Facial Expressions as Feedback Cue in Human–Robot Interaction — a Comparison between Human and Automatic Recognition Performances

Christian Lang[1,2], Sven Wachsmuth[1,2], Heiko Wersing[3], and Marc Hanheide[1,2,4]

[1] Research Institute for Cognition and Robotics (CoR–Lab), Bielefeld University, Germany
[2] Applied Informatics, Bielefeld University, Germany
[3] Honda Research Institute Europe, Offenbach, Germany
[4] School of Computer Science, University of Birmingham, England
E–Mail Contact: clang@cor-lab.uni-bielefeld.de

Abstract

Facial expressions are one important nonverbal communication cue, as they can provide feedback in conversations between people and also in human–robot interaction. This paper presents an evaluation of three standard pattern recognition techniques (active appearance models, gabor energy filters, and raw images) for facial feedback interpretation in terms of valence (success and failure) and compares the results to the human performance. The used database contains videos of people interacting with a robot by teaching the names of several objects to it. After teaching, the robot should term the objects correctly. The subjects reacted to its answer while showing spontaneous facial expressions, which were classified in this work. One main result is that an automatic classification of facial expressions in terms of valence using simple standard pattern recognition techniques is possible with an accuracy comparable to the average human classification rate, but with a high variance between different subjects, likewise to the human performance.

1. Introduction

Facial expressions provide one important nonverbal communication channel. People often give implicit feedback about a conversation by means of facial expressions, for instance by appearing to be interested or seeming to understand. One important goal of the research on automatic facial expression recognition in recent years is to enable a robot to communicate with humans in a fairly natural way. In order to achieve this goal, besides the understanding of speech, also the recognition and interpretation of facial expressions and other nonverbal cues is important, as they can provide useful imformation about the interaction situation.

We think that the six emotional facial expressions happiness, anger, disgust, fear, surprise, and sadness according to Ekman [7] are not the most important ones in this context. According to experiences in this field (as reported by Caridakis *et al*. [3] and Lang *et al*. [12], for instance), most of these emotional expressions occur much less frequently in human–robot interaction than facial expressions that carry some communicative semantics. Some examples of this kind of “communicative” facial expressions are looking disappointed or puzzled, appearing to agree or disagree with the interlocutor, or seeming satisfied with or frustrated by the situation. “Facial expressions” are considered in a broader sense in this context, also including head poses and head gestures, as they often carry a communicative meaning as well. However, emotional and communicative facial expressions are not disjunct. A repetitive failure of the robot might cause anger or the behavior of the robot could be surprising, so that the user might show the corresponding emotional facial expressions, which also imply a communicative meaning in these situations.

In this paper, we investigate the automatic recognition by means of standard pattern recognition techniques of a special type of communicative facial expressions: the recognition of *valence* in terms of *success* and *failure*, following the approach we used in previous work [12]. In this context, *success* means that a particular interaction with the robot could be performed as desired, whereas *failure* means that some problem occured. We think that in many practical interactions with robots, the detection of failure situations by means of facial expression interpretation would improve the interaction experience notably, even without a finer interpretation of the perceived facial expressions. For instance, the robot could change into a “problem solving” state and offer options that are applicable for many types of failures. To achieve this, the interpretation of a facial expression as signalling a failure would be sufficient, a finer classification ("angry", "sad", "disappointed", "puzzeld", etc.) is not essentiell in many cases (and would probably be very challenging). For the evaluations in this paper, we used a database of object–teaching scenes where several subjects showed objects to a robot and taught their names. One advantage of a pragmatic facial feedback recognition in terms of valence is that there is often an implicitly given ground truth label for the data, as one can usually decide whether an interaction succeeded or a problem occured based on comparatively objective criteria (for example whether the robot termed an object correctly or not), whereas the context–less acquisition of reliable ground truth data soley based on the visual appearance might be very difficult and subjective.

This paper is organized as follows. The next section briefly discusses some related works. Afterwards, the used database of object–teaching scenes is introduced. In section 4, the results of the investigated automatic recognition methods are presented and compared to the human performance in section 5. Finally, the last section draws conclusions and adds remarks about future work.

2. Related Work

Most work considers the classification into the six basic emotion categories according to Ekman [7] or the recognition of facial actions in terms of the facial action coding system proposed by Ekman & Friesen [8]. Fasel & Luettin [9] and Pantic & Rothkrantz [13] presented surveys on facial expression recognition techniques. Buenaposada *et al*. [2] presented a real–time capable system that can classify basic emotions. Bartlett *et al*. [1] 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 were usually used. In recent years, spontaneous facial expressions received increasing research attention. Sebe *et al*. [16] also created a database of spontaneous, authentic facial expressions. Zeng *et al*. [18] recently presented a survey that focusses on the recognition of spontaneous facial expressions.

Sebe *et al*. [15] added interest, boredom, confusion, and frustration to the six basic emotions and the neutral expression and investigated joint visual and audio emotion recognition, which performed significantly better than each modality alone. Also facial expression recognition using a dimensional emotion model (including dimensions such as "evaluation" and "activation") has been considered [3]. To our knowledge, there are so far only a few works that consider the direct automatic interpretation of facial expressions in terms of valence categories. Caridakis *et al*. [3] and also Fragopanagos & Taylor [10] used the two–dimensional "evaluation–activation space" and investigated the recognition of valence and activation level with neural networks.

3. Video Database

The video database used in this work is the object teaching corpus presented in a previous paper [12]. It contains videos of people interacting with the robot "Biron" [11] in an object–teaching scenario. The subjects taught the names of several objects to the robot, who should term the objects correctly afterwards. Figure 1 depicts some example images from the database. All object–teaching scenes in the videos were annotated and subdivided into four phases:

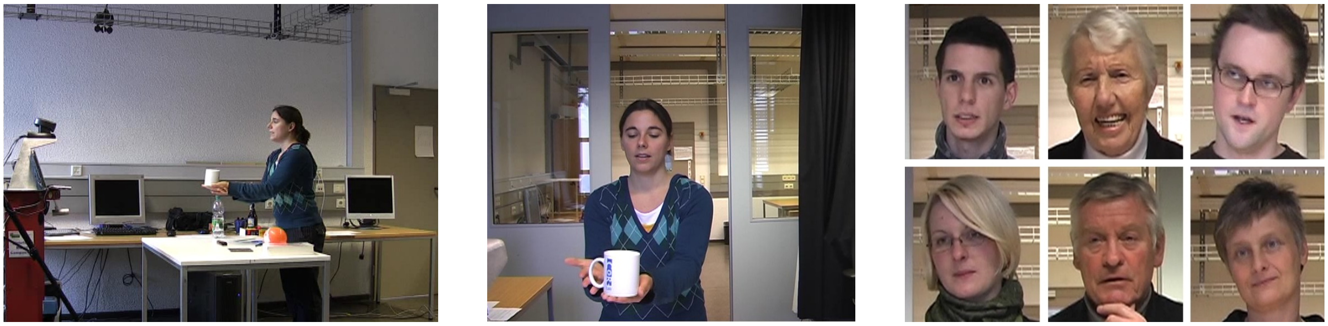
1. *present:* The subject presented the object to Biron and said its name or asked for the name.
2. *waiting:* The subject waited for the answer of the robot (not mandatory).
3. *answer:* The robot answered the subject, for instance, by classifying the object or asking a question.
4. *react:* The subject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

Furthermore, each object teaching scene was classified into a specific category, depending on the answer of the robot. Two categories are *success* and *failure*, meaning that the robot said the correct or a wrong object name in the answer phase. There are several other categories, which are not used in this paper. In total, there are 221 success and 227 failure scenes, distributed over 11 subjects, nine of which had never interacted with the robot before. The facial expressions that the subjects showed during the react phase can be considered as beeing authentic, because the subjects did not know beforehand that a Wizard of Oz study was performend and that facial expressions are important at all, but assumed that the object classification performance of an autonomously acting robot was to be evaluated.

In the previous paper where this database was presented we also evaluated the human interpretation performance in terms of valence recognition by letting other people watch and judge videos from the database. We extracted a subpart of each object–teaching scene, starting near the end of the answer phase, exactly when the robot started to say the object name, and ending at the end of the react phase. This starting point 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 used the same video segmentation for the evaluations in this paper. The sequence length is typically in the range from two to seven seconds (25 frames per second), a few videos are notedly longer.

4. Automatic Classification

This section reports the results of the conducted facial expressions classification in terms of valence using standard pattern recognition techniques. For each *success* and *failure* scene of the database an automatic face detection based on the approach of Castrillón *et al*. [4] was applied.



(a) scenario overview

(b) object–teaching scene

(c) examples of facial expressions

Figure 1. Example images from the used object–teaching video database.

It succeeded for 98% of the scenes, the remaining 2% were excluded from the evaluation.

Three different types of features were used: For each subject, we built an individual, hand–annotated active appearance model (aam) [5] with 55 landmarks placed over the face. The parameter vectors of the aam (when fitted to the images in the video sequences) were used as feature vectors for each frame. The aam fitting was initialized based on the method described by Rabie *et al*. [14]. As second feature extraction method, we applied a bank of 40 gabor energy filters, consisting of eight equally spaced orientations and five spatial frequencies with wavelengths of 1.17, 1.65, 2.33, 3.30, and 4.67 standard iris diameters (seventh part of the distance between the eye centers), as used by Whitehill *et al*. [17]. This filter design was found to be well suited for face recognition [6, 17]. We also used the face images directly as features.

4.1. SVM Majority Voting over Frames

We used a support vector machine (svm) classifier with radial basis function (rbf) kernel. The evaluation was conducted by a leave–one–out cross validation for all videos of a subject: all frames of all videos except one were used for the training, then the excluded video was classified via a majority voting over the single frames.

4.1.1 Meta Parameter Selection

In order to evaluate the effectiveness of the svm classifier in the given scenario, we performed a grid search to find good meta parameters (rbf parameter σ and regularization cost C), using a 10–fold cross validation for each parameter combination, over all frames of all videos of a subject. Afterwards the training and test of the classifier was executed as described in section 4.1. The results for different variants of the features are summarized in table 1: aams with 95% and 99% pca variance preservation (aam-95 and aam-99), gabor energy filters with response images scaled down to $4x4$, $8x8$, and $12x12$ (gab-*size*), and the raw face images scaled down to $8x8$, $16x16$, and $25x25$, for both gray level and rgb images (gray-*size* and rgb-*size*).

feature	all scenes		success		failure	
variants	mean	std	mean	std	mean	std
aam-95	63.6	23.1	54.8	27.1	69.8	23.9
aam-99	**76.1**	10.3	66.6	18.0	83.2	12.2
gab-4	**72.8**	12.3	65.7	28.4	76.3	15.5
gab-8	73.1	11.6	66.5	27.9	76.0	15.4
gab-12	71.3	12.9	64.4	27.9	74.9	17.3
gray-8	73.3	13.1	69.3	23.7	73.3	19.6
gray-16	**75.1**	12.5	67.5	25.3	79.0	14.6
gray-25	74.8	14.1	66.5	30.4	78.9	16.1
rgb-8	72.5	13.9	63.8	28.1	77.6	14.6
rgb-16	72.1	14.3	65.9	28.0	74.2	17.8
rgb-25	68.2	16.7	60.9	29.7	70.8	27.1
img-aam	70.5	11.1	64.0	21.2	73.3	18.3

Table 1. Mean value and standard deviation of the classification performance for all videos, only success and only failure videos (distribution over subjects), each for different features. Please refer to sections 4.1.1 and 4.1.2.

On the one hand, the classification rates are rather low for a two–class problem. On the other hand, the classification problem is expected to be hard, as the average human performance is only 82% [12] (please see section 5). For the subsequent investigations, we used only the best performing variant of each feature (marked in bold in table 1), except for the gabor energy filters, where we used variant "gab-4" instead of "gab-8" because of the lower feature vector dimensionality (640 compared to 2,560) and the only marginal difference in the classification rate (0.3% means just one more video classified correctly).

In real applications, it is not possible to use all feature vectors to find optimal meta parameters, as the test data is unknown and not available for meta parameter optimization. Therefore we conducted new grid searches, this time prior to each training, using only the respective training set of the svm for the search. Furthermore, it is not desirable

feature variants	all scenes mean	 std	success mean	 std	failure mean	 std
aam-gs	74.2	11.0	63.0	19.1	82.5	14.3
aam-av	75.4	10.1	64.1	19.3	84.0	12.5
img-gs	74.5	12.9	67.3	24.8	78.1	16.6
img-av	74.4	12.9	67.7	25.5	76.8	18.0
gab-gs	72.1	12.7	65.3	27.9	75.4	15.6
gab-av	72.4	12.5	65.4	28.1	76.0	15.4

Table 2. Mean value and standard deviation of the majority voting classification performance for all videos, only success and only failure videos (distribution over subjects), each for different features and meta parameters (*feature*-gs: individual grid search for each training process, *feature*-av: average meta parameters of these grid searches) Please refer to section 4.1.1.

for each training process to have its own set of meta parameters, as usually a certain stability of these parameters is required for practical usage. In order to estimate this stability, we tested the classifiers for the third time, using the mean σ and C values from the second test for all training processes. The results of these tests are listed in table 2. The classification rates are only slightly lowered, and the best meta parameters in the second grid search test were usually clusterd in a certain region in the search space. This supports the assumption that good meta parameters can be selected without knowing all data beforehand. For the subsequent evaluations, the results from the first grid search were used.

4.1.2 Feature Comparison

The raw image features compare surprisingly well to the active appearance models. The reason behind is that about 19% of the frames needed to be rejected from the aam classification, because the model fitting was too poor, mainly due to too large head rotations. If the raw image feature performance is evaluated only on those frames that are used for the aam tests, the classification rates decrease notably, as listed in the last row of table 1.

Surprisingly, the gabor energy filters yielded the lowest classification rates. Theoretically, they are expected to outperform the raw image features. We surmise that compared to the amount of available training data, the dimension of the feature vectors is too high, even though the gabor responses are highly downscaled (which might be a problem in its own), making it difficult to find appropriate class borders. It might be beneficial to use less filters with a higher resolution for future tests. In the remainder of the paper, we continue the investigations for aams and images only.

4.1.3 Classification Details

The classification performances of the aam and image features for each of the 11 subjects are listed in the left columns of table 3. The variance of the classification rates is very high, ranging from very good to very poor, even systematic misclassifications occur. We think that this difficulty of the classification problem is due to the high intraclass variance, compared to the interclass variance. As a rough estimate of these variances, we computed the mean pairwise euclidean distances between all success and all failure frames separately (mean intraclass distance), and also the mean pairwise euclidean distance between all success and all failure frames of each subject (mean interclass distance). The distances are listed in the right columns of table 3. The mean intra- and interclass distances are of comparable sizes, which is an indication of the difficulty of the classification problem. There is a significant correlation between the classification rate and the ratio of interclass to itraclass distance, the latter represented as the sum of the intraclass distances of the two classes (Spearman's rank correlation coefficient, $r = 0.77$, $p = 0.0059$ for aam features and $r = 0.61$, $p = 0.0484$ for image features). This supports our conjecture that a low interclass to intraclass variance ratio is the main reason for misclassifications in the investigated scenario.

For the aam features, the classification performance is also correlated to the percentage of selected support vectors (17% on average, 7.3 standard deviation) to some extent (close to significance, Spearman test, $r = -0.58$, $p = 0.0590$), which reflects the problem difficulty also in terms of model complexity. This correlation does not hold for the image features (19% support vectors on average, 8.4 standard deviation, $r = -0.19$, $p = 0.5703$).

4.2. SVM Mean Feature Vector Classification

In section 4.1, the results of a simple majority voting over single frames are reported. This section considers an even simpler approach: each video is represented by one feature vector only, namely the mean vector of its frames. This simple classification method yielded surprisingly good results, outperforming the previous majority voting scheme.

4.2.1 Classification Performance

The classification performances for the mean feature vectors are summarized in table 4. The aam features with meta parameters selected via a cross validation grid search over all training data performed best, but also the classification rate of the image features improved, compared to the majority voting. The results for meta parameters selected by an individual grid search over the training data prior to each training and the mean parameters of these grid searches (please refer to section 4.1.1) are a few percent lower. This difference is greater than in the majority voting case, showing a higher sensitivity to the meta parameter selection. We attribute this to the drastically decreased number of feature

subject	classification rates			mean distance values		
	all	succ	fail	succ	fail	inter
aam-01	85	80	89	25.5	22.1	26.8
aam-02	72	65	83	23.0	17.6	21.3
aam-03	83	82	83	26.9	19.3	24.5
aam-04	95	90	100	30.9	21.8	29.9
aam-05	84	75	94	39.9	28.7	37.5
aam-06	64	67	62	44.3	47.2	46.8
aam-07	64	48	77	27.4	29.3	29.0
aam-08	67	72	62	29.8	20.3	27.4
aam-09	69	25	91	22.1	21.8	23.0
aam-10	71	58	83	23.0	33.0	29.4
aam-11	83	71	91	25.4	17.1	24.6
img-01	91	93	89	3.03	2.09	2.73
img-02	66	76	50	1.89	1.54	1.80
img-03	80	86	72	2.52	2.16	2.43
img-04	97	95	100	3.11	2.08	2.92
img-05	88	81	94	2.85	2.83	2.95
img-06	71	73	69	3.14	3.24	3.23
img-07	59	32	81	2.17	2.15	2.19
img-08	72	81	62	2.68	1.85	2.43
img-09	60	17	83	1.43	1.44	1.48
img-10	71	58	83	2.42	2.89	2.75
img-11	71	50	86	2.39	1.70	2.25

Table 3. Classification details for majority voting classification. Left: Classification rates for all videos, only success and only failure videos for all 11 subjects, each for aam and image features. Right: Mean pairwise inter- and intraclass feature vector distances. Please refer to section 4.1.3.

feature variants	all scenes		success		failure	
	mean	std	mean	std	mean	std
m-aam	**82.1**	10.1	76.1	14.0	86.8	10.2
m-img	80.0	10.0	76.4	21.4	80.7	12.6
m-aam-gs	76.0	11.5	70.3	19.2	79.2	13.3
m-aam-av	77.6	11.6	73.5	16.7	80.3	10.1
m-img-gs	73.5	10.5	66.2	25.5	77.5	11.6
m-img-av	76.3	11.2	68.7	25.7	80.4	13.3

Table 4. Mean classification performance and standard deviation for mean vector features, each for all scenes, only success, and only failure scenes. Abbreviations likewise to table 2.

subject	classification rates			mean distance values		
	all	succ	fail	succ	fail	inter
aam-01	91	80	100	12.9	16.1	18.4
aam-02	79	76	83	14.4	11.2	13.8
aam-03	87	89	83	24.9	23.6	25.4
aam-04	97	95	100	20.2	12.0	21.0
aam-05	88	88	88	32.6	22.8	30.3
aam-06	68	67	69	39.0	38.5	38.5
aam-07	66	57	73	19.2	20.3	20.5
aam-08	81	72	92	25.3	15.5	23.2
aam-09	74	50	87	19.2	13.8	17.6
aam-10	79	75	83	18.8	17.1	18.2
aam-11	93	88	97	18.3	11.6	20.0
img-01	94	87	100	1.73	1.31	1.70
img-02	83	88	75	1.22	1.12	1.22
img-03	76	86	61	1.69	1.83	1.80
img-04	89	85	94	1.62	1.03	1.79
img-05	88	94	81	1.64	1.45	1.63
img-06	79	93	62	1.43	1.71	1.61
img-07	63	28	90	1.48	1.31	1.42
img-08	76	75	77	1.52	1.29	1.62
img-09	63	42	74	1.04	0.83	0.97
img-10	83	83	83	1.78	1.60	1.67
img-11	86	79	91	1.67	1.26	1.67

Table 5. Classification details for mean feature vector classification. Left: Classification rates for all videos, only success and only failure videos for all 11 subjects, each for aam and image features. Right: Mean pairwise inter- and intraclass feature vector distances. Please refer to section 4.2.2.

vectors. However, even most of these classification rates are better than the corresponding rates in the majority voting.

4.2.2 Classification Details

Table 5 shows the mean classification rates, intra- and interclass distances for all subjects, likewise to table 3 in the majority voting case. The average classification performance improved for all 11 subjects for the aam features and for eight subjects for the image features. The correlation between classification performance and ratio of inter- to intraclass distance is now stronger for the aam features (Spearman test, $r = 0.85$, $p = 0.0010$) and is weakened beyond the significance level for the image features ($r = 0.54$, $p = 0.0896$). For both feature types, the correlations between percentage of selected support vectors and classification performance are not significant ($r = -0.45$, $p = 0.1686$ for aam features and $r = -0.57$, $p = 0.0686$ for image features). Due to the much smaller number of training vectors, a higher percentage is choosen as support vectors (aam features: 68%, 17.9 standard deviation; image features: 69%, 15.7 standard deviation). The training error is usually in the range of 15% – 30%, whereas it is below 1% in very most cases in the majority voting classification. This is natural to some degree due to the differences in the amount of training data, but it might also indicate some overfitting in the majoriy voting.

4.2.3 Comparison to Majority Voting

In order to investigate why the mean feature vectors performed better than the majority voting over frames, we considered those videos where the two approaches disagreed in their classification. This was the case for 77 scenes for the aam features and 86 scenes for the image features. The classification of the mean features was correct in 51 and 53 cases, respectively. In an inspection of these scenes it was found that very often there are one or two subsequences where almost all frames are wrongly classified, and also one or two subsequences where almost all frames are classified correctly. In cases where the former outnumbered the latter ones in terms of total length, the scene was necessarily misclassified by the majority voting scheme. In contrast, the mean vector of all frames could still capture important characteristics of the associated class and hence allow for correct classification. Visual inspection of the videos also led to the conclusion that often only a (possibly short) subsequence of the video is discriminative in terms of valence interpretation, although the videos were already segmented to contain only important information, according to the given annotations. For those scenes, majority voting over the complete video sequence is not well suited. Instead, an automatic detection of important subsequences would be very beneficial and remains for future work.

5. Comparison to the Human Performance

In the paper were the database was introduced [12], we also evaluated the human recognition performance in facial feedback interpretation in terms of valence. We randomly chose 88 videos from the database (four success and four failure videos of each subject) and showed them to 44 new subjects who should interprete the videos in terms of valence. All videos were presented without sound and in four different context conditions: showing the full scene or only the face region of the video sequence, each combined with showing the video sequence over the full length or only the first half of the video. The condition where only the face of the subject is shown over the full length of the scene (according to the annotations) matches best with the information the automatic classification approaches considered in this paper can use, as they also process the whole video without any visual context. The average human recognition performance for this condition was 82%, with a high variance over both observing subjects and shown videos. The results are summarized in table 6.

The best performing automatic recogniton approach considered in this paper, the mean aam features, reached the average human recognition performance. When evaluated on the above mentionend 88 videos only (instead of all available videos), the performance of the mean aam features even exceeded the human performance, as shown in table 6. For comparison, also the performances of the image features and the majority voting are listed. The mean image features also reached the human performance when evaluated on the video subset shown to the human observers. Also in the majority voting case, the aam features performed better on this subset, whereas the image features yielded worse results. Further commonalities between human and automatic recognition performances are that on average failure scenes were easier to classify than success scenes, a higher variance for success than for failure scenes, and a high variance of the classification rate depending on the subject resp. video in general.

class-	all scenes		success		failure	
ifier	mean	std	mean	std	mean	std
human	82.0	19.1	78.1	21.2	86.0	16.1
m-aam-a	82.1	10.1	76.1	14.0	86.8	10.2
m-aam-s	86.6	8.8	79.5	15.1	93.2	11.7
m-img-a	80.0	10.0	76.4	21.4	80.7	12.6
m-img-s	82.1	15.0	79.5	18.8	84.1	16.9
aam-a	76.1	10.3	66.6	18.0	83.2	12.2
aam-s	78.7	14.9	70.5	27.0	86.4	17.2
img-a	75.1	12.5	67.5	25.3	79.0	14.6
img-s	71.8	22.3	59.1	35.8	84.1	16.9

Table 6. Comparison of the performances of human recognition, mean aam features evaluated on all videos (m-aam-a) and only on those 88 videos the human subjects judged (m-aam-s), likewise for mean image features (m-img-a and m-img-s). Also the performances for the majority voting for aam and image features when evaluated on all videos (aam-a and img-a) and only these 88 videos (aam-s and img-s) are listed. Please refer to section 5.

In order to evaluate whether the human observers and the svm classification using the mean aam features tended to make the same classification errors, we binarized the classification results for the 11 observing subjects[1] for each video by setting the classification result to 1 if six or more subjects classified it correctly, and to 0 otherwise. This binarization was done to become compatible with the results of the automatic recognition, which yield only one binary value (correct or false classification) for each video. It turned out that there is a significant correlation between these classification results on the 88 videos (Pearson's correlation coefficient, $r = 0.25$, $p = 0.0187$), indicating that indeed the human observers and the automatic classification tended to make the same classification errors to some extent.

6. Conclusions and Future Work

We demonstrated that it is possible to reach the human performance in facial expression interpretation in human–robot interaction in terms of valence with a surprisingly simple approach using standard pattern recognition techniques, when a subject–specific classification is performed. However, the classification with the best performing mean aam and image features requires an appropriate meta parameter selection. Likewise to the human classification, the

[1] There were 44 observing subjects, who are distributed over the four context conditions, thus resulting in 11 observing subjects for each context condition, not to be confused with the 11 subjects shown in the videos.

variance of the recognition performance is very high, and on average failure scenes are easier to classify than success scenes. The investigation of the surprisingly good performance of the mean feature vectors compared to the majority voting over frames indicated that the detection and usage of descriminative subsequences of the videos might be very benefical and shall be investigated in future work. Despite the achievement of the human average recognition performance, the classification rates are rather low for a two–class problem, espescially for the success class. We assume that this can be improved be using more sophisticated classification approaches, for instance based on subsequence analysis. One main problem to deal with is the comparatively low interclass to intraclass variance ratio, measured on frame level.

The good performance of the raw images compared to the active appearance models show that also the video parts with large out–of–plane head rotations (which are problematic for the aam fitting and were thus rejected in the aam tests) convey useful information and should be considered for the interpretation. Although shown to yield good results on other facial analysis problems, a bank of 40 gabor filters performend worse than aam and image features in our evaluations. All features that were used in this paper operate on single frames. In future work pattern recognition methods that consider the temporal dynamics shall be evaluated. Furthermore, this paper considered only subject–specifc interpretation of facial expressions. Generalization between different subjects is expected to be much more difficult and a main target of future work. Also the automatic segmentation of interesting video segments is to be investigated, as so far presegmented scenes were used.

7. Acknowledgements

Christian Lang gratefully acknowledges the financial support from Honda Research Institute Europe for the project "Facial Expressions in Communication".

References

[1] 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*, pages 223–230, 2006. 2

[2] J. M. Buenaposada, E. Muñoz, and L. Baumela. Recognising facial expressions in video sequences. *Pattern Analysis & Applications*, 11(1):101–116, 2008. 2

[3] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *8th International Conference on Multimodal Interfaces*, pages 146–154, 2006. 1, 2

[4] M. Castrillón, O. Déniz, and M. Hernández. The ENCARA System for Face Detection and Normalization. *Lecture Notes in Computer Science*, 2652:176–183, 2003. 2

[5] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In H. Burkhardt and B. Neumann, editors, *Proceedings European Conference on Computer Vision*, volume 2, pages 484–498. Springer, 1998. 3

[6] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999. 3

[7] P. Ekman. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2):268–287, 1994. 1, 2

[8] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978. 2

[9] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003. 2

[10] N. Fragopanagos and J. Taylor. Emotion recognition in humancomputer interaction. *Neural Networks*, 18(4):389–405, 2005. 2

[11] 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 E. Prassler, G. Lawitzky, P. Fiorini, and M. Haegele, editors, *Proceedings of the International Workshop on Advances in Service Robotics*, pages 27–32, Stuttgart, May 2004. Fraunhofer IRB Verlag. 2

[12] C. Lang, M. Hanheide, M. Lohse, H. Wersing, and G. Sagerer. Feedback Interpretation based on Facial Expressions in Human–Robot Interaction. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 189–194, 2009. 1, 2, 3, 6

[13] 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*, 22(12):1424–1445, 2000. 2

[14] A. Rabie, C. Lang, M. Hanheide, M. Castrillón-Santana, and G. Sagerer. Automatic Initialization for Facial Analysis in Interactive Robotics. In *Proceedings of the International Conference on Computer Vision Systems*, pages 517–526, Santorini, Greece, May 2008. Springer. 3

[15] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang. Emotion Recognition Based on Joint Visual and Audio Cues. In *18th International Conference on Pattern Recognition*, volume 1, pages 1136–1139, 2006. 2

[16] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang. Authentic Facial Expression Analysis. *Image and Vision Computing*, 25(12):1856–1863, December 2007. 2

[17] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward Practical Smile Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2106–2111, 2009. 3

[18] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. 2