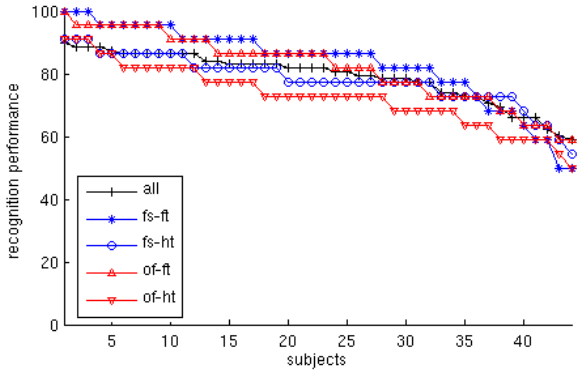
TABLE II: Mean value and standard deviation of the recognition performance for different video subsets (distribution over subjects). Please refer to section V.

Similar to Figure 2 and Table II, Figure 3 and Table III show the recognition performace distributed over all videos. The variance between different videos was even greater than

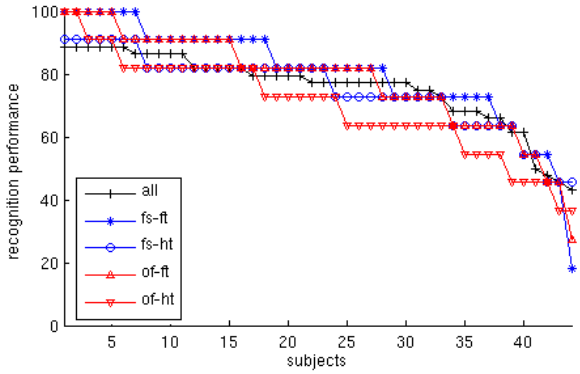
sub-set	all videos		success videos		failure videos	
	mean	std	mean	std	mean	std
all	79.1	17.9	75.8	19.4	82.4	15.8
fs-ft	83.4	18.1	80.2	21.1	86.6	14.1
fs-ht	78.2	24.0	75.0	27.2	81.4	20.1
of-ft	82.0	19.1	78.1	21.2	86.0	16.1
of-ht	72.8	23.9	69.8	25.8	75.8	21.7

TABLE III: Mean value and standard deviation of the recognition performance for different video subsets (distribution over videos). Please refer to section V.

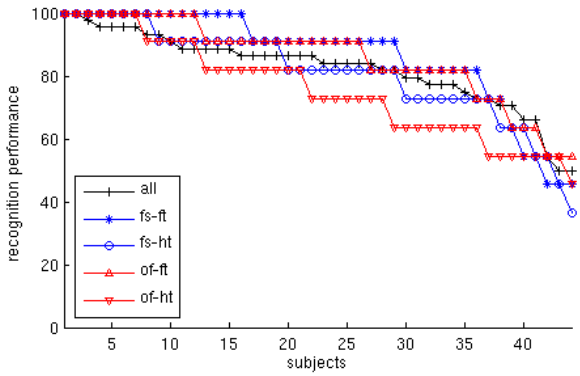
[2]The faces were located as rectangular regions using an automatic face detection approach (based on the work of Castrillón et al. [16]) that led to a kind of "glint" around the faces (as the face size varies somewhat) in some cases, also in a few cases the face detection got lost for a few frames. Videos where the face detection was too poor were rejected beforehand.



(a) Recognition performance for all videos

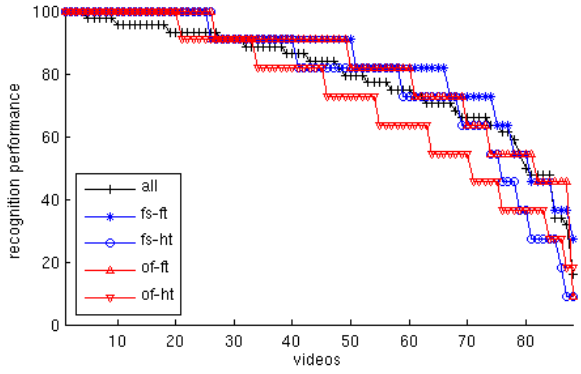


(b) Recognition performance for success videos

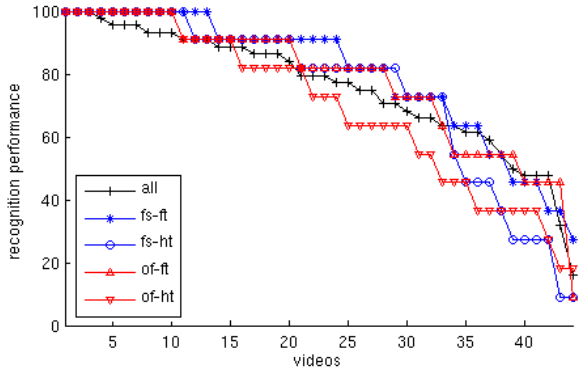


(c) Recognition performance for failure videos

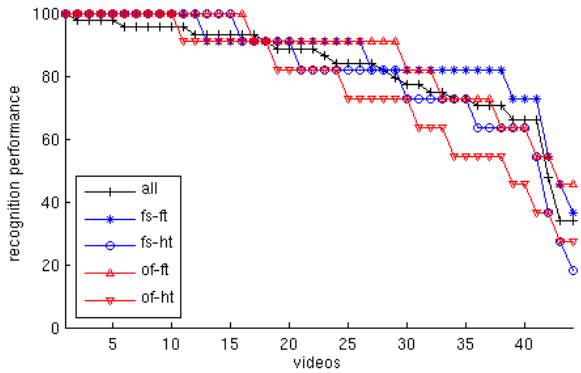
Fig. 2: Recognition performance for different video subsets (distribution over subjects). Please refer to section V.



(a) Recognition performance for all videos



(b) Recognition performance for success videos



(c) Recognition performance for failure videos

Fig. 3: Recognition performance for different video subsets (distribution over videos). Please refer to section V.

the variance between subjects: some videos were correctly classified in almost every case, whereas some other videos were systematically misclassified. The most recognized video showed a clearly visible nodding, the second most recognized video contained clear signs that the subject was perplexed. In the most poorly recognized video, the subject spoke a lot without clear affirmation, which was misinterpreted by most subjects as correcting the robot. The subject shown in the second most poorly recognized video displayed hardly any prominent facial expression at all. The variance for success videos was higher than for failure videos, which is consistent with the observation that failure videos were easier to classify on average, but there were also some success videos that were correctly classified in almost every case.

VI. Discussion and Comparison

Barkhuysen et al. [5] conducted three experiments where subjects watched film fragments of speakers interacting with an oral train timetable dialog system. The subjects decided whether or not there was a communication problem present in the shown situation. In the first experiment, the subjects saw a silent person listening to a confirmation question

of the system, where the system's confirmation was either correct or incorrect. About 75% of the subjects classified the videos correctly, and about 70% of the videos were significantly classified correctly. As in our study, some videos were significantly misclassified due to untypical behavior of the shown subject. These results and ours match fairly well. In the second experiment, the subjects watched videos of a speaker saying "no", either in response to a yes–no question or to indicate a misinterpretation of the system. Here the recognition performance was only slightly above chance level. This task seemed to be very hard, perhaps partially due to the short duration of the video sequences. Again, they observed great differences between different speakers. In the third experiment, the speakers uttered a destination, either in answering a question or to correct a misunderstanding. About two thirds of the subjects classified the videos correctly, and most of the videos were significantly classified correctly.

The partially higher recognition performances in our study might be due to the different settings. In our object–teaching scenario, the videos seem to contain more "implicit" context that could be used by the subjects. Asked about the features they (believed to) have used to classify the videos, some subjects mentioned aside from head gestures, "lipreading", and facial expressions also some "implicit" contextual features that were present in all four context variants to some degree: the length of the sequence respectively how much the person talked and whether the person seemed to put down the object at the end of the video.

In spite of the similarities of the experiments, there are also some important differences. Whereas Barkhuysen and her colleagues varied the shown video sequences, we used the same video sequences and varied the amount of displayed visual and temporal context. They presented the videos with sound, whereas we removed the sound from all videos. Barkhuysen et al. investigated differences between the videos respectively speakers shown in the videos, we also reported about differences in the recognition performance of the observing subjects.

VII. Conclusions and Future Work

In the Wizard of Oz object–teaching user study, we succeeded in creating a suitable corpus for evaluation in terms of nonverbal feedback by means of authentic, communicative facial expressions. This video corpus contains hundreds of interaction situations where subjects try to teach objects to a robot and give verbal und also nonverbal feedback; these sequences can be used in future investigations.

As a first evaluation of the corpus, we presented a feedback interpretation user study. The subjects in this study were able to distinguish problematic from unproblematic interaction situations with recognition performances between 73% and 83% on average, but there were in part large differences depending on the videos, the observing subjects, and the amount of context displayed. Our results are consistent with the results of Barkhuysen et al. [5], who conducted related experiments. We attribute the partially higher recognition rates in our studies to differences in the settings.

Future work will concentrate on the investigation of appropriate automatic recognition approaches for feedback interpretation in human–robot interaction in general and for problem detection in particular. Such an approach could be used to increase the ability of a robot to react and adapt to its human interaction partner and thus make the interaction more human–like. Important results to consider are that humans are capable of feedback interpretation to some extent even with very little context and that the temporal context seems to help the interpretation more than the visual context.

VIII. Acknowledgments

Christian Lang gratefully acknowledges the financial support from Honda Research Institute Europe for the project "Facial Expressions in Communication". The authors thank Sascha Hinte and Anton Helwart for their help in annotating the object–teaching user study videos and in performing the feedback interpretation study, respectively. We also thank the participants of the user studies and the respective preliminary studies.

References

[1] M. Pickering and S. Garrod, "Towards a mechanistic Psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[2] M. Lohse, K. J. Rohlfing, B. Wrede, and G. Sagerer, "Try something else! When users change their discursive behavior in human–robot interaction," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Pasadena, CA, USA, May 2008.

[3] K. Bergmann and S. Kopp, "Coexpressivity of speech and gesture: Lessons for models of aligned speech and gesture production," in *Proceedings of the Artificial and Ambient Intelligence Convention (AISB), Language, Speech and Gesture for Expressive Characters*, 2007.

[4] P. Ekman, "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique." *Psychological Bulletin*, vol. 115, no. 2, pp. 268–287, 1994.

[5] P. Barkhuysen, E. Krahmer, and M. Swerts, "Problem Detection in Human–Machine Interactions based on Facial Expressions of Users," *Speech communication*, vol. 45, no. 3, pp. 343–359, 2005.

[6] E. Krahmer, M. Swerts, M. Theune, and M. Weegelsa, "The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates," *Speech Communication*, vol. 36, no. 1–2, pp. 133–145, 2002.

[7] J. Hirschberg, D. Litman, and M. Swerts, "Identifying user corrections automatically in spoken dialogue systems," in *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 2001, pp. 1–8.

[8] L. Zhou, Y. Shi, D. Zhang, and A. Sears, "Discovering Cues to Error Detection in Speech Recognition Output: A User-Centered Approach," *Journal of Management Information Systems*, vol. 22, no. 4, pp. 237–270, 2006.

[9] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey," *Pattern Recognition*, vol. 36, pp. 259–275, 2003.

[10] M. Pantic and L. J. M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[11] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978.

[12] J. M. Buenaposada, E. Muñoz, and L. Baumela, "Recognising facial expressions in video sequences," *Pattern Analysis & Applications*, vol. 11, no. 1, pp. 101–116, 2008.

[13] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 223–230.

[14] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic Facial Expression Analysis," *Image and Vision Computing*, vol. 25, no. 12, pp. 1856–1863, December 2007.

[15] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, "BIRON — The Bielefeld Robot Companion," in *Proceedings of the International Workshop on Advances in Service Robotics*, E. Prassler, G. Lawitzky, P. Fiorini, and M. Haegele, Eds. Stuttgart: Fraunhofer IRB Verlag, May 2004, pp. 27–32.

[16] M. Castrillón, O. Déniz, and M. Hernández, "The ENCARA System for Face Detection and Normalization," *Lecture Notes in Computer Science*, vol. 2652, pp. 176–183, 2003.