Bielefeld University
Faculty of Technology
Research Institute for Cognition and Robotics / Applied Informatics

Technical Report

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

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

February 2012

[1] Research Institute for Cognition and Robotics, Bielefeld University, Germany
[2] Applied Informatics, Bielefeld University, Germany
[3] School of Computer Science, University of Lincoln, UK
[4] Honda Research Institute Europe, Offenbach, Germany

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. 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 rates are comparable to the average human recognition performance and outperformed our previous results on this task.

1 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, besides the understanding of speech, also 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 classification method. The following Sec. 2 briefly introduces related work about FCS recognition in general. Sec. 3 describes the motivation for our valence-based approach to FCS recognition, then the scenario and video database used for its evaluation is introduced in Sec. 4. The utilized face detection and feature extraction techniques are addressed in Sec. 5. Subsequently, the main contribution of this paper—the recognition approach based on reference subsequence selection—is explained in Sec. 6 and evaluated in Sec. 7. Finally, Sec. 8 concludes and remarks on future work.

2 Related Work

A large number of visual head pose estimation techniques are reviewed in the comprehensive survey of Murphy-Chutorian and Trivedi [30]. Among the state of the art approaches are the work of Wang and Sung [46] using geometric relations of facial features in an EM-framework, the utilization of active appearance models (AAMs) [8] for non-rigid head tracking as investigated by Baker *et al.* [1], and the 3D head tracking method method of Zhao *et al.* [50] based on SIFT features [27].

Morimoto and Mimica [29] presented a review of several eye tracking approaches. Wang and Sung's [45] system evaluated geometric relations of iris and eye corners to robustly estimate the eye gaze, whereas Newman *et al.* [31] used template matching. Ishikawa *et al.* [20] and also Ivan [21] utilized AAMs for eye gaze estimation, Varchmin *et al.* [44] combined eigeneye analysis and neural networks for that purpose.

Fasel and Luettin [12] and also Pantic and Rothkrantz [33] presented surveys on facial expression recognition techniques. A huge body of research investigated an interpretation in terms of discrete classes, most often basic emotions [10]. The applied methods include haar-like features and dynamic binary patterns [47], local facial feature deformations [39], and also flexible models [26] and the related AAMs [9, 36]. Buenaposada *et al.* [4] built linear subspace deformation and illumination models and used a k-nearest-neighbor-based classifier with a certain temporal history for the classification of basic emotions.

Many approaches (e.g. [43, 28]) investigated the recognition of action units (AUs) [11], for instance by means of gabor filters and support vector machines (SVMs) [3] or geometric facial feature modeling and neural networks [41]. Relatively few works considered the interpretation of facial expressions in terms of emotional dimensions. Fragopanagos and Taylor [15] and

also Caridakis *et al.* [5] investigated valence and activation recognition with neural networks, whereas Gunes and Pantic [19] used hidden markov models (HMMs) and SVMs for the continuous prediction of valence and four other dimensions.

The importance of temporal dynamics for facial analyses was already noticed very early [17] and has often been addressed by various bayesian approaches [49, 40] or related graphical models such as HMMs [7].

While most early works considered posed facial expressions, there is a growing interest in authentic, spontaneous facial expressions nowadays [43, 39, 3]. Zeng *et al.* [48] presented a comprehensive survey on this topic.

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.

3 Motivation

The way people use FCSs in human-human interaction has been investigated in a vast amount of psychological research; please see [24] for a discussion. Due to the complex and in 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 [24]. 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 [10] have been used as classification categories, due to their universality (although this is controversial [37]). 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 [41, 5, 23]. As a consequence, many works used posed facial expressions (e.g. [4, 47]), which are quite different from authentic, spontaneous ones (e.g. [43]).

On the contrary, facial expressions that carry some communicative semantics as proposed by Fridlund [16] 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 [16] 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 interprete in 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

[1] In a pre-study of previous work [23], 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.

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. 4), 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 [25], 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 ("angry", "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, their is no guarantee that a meaningful FCS is displayed at all, however, studies [2, 23] 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 [15, 5, 19], 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.* [2], 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 [23] for a comparison of these studies to our object-teaching study, which is briefly introduced in the following section.

4 Scenario and Video Database

For the evaluation of our approach, we used the object-teaching scenario introduced in previous work [23]: A person teaches the names of several objects to a robot, which is expected to term them correctly afterwards (please see Fig. 1(a) for a scenario overview). In its verbal answer, the robot can say the correct or a wrong object name. The facial display of the human teacher 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 FCSs the teachers showed during the interaction are authentic and spontaneous, as the participants did not know beforehand that a Wizard of Oz study was performend and that FCSs are important at all, but assumed 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 [23].

The interpretation of the FCSs in the object-teaching scenes in terms of valence turned out to be a difficult classification problem, as the human recognition performance was only 82.0% on

(a) scenario overview

(b) object-teaching scene

(c) examples of facial expressions

Figure 1: Example images from the used object-teaching video database. Please refer to Sec. 4.

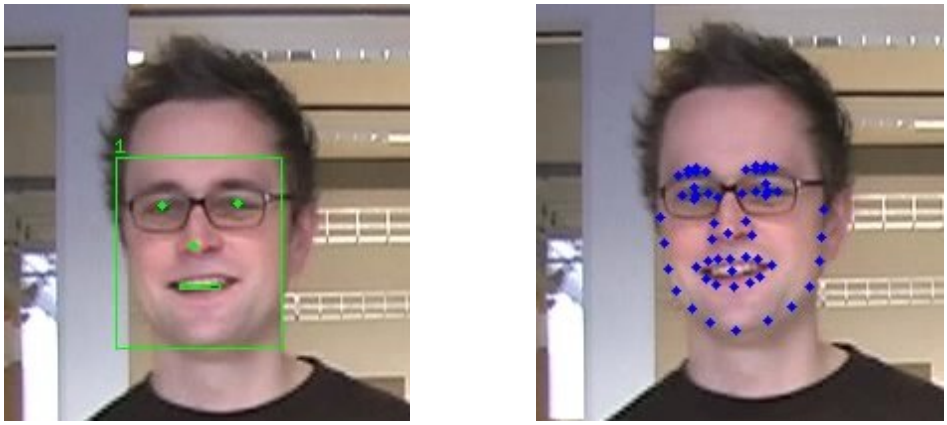

Figure 2: Example for the utilized face detection and feature extraction methods. Please refer to Sec. 5.

average (78.1% for *success* scenes and 86.0% for *failure* scenes, with a high variance in each case), which is comparatively low for a two-class problem. Further details about this can be found in [23].

5 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.* [6] was applied. It succeeded for 98% of the scenes, the remaining 2% were rejected by the system due to too poor face detection and were thus excluded from the experiments described in Sec. 7. The feature extraction on frame level is done by an active appearance model (AAM) [8]. For each human teacher, 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 [18]. 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.* [35]. 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. Fig. 2 shows an example of the face detection and feature extraction.

6 Reference Subsequence Selection For Facial Communicative Signal Classification

An evaluation of previous recognition results using a SVM classification [25] revealed that apparently only a short subsequence of a scene video is actually discriminative in terms of *success* and *failure* in many cases, although the videos are already segmented to contain only the rele-

vant interaction part, i.e. the reaction of the teacher to the answer of the robot.[2] Furthermore, the visual impression from watching these videos suggest that the temporal dynamics seem to be especially important.

This motivates the search for comparatively short video subsequences with high discriminative power und their usage as prototypical representatives for the two classes in a classification approach considering temporal dynamics. A subsequence is a set of several consecutive frames of a video, where each frame is represented by the corresponding AAM parameter vector. The "discriminative power" of a subsequence refers to its suitability to distinguish *success* from *failure* videos (please see below). The presented method consists of the following major steps, which are explained in detail in the subsequent sections, after a short comparison to some related approaches in Sec. 6.1:

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) [38] is used as distance measure between subsequences. [→Sec. 6.2]

2. From all considered subsequences, a certain number of subsequences with high discriminativity-values is chosen as reference subsequences for each class. [→Sec. 6.3]

3. These reference subsequences are used as prototypes in a nearest-neighbor-based classification. [→Sec. 6.4] 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. 6.5]

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. 6.6]

6.1 Related Approaches

Buenaposada *et al.* [4] also used a nearest-neighbor-based classifier to classify facial expressions. In their method, the prototypes are points in a linear subspace corresponding to a single face image each, where the temporal dynamics are considered at classification level, whereas in our approach the prototypes correspond to sequences of face images (instead of single images) and the DTW distance function accounts for the temporal dynamics. Also the prototype selection method is entirely different.

[2] More concretely, the relevant interaction part starts when the robot starts to utter the object name and ends when the human teacher finishes her or his reaction to the robot. Those intervals were annotated by human coders in the present database [23], but can in principle be determined automatically by exploiting knowledge about the task and the typical turn-taking behavior.

Most works on the detection of specific subsequences in sequential input data stem from datamining research. Tiwari *et al.* [42] presented a survey on methods to find frequently occurring patterns in large datasets. In typical state of the art datamining techniques for discriminative subsequence detection (e.g. [22]) and related pattern matching problems (e.g. [14]), the data are usually sequences of ordinal, univariate items (alphabet). This allows for effective pruning strategies as integral parts of the respective methods where large parts of the generated search trees can be discarded as they cannot contain a desired subsequence. Due to the continuous, multivariate feature vector data and the traits of the DTW distance measure, those pruning strategies are not applicable in our case. Nowozin *et al.* [32] used discriminative subsequences to classify human actions. They used complex features based on gabor filters that are transformed in sequences of sets of integers for a subsequent classification with a boosting algorithm. To perform the pattern search for suitable subsequences, they developed an extension of the PrefixSpan algorithm [34] that relies on an effective search tree pruning, which is not applicable for our features again.

Tiwari *et al.* [42] pointed out that most pattern mining approaches focus on the performant computation of frequent patterns, leaving the quality assessment for a specific use case 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.

6.2 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:

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

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:

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

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 prunes the search space for the calculation of $s_{m,i}$. In the experiments described in Sec. 7, $f = 1.2$ 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. 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 pruning of relevant subsequences.

The distance $d_m^n(i,j)$ of two subsequences is computed via dynamic time warping (DTW) [38] 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 (1).

Equation (1) yields high discriminativity-values for subsequences with low minimum distances to subsequences of videos representing the own class (denominator) and high minimum distances to subsequences of videos representing the other class (numerator). This is similar to the Fisher criterion [13], 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 discriminativity-values of other subsequences of the given video set), the better it is suited as a representative of the respective class.

6.3 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. This motivates the following selection method: S_c is the index set of the best v subsequences for each training video of class c:

$$S_c = \bigcup_{m|c_m=c} \{ \arg\underset{(m,i)}{s_{m,i}} \mid i \in R_v^m \}, \tag{3}$$

where R_v^m is the index set of the v non-overlapping subsequences with the highest discriminativity-values in the m-th video. Further, Q_c contains the indices of all subsequences that are not part of S_C:

$$Q_c = \bigcup_{m|c_m=c} \{(m,i) \mid i \in M_m \ , \ (m,i) \notin S_c\}, \tag{4}$$

The index set R_c of the final reference subsequences for class c is given by a combination of the elements of S_c and Q_c:

$$R_c = \begin{cases} t\_\underset{(m,i)}{\arg\max} \ \{s_{m,i} \mid (m,i) \in S_c\} & \mid \text{if } \|S_c\| \geq t \\ S_c \quad \cup \quad T_c & \mid \text{if } \|S_c\| < t \end{cases} \tag{5}$$

$$\text{with} \quad T_c = (t - \|S_c\|)\_\underset{(m,i)}{\arg\max} \ \{s_{m,i} \mid (m,i) \in Q_c\}, \tag{6}$$

where $k\_\arg\max\{X\}$ denotes the arguments associated with the k largest values of the set X.

6.4 Nearest-Neighbor-based Classification

The classification of a test video sequence (index m) starts with the computation of the minimum distance $d^*_{m,(n,j)}$ of every reference subsequence (index (n,j)) to all subsequences (index i) of the test video, considering a similar pruning condition for the involved subsequence lengths as in (2):

$$d^*_{m,(n,j)} = \min \{d^n_m(i,j) \mid i \in M_m\}, \tag{7}$$

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 $d_{m,c}$:

$$d_{m,c} = \sum_{\gamma \in \Gamma} \frac{1}{\gamma^w} \quad , \ \Gamma = u\_\min \{d^*_{m,(n,j)} \mid (n,j) \in R_c\}, \tag{8}$$

where the parameter w weights the influence of large distances compared to small ones. The test video sequence is classified into the class with the highest classification score. This is a k-nearest-neighbor-based classification, as the best distances to a certain number of reference subsequences are combined to form the final classification.

6.5 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 might show a clear bias, meaning that the absence of failure signs could reasonably be interpreted as success, or vice versa. Taking this into account, we introduce a bias b on the classification scores:

$$d'_{m,\text{success}} = d_{m,\text{success}} \quad , \quad d'_{m,\text{failure}} = b \cdot d_{m,\text{failure}}, \tag{9}$$

where d'_c is the new classification score for class c. The value of b is chosen such that the training error is minimized. The candidate values for b are computed as follows: For each classification of a training video (index g), the quotient z_g of the classification scores is calculated:

parameter / description	**grid values**
$(l_{\text{min}}, l_{\text{max}})$: considered subsequence lengths [Sec. 6.2]	(5,5), (10,10), (15,15), (5,20)
k: # distances for subsequence scores [Eq. (1)]	1, 2, 5, 10
t: # reference subsequences in total [Sec. 6.3]	10
v: # reference subsequences per video [Eq. (3)]	0, 1, 2
u: # distances for classification scores [Eq. (8)]	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
w: distance weight [Eq. (8)]	1, 2
b: classification bias [Sec. 6.5]	1.0, b^*

Table 1: Overview of all parameters. Please refer to Sec. 6.6 and Sec. 7.

$$z_g = \frac{d_{m,\text{success}}}{d_{m,\text{failure}}}, \tag{10}$$

As $d'_{g,\text{success}} = d'_{g,\text{failure}}$ holds for $b = z_g$, these z_g values are the points where changes in the classification results of the training data occur when one alters b. Thus, when one sorts all z_g values in ascending order, for any two neighboring values z_{g_1} and z_{g_2}, all selections of b from the interval (z_{g_1}, z_{g_2}) will yield the same classification result, hence only one value $b \in (z_{g_1}, z_{g_2})$ needs to be considered as candidate for the optimization. We choose the mean values of the interval borders in each case (because they have maxium distance to the "change points" for the classification). Together with one value slightly below the minimum z_g value and another value slightly above the maximum z_g value, they constitute the candidate values for the optimization. Finally, we select the value $b = b^*$ that yields the best classification result on the training data. If there are several best values, the median of them is chosen.

6.6 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 rate 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:

$$\psi = \frac{\sum_{r_m = c_m} |d'_{m,\text{success}} - d'_{m,\text{failure}}|}{\sum_{r_m \neq c_m} |d'_{m,\text{success}} - d'_{m,\text{failure}}|}, \tag{11}$$

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. 7 is given in Tab. 1. Parameters that influence the training are

experiment	Sec.	1	2	3	4	5	6	7	8	9	10	11	mean	std.
per-scene-opt:	7													
– all scenes		91	79	83	97	71	67	82	85	62	78	79	79.5	10.2
– only success		86	88	81	94	73	90	79	85	45	83	74	79.8	13.1
– only failure		94	67	86	100	70	38	84	86	70	73	83	77.2	17.0
med-scenes:	7.1													
– all scenes		94	93	83	97	88	83	84	85	68	87	90	86.5	7.7
– only success		93	94	94	94	91	90	80	85	45	92	87	85.8	14.1
– only failure		94	92	71	100	85	75	87	86	78	82	91	85.7	8.6
med-persons:	7.1													
– all scenes		91	79	74	85	75	72	75	60	56	65	86	74.3	10.7
– only success		79	76	65	69	64	70	88	27	100	92	78	73.3	19.1
– only failure		100	83	86	100	85	75	65	100	35	37	91	77.8	23.6
svm:	7.2													
– all scenes		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

Table 2: Results of the experiments. The columns show the experiment, the section reporting about the respective details, the classification rate (in percent) for each person, and the mean and standard deviation of these classification rates.

listed in the upper block, those only affecting the classification of test data in the lower one.

7 Evaluation

This section presents an evaluation of our approach on the database introduced in Sec. 4. 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. 6. Table 1 shows the used grid values for the parameter optimization. The classification results are listed in Tab. 2, row "per-scene-opt". On average, the classification rate was 79.5% in total, 79.8% for *success* and 77.2% for *failure* scenes. This is only slightly below the average human recognition performance on this task (Sec. 4) and better than our previous results (Sec. 7.2). The standard deviation of the classification rates is very high and even systematic misclassifications occur (persons six and nine), both holds for the human performance and the previous results as well [23, 25]. Fig. 3 depicts example images taken from the most discrimative reference subsequence of some people.

7.1 Parameter Stability

The leave-one-out training and test procedure results in an individual parameter set for each classification, which is fine for a performance evaluation of the classifier. For the practical usage

Figure 3: Example images from the selected reference subsequences. Top row: signaling *success* via head gestures (left) and gaze direction (right). Bottom row: signaling *failure* via facial expressions. In each case, the first, middle, and last image of a reference subsequence is shown. Please refer to Sec. 6 and Sec. 7.

in a classification system, a certain stability of these parameters is required, as a classifier trained with one specific parameter set is usually expected to give reasonable results on various test data.

In order to estimate the parameter stability of the presented 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 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 rates resulting from a training with this median parameter set are shown in Tab. 2, row "med-scenes". Compared to the "per-scene-opt" results, the classification rate improved for almost all persons. However, these numbers are not meant to be taken for the evaluation of the classifier in terms of classification rates (for this purpose, the "per-scene-opt" 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 (Tab. 3). 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 between the median parameter sets for different persons (Tab. 3) 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

person	$(l_{\min}, l_{\max})$	k	t	v	u	w	b
1	(5,5)	2	10	2	10	2	1.1837
2	(5,20)	1	10	0	10	1	0.9241
3	(15,15)	10	10	1	1	1	0.8748
4	(5,5)	1	10	0	1	1	0.8245
5	(5,20)	2	10	2	4	1	0.8165
6	(5,5)	5	10	0	8	1	0.9749
7	(5,20)	5	10	1	8	2	1.1378
8	(5,20)	10	10	0	4	2	0.6705
9	(5,10)	5	10	1	3	2	2.0374
10	(10,10)	1	10	0	3	1	1.1012
11	(5,20)	5	10	1	9	1	1.0100
m.o.p.	(5,15)	5	10	1	4	1	1.0075

Table 3: Median parameters for each person (rows 1–11) and median over all persons (last row). Please refer to Sec. 7.1.

notably, as the row "med-persons" in Tab. 2 shows. Thus, suitable parameters of the classifier appear to be person-specific and do not generalize well to other persons.

7.2 Comparison to Previous Results

In previous work [25], we evaluated the classification performance of a SVM classification of AAM feature vectors, neglecting any temporal dynamics. The previous results are shown in Tab. 2, row "svm". The classification based on reference subsequences outperformed the SVM classification in terms of classification rate on average and for seven of the eleven persons. There was no significat correlation between the classification rates for different persons of the two approaches (Spearman test, $r = 0.43, p = 0.19$ for all scenes, $r = 0.27, p = 0.42$ for *success* scenes, and $r = 0.26, p = 0.45$ for *failure* scenes).

8 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 and not by 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 rate of 79.5% was achieved for a person-dependent classification, which is comparable to the human performance of 82.0% and outperforms our previous results based on a SVM classification.

We showed that stable classifier parameters can be found for each person in the database. However, these parameters are person-specific and do not generalize well to new persons, which is natural to some degree due to the large variations regarding the display of FCSs between

different people. This is a major challenge for a person-independent classification that is a main target of future work.

As the training of the presented classifier involves an exhaustive search over candidate subsequences, it is very time-consuming. Possibilities to speed up this search, for instance by means of a more sophisticated pruning and concentration on the most relevant parts of the search space, shall be investigated. Future work could 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 the statistics of the training data.

9 Acknowledgements

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

References

[1] 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.

[2] P. Barkhuysen, E. Krahmer, and M. Swerts. Problem Detection in Human-Machine Interactions based on Facial Expressions of Users. *Speech communication*, 45(3):343–359, 2005.

[3] 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.

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

[5] 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.

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

[7] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. S. Huang. Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. *Computer Vision and Image Understanding*, 91(1–2):160–187, July 2003.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[9] G. Edwards, T. Cootes, and C. Taylor. Face Recognition Using Active Appearance Models. In H. Burkhardt and B. Neumann, editors, *Proceeding of the European Conference on Computer Vision*, volume 2, pages 581–695. Springer, 1998.

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

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

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

[13] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179–188, 1936.

[14] A. Floratou, S. Tata, and J. M. Patel. Efficient and Accurate Discovery of Patterns in Sequence Data Sets. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1154–1168, 2011.

[15] N. Fragopanagos and J. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005.

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

[17] N. H. Frijda. An understanding of facial expression of emotion. *Acta Psychologica*, 9:294–362, 1953.

[18] R. Gross, I. Matthews, and S. Baker. Generic vs. Person Specific Active Appearance Models. *Image and Vision Computing*, 23(12):1080–1093, 2005.

[19] 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*, pages 371–377, 2010.

[20] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive Driver Gaze Tracking with Active Appearance Models. In *11th World Congress on Intelligent Transportation Systems*, 2004.

[21] P. Ivan. Active Appearance Models for Gaze Estimation. Master's thesis, Vrije Universiteit Amsterdam, Faculty of Sciences, Business Mathematics & Informatics, 2007.

[22] X. Ji, J. Bailey, and G. Dong. Mining Minimal Distinguishing Subsequence Patterns with Gap Constraints. *Knowledge and Information Systems*, 11(3):259–286, 2007.

[23] 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.

[24] C. Lang, S. Wachsmuth, M. Hanheide, and H. Wersing. Facial Communicative Signals - Valence Recognition in Task-Oriented Human-Robot Interaction. *International Journal of Social Robotics - Special Issue on Measuring Human-Robot Interaction*, to appear, 2012.

[25] 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 *Third IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, pages 79–85, 2010.

[26] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic Interpretation and Coding of Face Images Using Flexible Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[27] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[28] S. Lucey, A. B. Ashraf, and J. F. Cohn. *Face Recognition*, chapter Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face, pages 275–286. I-TECH Education and Publishing, 2007.

[29] C. H. Morimoto and M. R. Mimica. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4–24, 2005.

[30] E. Murphy-Chutorian and M. M. Trivedi. Head Pose Estimation in Computer Vision: A Survey. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[31] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. Real-Time Stereo Tracking for Head Pose and Gaze Estimation. In *International Conference on Automatic Face and Gesture Recognition*, pages 122–128, 2000.

[32] S. Nowozin, G. Bakir, and K. Tsuda. Discriminative Subsequence Mining for Action Classification. In *International Conference on Computer Vision*, pages 1–8, 2007.

[33] 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.

[34] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, 2004.

[35] 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.

[36] 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*, volume 1, pages 598–602, 2009.

[37] J. A. Russell. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141, 1994.

[38] S. Sakoe, H.; Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[39] 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.

[40] C. Shan, S. Gong, and P. McOwan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *Proc. BMVC*, volume 1, pages 297–306. Citeseer, 2006.

[41] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing Action Units for Facial Expression Analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[42] A. Tiwari, R. Gupta, and D. Agrawal. A Survey on Frequent Pattern Mining: Current Status and Challenging Issues. *Information Technology Journal*, 9(7):1278–1293, 2010.

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

[44] A. C. Varchmin, R. Rae, and H. Ritter. Image based recognition of gaze direction using adaptive methods. In *Gesture and sign language in human-computer interaction. International Gesture Workshop*, Bielefeld, Germany, September 1997.

[45] J.-G. Wang and E. Sung. Study on Eye Gaze Estimation. *IEEE Transactions on Systems, Man, and Cybernetics*, 32(3):332–350, 2002.

[46] J.-G. Wang and E. Sung. EM enhancement of 3D head pose estimated by point at infinity. *Image and Vision Computing*, 25:1864–1874, 2007.

[47] 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.

[48] 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.

[49] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):699–714, May 2005.

[50] G. Zhao, L. Chen, J. Song, and G. Chen. Large Head Movement Tracking Using SIFT-Based Registration. In *15th International Conference on Multimedia*, pages 807–810, 2007.