Facial Communicative Signal Interpretation in Human-Robot Interaction by Discriminative Video Subsequence Selection

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

***Abstract*— Facial communicative signals (FCSs) such as head gestures, eye gaze, and facial expressions can provide useful feedback in conversations between people and also in human-robot interaction. This paper presents a pattern recognition approach for the interpretation of FCSs in terms of valence, based on the selection of discriminative subsequences in video data. These subsequences capture important temporal dynamics and are used as prototypical reference subsequences in a classification procedure based on dynamic time warping and feature extraction with active appearance models. Using this valence classification, the robot can discriminate positive from negative interaction situations and react accordingly. The approach is evaluated on a database containing videos of people interacting with a robot by teaching the names of several objects to it. The verbal answer of the robot is expected to elicit the display of spontaneous FCSs by the human tutor, which were classified in this work. The achieved classification accuracies are comparable to the average human recognition performance and outperformed our previous results on this task.**

I. Introduction

Facial communicative signals (FCSs) such as head gestures, eye gaze, and facial expressions are one important means of nonverbal communication. People often use them to give implicit feedback about a conversation, for instance by appearing to understand or seeming to be puzzled. In order to move towards a fairly natural communication and collaboration between humans and robots, the recognition and interpretation of FCSs are important capabilities a robot should have, as they can provide useful information about the current interaction.

This paper presents an approach for the recognition of FCSs in task-oriented human-robot interaction based on the selection of prototypical reference subsequences for a k-nearest-neighbor-based (k-NN) classification method. The following Sec. II briefly introduces related work. Sec. III describes the motivation for our valence-based approach to FCS recognition, then the scenario and video database used for its evaluation are introduced in Sec. IV. The utilized face detection and feature extraction techniques are addressed in Sec. V. Subsequently, the main contribution of this paper—the recognition approach based on reference subsequence selection—is explained in Sec. VI and evaluated in Sec. VII. Finally, Sec. VIII concludes and remarks on future work.

[1] Research Institute for Cognition and Robotics, Bielefeld University
[2] Applied Informatics, Bielefeld University
[3] School of Computer Science, University of Lincoln
[4] Honda Research Institute Europe, Offenbach
E-Mail Contact: marc@hanheide.net

II. Related Work

Several researchers provided comprehensive surveys of automatic approaches for face detection [1], visual head pose estimation [2], eye tracking [3], and facial expression recognition [4], [5], [6]. Please refer to [7] and these surveys for a deeper discussion. Our approach is based on prototypical video subsequences in a k-NN-based classification scheme. The shapelet method of Ye and Keogh [8] also uses prototypical reference subsequences, but considers real-valued time series only and utilizes another distance measure and an entirely different prototype selection. Nowozin *et al.* [9] used discriminative subsequences in a classification approach based on Gabor filters, visual words, and boosting classifiers. Apart from the usage of discriminative subsequences, their approach is quite different from ours, and it is intended for human action classification, whereas we target FCS recognition. Buenaposada *et al.* [10] also used a NN-based classifier, but addressed temporal dynamics and prototype selection in fully different ways. Most methods that find specific subsequences focus on the performant computation of frequent patterns [11], leaving their quality assessment for subsequent processing steps. Our approach does not search for frequent patterns first, but directly tries to estimate the quality of the considered subsequences in terms of expected discrimination power.

We use active appearance models (AAMs) [12] for the extraction of facial features, which were also utilized by many others [13], [14], [15]. For comparison, we also performed some experiments with constrained local models (CLMs) [16], [17], which are also widely used for face recognition. As we explain in Sec. III, we investigate FCS recognition in terms of *valence*. Several other researchers also performed visual valence recognition. The utilized methods include neural networks [18], [19], fisher features and boosting [20], rule-based neurofuzzy networks [21], and facial action unit spectrograms [22]. While most early works considered posed facial expressions, there is a growing interest in authentic, spontaneous facial expressions nowadays [23], [24], [25]. An important issue with the recognition of authentic, spontaneous FCSs is the definition of the categories in whose terms the interpretation is performend. We take a different approach here than the research cited above and motivate and explain it in the next section.

III. Motivation

The way people use FCSs in human-human interaction has been investigated in a vast amount of psychological research; please see [7] for a discussion. Due to the complex and in

(a) scenario overview (b) object-teaching scene (c) examples of displayed FCSs

Fig. 1. Example images from the used object-teaching video database. Please refer to Sec. IV.

large parts controversial nature of these signals, we suggest to take a pragmatic view in human-robot interaction and to focus on scenario-specific investigations instead of trying to build general purpose systems for comprehensive FCS recognition, at least for the midterm development of the field [7]. We consider facial expression recognition to illustrate this point: Often the six basic emotional expressions happiness, anger, disgust, fear, surprise, and sadness according to Ekman [26] have been used as classification categories, due to their universality (although this is controversial [27]). However, these facial expressions are not the most important ones in interaction situations as most of them rarely occur in everyday life in a pronounced way and even less in human-robot interaction [28], [19], [29]. Thus, many works used posed facial expressions (e.g. [10], [30]), which are quite different from authentic, spontaneous ones (e.g. [31], [23]).

On the contrary, facial expressions that carry some communicative semantics as proposed by Fridlund [32] are much more frequently displayed in those interaction situations. 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. Fridlund [32] argued that there are no prototypical displays of certain communicative facial expressions as their meaning depends heavily on the context. Hence, we suggest to investigate the automatic recognition of FCSs in specific interaction scenarios, i.e. in a certain context.

Another problem is the definition of classification categories and the acquisition of reliable ground truth data. Spontaneously displayed FCSs are often difficult to interpret in terms of precise categories. Thus obtaining ground truth labels by human raters judging recorded interaction videos might be very subjective and ambiguous.[1] Also interviewing the participants about the intended meaning of their facial displays is not feasible in many cases.

To cope with this problem, we used an approach different from the usual practice: we defined the ground truth labels in terms of the interaction situation. In our scenario (please see Sec. IV), a particular interaction with the robot can either be *successful* or *problematic*, and this can be objectively determined from the situation. The FCSs displayed in these situations are treated as examples for one of two classes (*success* and *failure*). As already argued earlier [33], we think that in many practical interactions with robots, the detection of failure situations by FCS interpretation would improve the interaction experience notably, as the robot could change into a "problem solving" state and offer options that are applicable for many types of failures, for instance. A finer classification of the displayed FCS ("sad", "disappointed", "puzzeld", etc.) is not essentiell to achieve this.

While this approach yields reliable ground truth labels, it faces another problem: As the definition of these labels is independent of the visual appearance, there is no guarantee that a meaningful FCS is displayed at all, however, studies [34], [29] suggest that usually a meaningful display occurs. Thus, the research question investigated in this work is not the standalone interpretation of FCS in itself (as in most work on facial expression recognition), but their interpretation as feedback about the interaction in terms of valence, and the question to which degree this feedback can be gained from FCSs at all. One can regard this as interpretation on *pragmatic* level, while the former is on *semantic* level. This definition of valence is also different from the definition used in most other works on valence recognition [18], [19], [35], where the visual appearance of the face is rated by human coders in order to get a ground truth valence value. An exception is the work of Barkhuysen *et al.* [34], who used the correct or wrong understanding of a spoken dialog system to define a positive or negative ground truth value, which is very similar to our approach. They conducted several user studies, but did not report results of automatic recognition approaches. Please refer to [29] for a comparison of these studies to our object-teaching study, which is briefly introduced in the following section.

IV. Scenario and Video Database

For the evaluation of our approach, we used the object-teaching scenario introduced in previous work [29]: A person teaches the names of several objects to a robot, which

[1] In a pre-study of previous work [29], several people judged videos of participants teaching objects to a robot. These human raters did neither agree on the number of FCSs nor on the labels that should be used to describe the observed FCSs.

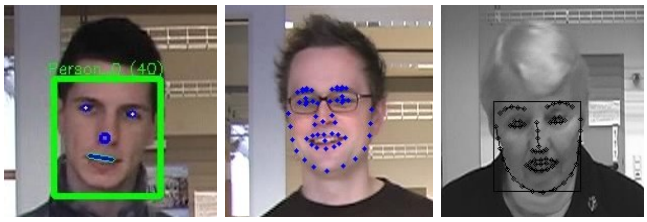

Fig. 2. Example images for the face detection (left) and feature extraction with AAMs (middle) and CLMs (right) Please refer to Sec. V.

is expected to term them correctly afterwards (please see Fig. 1(a) for a scenario overview). In its verbal answer, the robot will either say the correct or a wrong object name. The facial display of the human tutor during the answer of the robot and her or his reaction to that answer constitutes video data of the respective category: *success* in case of a correct answer, or *failure* if the answer is wrong. The video database recorded in this scenario contains 221 *success* and 227 *failure* scenes, distributed over 11 participants (please see Fig. 1). The videos are segmented to contain only the relevant part of the interaction, i.e. the reaction of the tutor to the answer of the robot.[2] In this Wizard of Oz study the elicited FCSs are authentic and spontaneous, as the participants did not know beforehand that FCSs are the subject of study. They were deceived to believe that the object classification performance of an autonomously acting robot was to be evaluated. For further details on this scenario and the recorded video database, please refer to [29].

V. Face Detection and Feature Extraction

For each *success* and *failure* video in the database, an automatic face detection based on the approach of Castrillón *et al.* [36] was applied. The feature extraction on frame level is done by an AAM [12]. For each human tutor, we used an individual AAM, built from hand-annotated images with 55 landmarks placed over the face, because person-specific AAMs are known to yield better fitting results than generic ones [37]. In order to fit to an input image, an AAM needs a suitable initializion, which is provided by overlaying the mean AAM shape on the detected face, based on the method described by Rabie *et al.* [38]. The parameter vector of the AAM (when fitted to a particular face image in the input video sequences) is used as feature vector for the respective frame. For comparison, we also conducted experiments with generic CLMs [17]. Fig. 2 shows example images of these feature extraction methods.

VI. Reference Subsequence Selection For Facial Communicative Signal Classification

This section describes our FCS classification approach, where short video subsequences with high discriminative power are used as prototypical representatives for the two classes. This is motivated by a previous evaluation [33] revealing that apparently only a short subsequence of a scene video is actually discriminative in terms of *success* and *failure* in many cases. Furthermore, the visual impression from watching these videos suggest that the temporal dynamics seem to be especially important. Our method considers the temporal dynamics using dynamic time warping (DTW) [39], utilizing a k-NN-based classification scheme to mitigate the relatively small training set. This is motivated by comparative studies that showed a very good performance for NN-classifiers and the DTW-distance on various time series datasets [40], [41]. Our method involves the following major steps, which are explained in the subsequent sections:

1) For all possible subsequences (within a certain range of length) of all videos of the given training data, a "discriminativity"-value is computed. This value is high for subsequences that are similar to other subsequences of the same class, but are rather different to any subsequence of the opposite class. Thus, a high discriminativity-value indicates a subsequence with high discriminative power. To account for the temporal nature of the subsequences, dynamic time warping (DTW) [39] is used as distance measure between subsequences. [→Sec. VI-A]
2) From all considered subsequences, a certain number of subsequences with high discriminativity-values is chosen as reference subsequences for each class. [→Sec. VI-B]
3) These reference subsequences are used as prototypes in a k-NN-based classification. [→Sec. VI-C] To take into account the possibly different expressiveness of a person regarding positive and negative FCSs, this classification scheme is extended by introducing a bias that favors one class over the other. [→Sec. VI-D]
4) This classification approach involves several parameters which are optimized on the training data by means of model selection techniques. Therefore, the steps 1. to 3. are iterated over different parameter sets to perform a leave-one-out cross-validation on the training data for parameter optimization. [→Sec. VI-E]

A. Discriminative Subsequence Detection

The goal of the discriminative subsequence detection is to find (comparatively short) video subsequences within the input videos that are characteristic for either *success* or *failure* scenes and can thus be used as prototypical reference subsequences to classify a new scene. Each video is represented as a sequence $A = a_1 a_2 \dots a_N$ of AAM frame parameter vectors a_i of the face, normalized to zero mean and unit variance. In order to find suitable subsequences, an exhaustive search over all possible subsequences of length $l \in [l_{\min}, l_{\max}]$ (in frames) of all training video sequences is performed. For each subsequence of each video, a discriminativity-value $s_{m,i}$ is computed:

$$s_{m,i} = \frac{\sum k\_\min_{n,j}\{d_m^n(i,j) \mid c_m \neq c_n, \quad , j \in P_{m,i}^n\}}{\sum k\_\min_{n,j}\{d_m^n(i,j) \mid c_m = c_n, n \neq m, j \in P_{m,i}^n\}}, \tag{1}$$

[2]The relevant time interval can also be determined automatically in an online system as the starting point is defined by the robot's own prompting and the end point is given by the tutor's direct verbal answer.

where m and i are the indices of the i-th subsequence in the m-th video, $k\_\min\{X\}$ denotes the k smallest values of set X, $d_m^n(i,j)$ is the normalized distance of the i-th subsequence in the m-th video to the j-th subsequence in the n-th video, c_m denotes the class (*success* or *failure*) of the m-th video, and $P_{m,i}^n$ is the index set of all subsequences in the n-th video, the lengths of which are constrained by the length of the i-th subsequence in the m-th video:

$$P_{m,i}^n = \{j \mid \lfloor l_{m,i}/f \rceil \leq l_{n,j} \leq \lfloor l_{m,i} \cdot f \rceil \mid j \in M_n\}, \quad (2)$$

where $l_{m,i}$ is the length (in frames) of the i-th subsequence in the m-th video, M_n is the index set of all subsequences in the n-th video, and $f \geq 1$ is a factor describing the maxium allowed difference in length of two subsequences. This avoids comparison of subsequences of very different lengths and thus prunes the search space for the calculation of $s_{m,i}$. In the experiments described in Sec. VII, $f = 1.3$ was pragmatically chosen, as this value is expected to be a reasonable compromise between evaluating all relevant subsequences and pruning the search space to avoid needless computations. Values significantly higher are not expected to influence the resulting discriminativity-value $s_{m,i}$, as according to Eq. 1 only the k smallest distances are considered, and two subsequences with very different lengths are unlikely to have a small distance to each other, thus it seems safe to drop those comparisons. Nevertheless, a high f-value would substantially increase the computational effort because many irrelevant distances needed to be calculated. On the other hand, f should not be chosen too small to avoid the undesired pruning of some relevant subsequences. The distance $d_m^n(i,j)$ of two subsequences is computed via dynamic time warping (DTW) over the AAM parameter vector sequences. The resulting distance value is normalized by the length $l_{m,i}$ to allow for fair comparison of subsequences of different lengths in Eq. 1. Equation 1 yields high discriminativity-values for subsequences with low minimal distances to subsequences of videos representing the same class (denominator) and high minimal distances to subsequences of videos representing the opposite class (numerator). In other words, the discriminativity-value is high for subsequences that are very similar to other subsequences of the same class and at the same time rather different from even the most similar subsequences of the opposite class. This is similar to the Fisher criterion [42], which minimizes the within scatter while maximizing the between scatter of data from two classes to find an optimal discriminant function. Thus, the higher the discriminativity-value of a subsequence (compared to the values of other subsequences of the given video set), the better it is suited as a representative of the respective class for discrimination purposes. Figure 3 shows an example illustration of the discriminativity-value computation.

B. Reference Subsequence Selection

For each of the two classes, t non-overlapping subsequences with high discriminativity-values are selected as reference subsequences. It might be beneficial for the classification to not select the t subsequences with the t highest discriminativity-values overall, but to preferably select v subsequences per video, for the following reason: If a small number of videos of one class c is very similar to each other and also rather different to any video of the other class, the major part of the t subsequences with highest discriminativity-values overall might stem from these few videos. A larger number of videos of class c might be typical for this class as well, but not that similar to the aforementioned small group of videos. This larger group would be underrepresented by the reference subsequence selection. Thus, the resulting classifier would be able to classify videos similar to the small group very confidently, but would probably perform poor for videos similar to the larger group. To avoid this problem, a more uniform distribution of reference subsequences over the training videos is required.

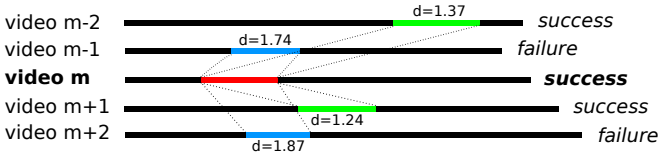

Fig. 3. Example depiction of the discriminativity-value computation. It illustrates the computation of $s_{m,i}$ for the i-th subsequence (shown in red) of the m-th video for $k = 2$. Concerning the videos of the same class as the m-th video, *success*, the two subsequences of the $(m+1)$-th and $(m-2)$-th video (shown in green) are found to have minimal distances (1.24 resp. 1.37) to the target subsequence. Similarly, the two subsequences of the $(m-1)$-th and $(m+2)$-th video (shown in blue) have the minimal distances (1.74 resp. 1.87) of the subsequences from all videos of the opposite class, *failure*. Hence $s_{m,i} = \frac{1.74+1.87}{1.24+1.37}$ according to Eq. 1. Please refer to Sec. VI-A.

This motivates the following selection method. For each video of class c, the v (non-overlapping) subsequences with the highest discriminativity-values are determined and collected in a set S_c. The t most discriminative ones of these subsequences are chosen as reference subsequences of class c. In case S_c contains less than t elements, the missing reference subsequences are taken from the best remaining (non-overlapping) subsequences of all training videos that are not part of S_c. In the following, the index set of the reference subsequences of class c is denoted by R_c.

C. Nearest-Neighbor-based Classification

For classification of a test video sequence (index m) the minimum distance $d_{m,(n,j)}^*$ of every reference subsequence (index (n,j)) to all subsequences (index i) of the test video is computed using a similar pruning condition for the involved subsequence lengths as in Eq. 2:

$$d_{m,(n,j)}^* = \min\{d_m^n(i,j) \mid i \in M_m\}, \quad (3)$$

where $(n,j) \in R_{\text{success}} \cup R_{\text{failure}}$. For each class c, the u best distances are combined to get a classification score $d_{m,c}$:

$$d_{m,c} = \sum_{\gamma \in \Gamma} \frac{1}{\gamma^w} \quad , \Gamma = u\_\min\{d_{m,(n,j)}^* \mid (n,j) \in R_c\}, \quad (4)$$

where the parameter w weights the influence of large distances compared to small ones. The test video sequence is

Parameter / Description	Grid Values
$[l_{\min}, l_{\max}]$: considered subsequence lengths [Sec. VI-A]	[5,5], [10,10], [15,15], [5,20]
k: # distances for subseq. scores [Eq. 1]	1, 2, 5, 10, 15
t: # ref. subseq. in total [Sec. VI-B]	1, 2, 5, 10, 15, 20, 25
v: # ref. subseq. per video [Sec. VI-B]	0, 1, 2
u: # distances for class. scores [Eq. 4]	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
w: distance weight [Eq. 4]	1, 2
b: classification bias [Sec. VI-D]	1.0, b^*

TABLE I
OVERVIEW OF ALL PARAMETERS [→ SEC. VI-E AND SEC. VII]

classified into the class with the highest classification score. This is a k-NN-based classification, as the best distances to a certain number of reference subsequences are combined to form the final classification. [3]

D. Biased Classification

The degree of expressiveness of positive compared to negative valence might vary considerably, depending on the individual characteristics of a person. While some people display both with approximately the same expressiveness, others show a clear bias, meaning that the absence of failure signs could reasonably be interpreted as success, or vice versa. Thus, we introduce a bias b on the classification scores:

$$d'_{m,\text{success}} = d_{m,\text{success}} \quad , \quad d'_{m,\text{failure}} = b \cdot d_{m,\text{failure}}, \qquad (5)$$

where d'_c is the new classification score for class c. During the training, a value $b = b^*$ is chosen such that the training error is minimized. The number of candidate values for b is linear in the number of training videos, because a change of the classification result for a training video p only occurs at a value b given by $d_{p,\text{success}} = b \cdot d_{p,\text{failure}}$. Thus, there is an optimal range of b-values, given by the interval borders between certain two of these "change points". We choose b^* as the mean value of this interval, such that b^* has maximum margin to the change points.[4]

E. Parameter Optimization

The presented approach involves several parameters. They are optimized on the training data by means of a grid search over different candidate parameter sets, where a leave-one-out cross-validation is performed for each set to test its suitability: For all possible combinations of parameters, each training video is treated as test data once, whereas all remaining videos are used to train the classifier. Finally, the parameter set yielding the best classification accuracy is selected and used to train the classifier on all training videos. In case of several parameter sets showing the same optimal performance, the set with the highest ψ value is selected:

$$\psi = \frac{\sum_{r_m = c_m} |d'_{m,\text{success}} - d'_{m,\text{failure}}|}{\sum_{r_m \neq c_m} |d'_{m,\text{success}} - d'_{m,\text{failure}}|}, \qquad (6)$$

where r_m is the classification result for the m-th video. This auxiliary value ψ is high for correctly classified videos with a high difference in classification scores ("confidently correct") and for misclassified videos with a low difference in classification scores ("near miss"). Thus, this parameter selection tries to improve generalization.

A complete list of all parameters together with their values used in the grid search in the experiments described in Sec. VII is given in Tab. I. Parameters that influence the training are listed in the upper block, those only affecting the classification of test data in the lower one.

F. Implementation

The training of our approach involves an exhaustive search for subsequences with high discriminativity-values and a thorough parameter optimization and is thus very demanding. For its practical usage, several optimizations (distance precomputations, efficient subsequence indexing, etc.) are beneficial, which are beyond the scope of this paper. Considering these optimizations, the runtime is $O(N^2 \cdot L^4)$, where N is the number of training videos and L the average length of a training video in frames. While the training is to be performed offline, the classification of a test video can be done online (for typical video lengths) and requires a runtime of $O(N_r \cdot L_r^2 \cdot L_t)$, where N_r is the number of reference subsequences, L_r the maximum length of the reference subsequences, and L_t the length of the test video.

VII. EVALUATION

This section presents an evaluation of our approach on the database introduced in Sec. IV. We performed the classification on each person separately in a leave-one-out cross-validation manner, i.e. each video of the respective person was chosen as test data once, whereas all remaining videos were used as training data. The training and classification for each scene was performed as described in Sec. VI. Table I shows the used grid values for the parameter optimization. The achieved classification accuracies are listed in Tab. II.

To provide a baseline, the first row of Tab. II summarizes the human recognition performance on this task, which we evaluated in previous work [29]. These results show that the interpretation of the FCSs in the object-teaching scenes in terms of valence is a difficult classification problem, as the human classification accuracy was only 82.0% on average (78.1% for *success* scenes and 86.0% for *failure* scenes), which is comparatively low for a two-class problem. The variance is very high in each case, which reflects the large differences in the facial displays of the people in the database. Further details about this can be found in [29].

A. Results Using Individual Models

The second row of Tab. II ("I-AAM: per-scene-optimization") shows the classification accuracies using individual AAMs as discussed in Sec. V. The achieved average

[3] The k-NN-based classification has been shown to outperform a simple NN-based classification ($u = 1$). The use of regression techniques would also be possible, but need an alternative DTW-like distance because DTW does not obey the triangular inequality.

[4] Although unlikely, it is also possible that the optimal b-value lies below the first resp. above the last chance point. In this case, b^* is chosen as a value slightly smaller resp. larger than this change point.

Experiment / Parameter Selection		Sec.	01	02	03	04	05	06	07	08	09	10	11	Mean	SD
Human performance	– all scenes	VII	82	75	85	92	68	73	94	67	78	95	92	82.0	19.1
	– only success		91	66	84	89	61	70	91	52	66	95	93	78.1	21.2
	– only failure		73	84	86	95	75	75	98	82	91	95	91	86.0	16.1
I-AAM: per-scene-optimization	– all scenes	VII-A	94	83	82	88	83	89	89	65	67	74	88	81.9	9.4
	– only success		86	82	95	88	82	90	96	58	37	83	87	80.2	17.7
	– only failure		100	83	64	88	85	88	84	73	83	64	89	81.7	10.9
G-AAM: per-scene-optimization	– all scenes	VII-B	73	90	68	73	63	46	80	76	54	83	68	70.3	12.7
	– only success		80	94	75	76	63	43	83	80	42	75	58	69.9	16.7
	– only failure		67	83	56	69	63	50	77	72	61	92	75	69.4	12.1
G-CLM: per-scene-optimization	– all scenes	VII-B	61	66	62	83	66	71	59	69	66	63	88	68.4	9.4
	– only success		60	71	72	89	56	67	52	73	50	58	88	66.9	13.3
	– only failure		61	58	44	76	75	77	65	64	74	67	89	68.2	11.8
SVM: per-scene-optimization	– all scenes	VII-C	76	83	80	95	84	57	62	74	66	71	88	76.0	11.5
	– only success		67	82	89	90	81	60	52	69	25	75	83	70.3	19.2
	– only failure		83	83	67	100	88	54	70	81	87	67	91	79.2	13.3
I-AAM: median-over-scenes	– all scenes	VII-D	97	83	91	91	92	83	89	71	82	87	90	86.8	6.9
	– only success		93	83	95	100	100	80	96	62	64	83	87	85.6	13.3
	– only failure		100	83	86	81	85	88	84	82	91	91	91	87.4	5.6
I-AAM: median-over-persons	– all scenes	VII-D	75	69	74	84	71	78	71	75	56	65	78	72.3	7.5
	– only success		93	59	75	69	45	100	71	65	100	100	83	78.2	18.5
	– only failure		61	83	71	100	92	50	71	86	35	27	74	68.4	23.2

TABLE II

RESULTS OF THE EXPERIMENTS. THE COLUMNS SHOW THE EXPERIMENT, THE SECTION REPORTING ABOUT THE RESPECTIVE DETAILS, THE CLASSIFICATION ACCURACY (IN PERCENTAGE) FOR EACH PERSON, AND THE MEAN AND STANDARD DEVIATION OF THESE RESULTS.

classification accuracy of 81.9% (80.2% for *success* and 81.7% for *failure*) is comparable to the average human recognition performance. However, there is no significant correlation between the respective classification accuracies of the individual persons of the humans and the automatic classification (Spearman correlation, $\rho \approx 0.24$, $p > 0.47$). Fig. 4 depicts example images taken from the most discrimative reference subsequence of some people. The selected reference subsequences comprise all three kinds of FCSs (head gestures, eye gaze, facial expressions), depending on the individual characteristics of the respective person.

B. Results Using Generic Models

We also evaluated our approach with generic AAMs, where each person's face is fitted by an AAM that was trained without images of this person, using only face images of the remaining people in the database. Unfortunately, this notably impaired the results, as row "G-AAM: per-scene-optimization" of Tab. II shows: the average classification accuracy decreased to 70.3%. This is in line with the results of Gross *et al.* [37]. For comparison, we repeated the experiments replacing the generic AAMs with generic CLMs [17], using Saragih's implementation [17] and its pretrained model, which is considered one of the best state of the art frameworks for generic face fitting. However, the results were comparable (row "G-CLM: per-scene-optimization" of Tab. II), yielding an average classification accuracy of 68.4%. Thus, a very precise facial feature fitting is essential in our scenario. Although most of the generic fitting results did not look that poor visually, the difference to the individual models apparently is vital.

C. Comparison to Previous Results

In previous work [33], we evaluated the classification performance of a SVM classification of the AAM feature vectors, neglecting any temporal dynamics. These previous results are shown in Tab. II, row "SVM: per-scene-optimization". The classification based on reference subsequences outperformed the SVM classification in terms of average classification accuracy. However, the results of the two approaches for individual persons are very different and do not appear to be correlated (Spearman correlation, $\rho \approx 0.005$, $p > 0.98$). A comparison to other results for automatic visual valence recognition is difficult, because the overall setting and evaluation procedure and the characteristics of the video database vary greatly between different studies. The following average classification accuracies were reported in the literature: 55% [43], 62% [22], 67% [19], 77% [18], 78% [21], and 84% [20]. However, due to the significant differences between the studies, these results are not directly comparable, neither to each other nor to our results.

D. Parameter Stability

The leave-one-out training and test procedure naturally results in an individual parameter set for each classification. For the practical usage in a classification system, a certain stability of these parameters is required, as a classifier trained with one specific parameter set is expected to give reasonable results on various test data.

In order to estimate the parameter stability of the classification approach, we computed, for each person separately, a single parameter set that consists of the median values of the parameters resulting from the per-scene-optimization (except for parameter b, where the geometric mean was used instead of the median). The reasoning behind this is that, if a sufficient stability is present, the slightly different training data sets in the leave-one-out cross-validation classifications of the single scenes should yield slightly different parameter sets, which on average capture some characteristics of the

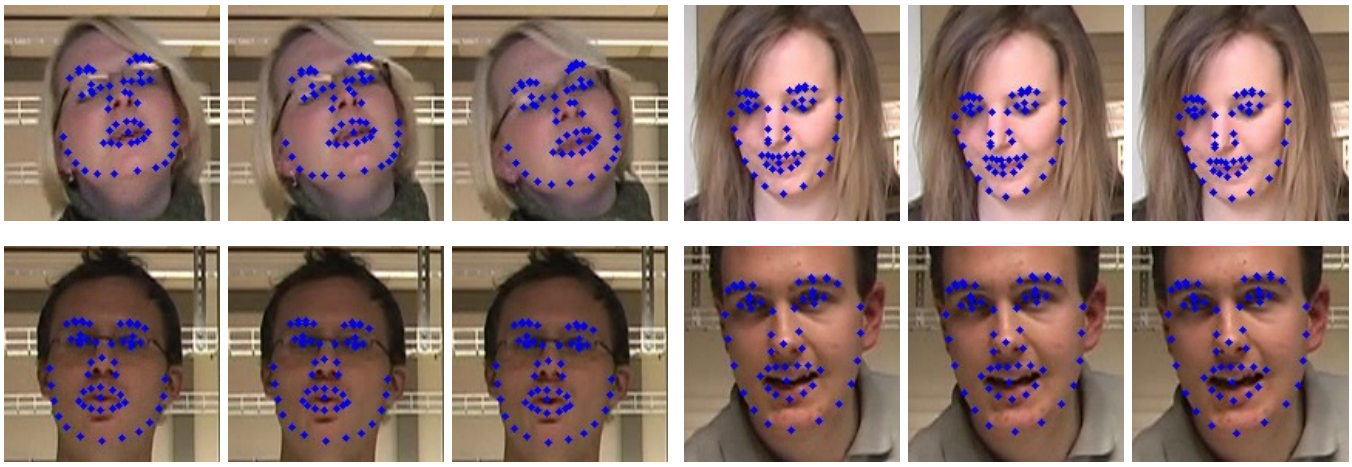

Fig. 4. Example images from the selected reference subsequences. Top: signaling *success* via head gestures (left) and gaze direction (right). Bottom: signaling *failure* via facial expressions. In each case, the first, middle, and last image of a reference subsequence is shown. Please refer to Sec. VI and VII.

respective person. Thus, taking the median value of each parameter should be a good guess for a single parameter set that yield good results for all scenes.

The classification accuracies resulting from a training with this median parameter set are shown in Tab. II, row "I-AAM: median-over-scenes". Compared to the "per-scene-optimization" results, the classification accuracy improved for almost all persons. However, these numbers are not meant to be taken for the evaluation of the classifier in terms of classification accuracy (for which the "per-scene-optimization" results are determinative). As the median operation is performed on the parameter sets of *all* scenes, it also processes information extracted from the respective test data, which is a likely reason for the performance improvement. The point here is that a single parameter set with plausible values (median values, see argumentation above) yielded a reasonable good performance for all scenes of a person. This is an indication that stable parameters exist for each person.

An important question is whether a single stable parameter set can also be selected for all persons. The partially large differences in the characteristics of the different persons let us doubt this. This negative expectation is confirmend by a tentative experiment where we computed again the median values of all the median parameter sets, resulting in a single parameter set for all persons. This parameter selection impairs the classification results notably, as the row "I-AAM: median-over-persons" in Tab. II shows. Thus, suitable parameters of the classifier are person-specific and do not generalize well to other persons.

VIII. Conclusion

We presented an approach for the interpretation of facial communicative signals (FCSs) in terms of valence by discriminative reference subsequence selection. In contrast to most related works, we defined the ground truth labels in terms of the interaction situation instead of the visual appearance of the face. We evaluated this approach on a database containing human-robot interaction videos in an object-teaching scenario. In the reported experiments, an average classification accuracy of 81.9% was achieved for a person-dependent classification with individual models, which is comparable to the human performance of 82.0% and outperforms our previous results based on a SVM classification.[5] Likewise to the human classification, the variance between different persons was very high.

We showed that stable classifier parameters can be found for each person in the database, but these parameters are person-specific, which is natural due to the large variations regarding the display of FCSs between different people. These large variations are a major challenge for a person-independent classification that is a main target of future work.[6] Further aiming at this target, we will evaluate new methods (e.g. [44]) that facilitate generic tracking of facial features more robustly in the context of FCS recognition.

Furthermore, possibilities to speed up the training, for instance by means of a more sophisticated search space pruning, shall be investigated.[7] Future work shall also evaluate other, possibly more sophisticated approaches for the reference subsequence selection, for instance by modifying the discriminativity-score to explicitly take into account the expected number of matches for a candidate subsequence in later classifications, based on statistics of the training data.

IX. Acknowledgements

Christian Lang gratefully acknowledges the financial support from Honda Research Institute Europe for the project "Facial Expressions in Communication". The authors thank Modesto Castrillón-Santana for his face detection software and Jason Saragih for the possibility to use his CLM code.

[5]However, it is not clear to which degree the humans performed a person-dependent or -independent classification. At least some adaptation to the shown people took place in the course of the experiment.

[6]First tentative generalization experiments yielded an average classification accuracy of only 64.0% in the best case.

[7]Due to the continuous, multivariate feature vector data and the traits of the DTW-distance, many typical pruning strategies of other time series recognition methods cannot be applied, because they usually require sequences of ordinal, univariate items (alphabet) or use properties of the Euclidean distance that are not fulfilled by DTW.

REFERENCES

[1] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," Microsoft Research, Tech. Rep. MSR-TR-2010-66, 2010.

[2] E. Murphy-Chutorian and M. M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[3] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 4–24, 2005.

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

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

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[7] C. Lang, S. Wachsmuth, M. Hanheide, and H. Wersing, "Facial Communicative Signals - Valence Recognition in Task-Oriented Human-Robot Interaction," *Journal of Social Robotics - Special Issue on Measuring Human-Robot Interaction*, vol. 4, no. 3, pp. 249–262, 2012.

[8] L. Ye and E. Keogh, "Time series shapelets: a novel technique that allows accurate, interpretable and fast classification," *Data Mining and Knowledge Discovery*, vol. 22, no. 1–2, pp. 149–182, 2011.

[9] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative Subsequence Mining for Action Classification," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[10] 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.

[11] A. Tiwari, R. Gupta, and D. Agrawal, "A Survey on Frequent Pattern Mining: Current Status and Challenging Issues," *Information Technology Journal*, vol. 9, no. 7, pp. 1278–1293, 2010.

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[13] G. Edwards, T. Cootes, and C. Taylor, "Face Recognition Using Active Appearance Models," in *Proceeding of the European Conference on Computer Vision*, H. Burkhardt and B. Neumann, Eds., vol. 2. Springer, 1998, pp. 581–695.

[14] S. Baker, I. Matthews, R. Xiao, J. Gross, T. Kanade, and T. Ishikawa, "Real-Time Non-Rigid Driver Head Tracking for Driver Mental State Estimation," in *11th World Congress on Intelligent Transportation Systems*, 2004.

[15] A. Rabie, B. Wrede, T. Vogt, and M. Hanheide, "Evaluation and Discussion of Multi-modal Emotion Recognition," in *Second International Conference on Computer and Electrical Engineering*, vol. 1, 2009, pp. 598–602.

[16] D. Cristinacce and T. Cootes, "Feature Detection and Tracking with Constrained Local Models," in *Proceedings of the British Machine Vision Conference*, vol. 3, 2006, pp. 929–938.

[17] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable Model Fitting by Regularized Landmark Mean-Shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[18] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[19] 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*, 2006, pp. 146–154.

[20] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. Huang, "Audio-visual affect recognition in activation-evaluation space," in *International Conference on Multimedia and Expo*, 2005.

[21] S. V. Ioannou, A. T. Raouzaiou, V. A. Tzouvaras, K. C. Mailis, Theofilos P.and Karpouzis, and S. D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network," *Neural Networks*, vol. 18, no. 4, pp. 423–435, 2005.

[22] D. McDuff, R. Kaliouby, K. Kassam, and R. Picard, "Affect Valence Inference From Facial Action Unit Spectrograms," in *Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 17–24.

[23] M. F. Valstar, H. Gunes, and M. Pantic, "How to Distinguish Posed from Spontaneous Smiles using Geometric Features," in *International Conference on Multimodal Interfaces*, 2007, pp. 38–45.

[24] 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.

[25] 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.

[26] P. Ekman, "Universals and Cultural Differences in Facial Expressions of Emotion," *Nebraska Symposium on Motivation*, vol. 19, pp. 207–283, 1971.

[27] J. A. Russell, "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, no. 1, pp. 102–141, 1994.

[28] Y.-l. Tian, T. Kanade, and J. F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, February 2001.

[29] 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)*, 2009, pp. 189–194.

[30] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas, "Facial Expression Recognition Based on Dynamic Binary Patterns," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[31] M. T. Motley and C. T. Camden, "Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting," *Western Journal of Speech Communication*, vol. 52, no. 1, pp. 1–22, 1988.

[32] A. J. Fridlund, *Human facial expression: An evolutionary view*. San Diego, CA: Academic Press, 1994.

[33] C. Lang, S. Wachsmuth, H. Wersing, and M. Hanheide, "Facial Expressions as Feedback Cue in Human-Robot Interaction - a Comparison between Human and Automatic Recognition Performances," in *Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, 2010, pp. 79–85.

[34] 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.

[35] H. Gunes and M. Pantic, "Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners," in *International Conference on Intelligent Virtual Agents*, 2010, pp. 371–377.

[36] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "ENCARA2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, vol. 18, no. 2, pp. 130–140, 2007.

[37] R. Gross, I. Matthews, and S. Baker, "Generic vs. Person Specific Active Appearance Models," *Image and Vision Computing*, vol. 23, no. 12, pp. 1080–1093, 2005.

[38] A. Rabie, C. Lang, M. Hanheide, M. Castrillón-Santana, and G. Sagerer, "Automatic Initialization for Facial Analysis in Interactive Robotics," in *Proc. of the Int. Conf. on Computer Vision Systems*. Santorini, Greece: Springer, May 2008, pp. 517–526.

[39] S. Sakoe, H.; Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[40] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast Time Series Classification using Numerosity Reduction," in *23rd Int. Conf. on Machine Learning*, 2006, pp. 1033–1040.

[41] H. Ding, G. Trajcevski, P. Scheuermann, and E. Wang, Xiaoyue Keogh, "Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures," *Proceedings of the VLDB Endowment (PVLDB)*, vol. 1, no. 2, pp. 1542–1552, 2008.

[42] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[43] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling," in *Interspeech*, 2010.

[44] G. Tzimiropoulos, J. Alabort, S. Zafeiriou, and M. Pantic, "Generic Active Appearance Models Revisited," in *Proc. Asian Conf. on Computer Vision (ACCV)*, Daejeon, Korea, Nov. 2012.