This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

Automatic Pain Assessment with Facial Activity Descriptors

Philipp Werner, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C. Traue

Abstract—Pain is a primary symptom in medicine, and accurate assessment is needed for proper treatment. However, today's pain assessment methods are not sufficiently valid and reliable in many cases. Automatic recognition systems may contribute to overcome this problem by facilitating objective and continuous assessment. In this article we propose a novel feature set for describing facial actions and their dynamics, which we call facial activity descriptors. We apply them to detect pain and estimate the pain intensity. The proposed method outperforms previous state-of-the-art approaches in sequence-level pain classification on both, the BioVid Heat Pain and the UNBC-McMaster Shoulder Pain Expression database. We further discuss major challenges of pain recognition research, benefits of temporal integration, and shortcomings of widely used frame-based pain intensity ground truth.

Index Terms—Automatic pain assessment, pain intensity, facial expression analysis, facial dynamics, recognition, health care.

1 Introduction

Pain is the primary symptom that prompts people to seek medical attention [1]. Uncontrolled pain not only causes suffering and reduces quality of life, but also compromises immune function, promotes tumor growth and can compromise healing after surgery [2]. Further, wrong treatment may lead to problems and risks for the patients [3]. So valid and reliable pain assessment is necessary to choose the adequate treatment.

Pain is a very complex phenomenon, which is not fully understood yet. The International Association for the Study of Pain (IASP) defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage." [4] So pain—as a personal experience—is always subjective. Usually it is caused by a painful stimulus, but the same stimulus may lead to different pain experiences. People differ in their pain sensitivity and various factors may amplify or attenuate their experience, e. g. beliefs, emotions, or situational factors [1], [5], [6]. When the severity of the pain experience exceeds a critical threshold it triggers behavioral reactions that can be observed, most notably facial expression [5], [6]. The intensity and frequency of facial expression increase with the stimulus intensity [7]. However, individuals differ in their expressiveness, i. e. they start to react at different stimulus intensities [5].

The current "gold standard" in pain assessment is self-report, i. e. the patient is asked to quantify the experienced pain. However, self-report cannot be applied for unconscious or newborn patients and is not always reliable and valid, e. g. for demented patients [8]. Further, self-report still may differ from the subjective pain experience, as it is a controlled and goal-oriented response to pain [7], which might be affected by reporting bias and variances in memory and verbal ability [9]. In contrast, facial expression can largely be characterized as a reflexive, automatic reaction to painful experience [7].

Most studies on facial expression are based on the Facial Action Coding System (FACS), which decomposes expressions into elementary action units [10]. Each present action unit is coded with onset, offset and (most of them) with an intensity on a 5-point scale. Numerous studies observed a set of core actions, which tend to occur consistently during acute pain and acute exacerbation of persistent pain [7]. Some authors suggest that there is a prototypical facial expression of pain [11], [12] with some variability in detail. However, a more recent study suggests that there are several "faces of pain" [13]. Apart from the details, pain researchers agree on the existence of facial activity patterns that are associated with pain.

Whereas FACS suits well to study facial expressions, it is too time-consuming for clinical pain assessment, as a trained FACS expert needs about two hours to code one minute of video [14]. Automatic recognition systems are a promising alternative. They can provide a continuous pain assessment, which may improve clinical outcomes, especially by facilitating early intervention for patients that cannot call for help by themselves. Automatic systems may be more objective than a human observer, whose assessment is influenced by own personal factors, the relationship to the sufferer [6] and even by aspects like the patient's attractiveness [9]. Further, these systems may help to gain new knowledge about pain, e. g. about the dynamics of facial expressions [15]. They can be more sensitive to slight changes and grasp more complex dynamics than FACS [16]. There are several efforts to create an automatic system for recognizing pain through computer vision and machine learning techniques. These build on

- *P. Werner and A. Al-Hamadi are with the Institute for Information Technology and Communications, University of Magdeburg, Germany. E-mail: Philipp.Werner@ovgu.de*
- *K. Limbrecht-Ecklundt is with the University Medical Center Hamburg-Eppendorf, Germany.*
- *S. Walter, S. Gruss, and H. C. Traue are with the University Clinic for Psychosomatic Medicine and Psychotherapy, Ulm, Germany.*

Manuscript received 11-Jun-2015, revised 11-Dec-2015, accepted 14-Feb-2016.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

many years of broad research in automatic facial expression recognition [17].

1.1 Datasets

Progress in automatic pain recognition is facilitated by databases for training, testing and comparison. The most widely used dataset is the UNBC-McMaster Shoulder Pain Expression Archive Database [18]. It consists of 200 videos recorded with 25 participants suffering from shoulder pain. The participants underwent a series of active and passive range-of-motion tests with their affected and unaffected limbs. The dataset also includes self-report and observer measures of the pain intensity at video sequence level. Further, at frame level it provides FACS coding and facial landmarks.

The recently published BioVid Heat Pain Database [19], [20], [21] was created in an experimental study with healthy participants who were stimulated with heat to induce pain in four intensities. To compensate for varying thermal pain sensitivities, the temperatures were individually adjusted according to the subject-specific pain threshold (feeling of heat turns into pain) and pain tolerance (pain gets unacceptable). Each of the four pain intensities was stimulated 20 times in randomized order. For each stimulus, the maximum temperature was held for 4 s, alternated with pauses of 8-12 s. The dataset contains videos of 87 participants. Including 20 baseline (no pain) and 4 x 20 pain samples per person, there is a total of 8,700 samples; each with color and depth video of 5.5 s length. The participants were explicitly allowed to move their head freely. The database also contains biomedical signals, see [21] for more details.

1.2 Recognition Approaches

Pain gains more and more attention in the facial expression recognition community. Many works have contributed on pain detection, i. e. on distinguishing pain from no pain [20], [22], [23], [24], [25], [26]. But there is a tendency towards the challenging task of pain intensity assessment [23], [27], [28], [29], [30], [31], [32]. Other authors classify genuine versus faked pain [14], [15] or pain versus basic emotions [33], [34].

Most works are evaluated with the UNBC-McMaster database [22], [23], [24], [25], [28], [29], [30], [31], [32]; some use the BioVid database [20], [26], [35], [36]. Other approaches are tested on non-public data sets [14], [15], [27], [33], [34], which makes it hard to compare with their results or reproduce them.

A variety of features have been used for pain recognition: Gabor features [14], [15], local binary patterns [24], [28], [30], [35], dense SIFT bag of words [25], as well as other generic shape and appearance features [22], [23], [29], [30], [31], [32]. In contrast, some works propose hand-engineered features, namely facial distances [33] which are usually combined with some kind of wrinkle measures to capture some additional changes in appearance [16], [20], [26], [27], [34], [36]. Hand-engineered features usually are lower-dimensional and easier to interpret.

Most approaches classify with Support Vector Machines (SVMs), either with linear [22], [23], [31], [36] or radial basis function kernel [14], [15], [16], [20], [27], [33]. However, several classifiers outperform SVM in the presented experiments: Random Forest [26], Multiple-Instance-Learning with Boosting [25], and Heteroscedastic Conditional Ordinal Random Fields [28], [29]. The latter work and some transfer learning approaches [24], [32] aim at reducing the problem of inter-subject variability, which is highly relevant in facial expression analysis. Some authors use regression techniques to estimate pain intensity in a continuous scale [30], [32].

The majority of contributions is focused on frame-based pain assessment without temporal integration [16], [23], [24], [27], [28], [29], [30], [31], [32], [33]. Several authors see the problems associated with the lack of temporal information. E. g. an eye blink, which is probably not directly related to pain, cannot be distinguished from eye closure, which is clearly related to pain [11]. Another drawback is that these methods rely on frame-level ground truth, which is very expensive—especially if it is FACS-based—but may still contain errors (as in the UNBC-McMaster database [25]). Further, it restricts the system to the prior knowledge used to create the labeling, which may be incomplete or inconsistent; so the opportunity to utilize or discover something new is lost. See Sec. 4.3 for more discussion on that.

Sequence-level pain assessment is based on coarser ground truth, which has been obtained from an observer [22], [23], [25] or from the experimental procedure [14], [15], [20], [26], [35], [36]. It calls for temporal integration of the frame-level information. This can be done by fusing frame-level decisions without incorporating dynamics [22], [23] or by fusing segment-level decisions, which are based on multiple frames [25].

In contrast, we hypothesize that temporal integration should be done at feature level, i. e. before decision, to optimally use dynamics. In this approach, the time series of frame-level features (feature signals) are condensed into descriptors, which are used to classify at sequence level. This technique has been demonstrated with hand-engineered frame-level features and a signal statistic's descriptor in our preliminary work [20], [26]. Other authors propose a two layer architecture in which the first layer is trained to extract facial action unit scores. The second layer classifies the sequence based on either score signal statistics [14], [37], or score signal descriptors, i. e. histograms of filter responses called bags of temporal features [14], [15].

Interestingly, sequence-level ground truth has also been used to learn a model for frame-level pain estimation and localization [22], [25].

Usually, sequence-level methods are applied to get one pain assessment per video, because ground truth is only available in this granularity. However, they can also be applied in a sliding time window to get a continuous prediction similarly to frame-based methods [35]. So we could also call them time window level methods.

Next to facial expression, scientists also work on pain recognition from biomedical signals [38], [39], [40], [41] and fusion of video and biomedical signals [26], [35], [36], but this work focuses on the video modality.

1.3 Contributions

This article discusses the relevance of proper temporal integration in the context of pain recognition, proposes a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

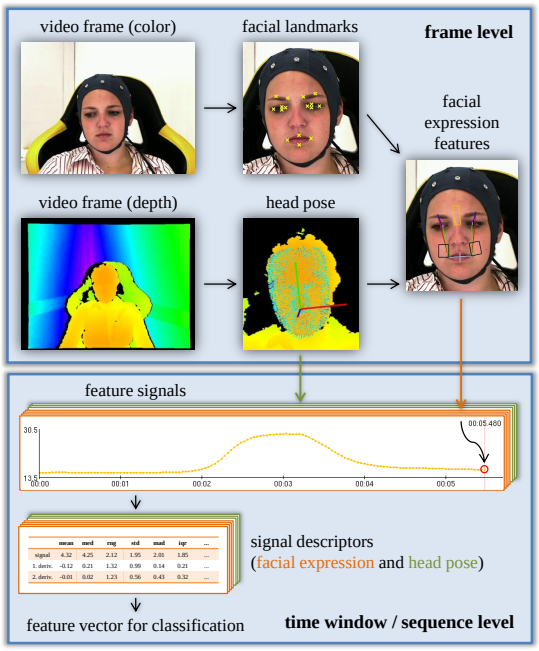


Fig. 1. Overview on the extraction of the facial activity descriptors.

new feature set for temporal integration, and shows its advantages in an experimental comparison. It advances our previous work on video based pain detection [20], [26] with the following key contributions:

- We extend our preliminary descriptor, yielding a new feature set called facial activity descriptors.
- We compare the predictive performance of several temporal integration methods on the task of pain intensity estimation, which is more challenging than pain detection. Experiments are conducted with two publicly available datasets and several types of frame-level features. Our proposed descriptors outperform state-of-the-art methods for sequence-based pain recognition in all experiments.
- We analyze the relevance of descriptor variables and feature signals. Further, we illustrate the variability regarding individuals and intensities, which are major challenges in pain recognition. We also discuss shortcomings of the Prkaching and Solomon Pain Intensity (PSPI) score, which is widely used as ground truth, and give our recommendations.

Sec. 2 introduces the facial activity descriptors with our primarily used frame-level features. Note that frame-level processing can be easily replaced as done for the experiments in Sec. 4.1.4 and Sec. 4.2. Sec. 3 gives a brief overview on the machine learning methods that we use in our experiments, which are reported and discussed in Sec. 4. Sec. 5 concludes the article and outlines future research directions.

2 Facial Activity Descriptors

In this section we describe our method to extract pain relevant features for classification. We propose to use descriptors of facial activity during a time window instead of frame-level information only. The motivation is threefold.

1) Behavioral studies showed that there is a beneficial effect of dynamic information on recognition accuracy in emotion, especially for subtle expressions and when available static information is limited [42]. Assuming that current automatic feature extraction methods can only grasp parts of the facial information that humans can, extracted static features can be regarded as limited. So dynamic information may help to compensate for this weakness [42]. Further, dynamics already proved their utility in assessing the authenticity of smiles [43], [44] and facial expressions of pain [14], [15]. As already mentioned, temporal information is also required to distinguish between eye closure and blinking, which is important for pain recognition.
2) Many aspects of dynamics cannot be modeled with image based classification and late decision-based fusion, but are simple to represent in a time-window descriptor. These aspects include speed, acceleration, smoothness, tendency or overall variation.
3) To the authors best knowledge, there is currently no method to acquire ground truth of self-report pain measures that is valid and reliable *on frame level*. One may argue that there are frame-level observational measures such as the FACS based pain expression score suggested by Prkachin and Solomon [12]. This kind of ground truth is very useful to train and test facial expression recognition systems. However, if we are interested in the underlying feeling of pain, this measure has several shortcomings, which will be discussed in Sec. 4.3. The feeling of pain and its facial expression are strongly related, but through a complex time-dependent process with a lot of influencing factors. This calls for modeling with temporal integration, which we realize through time-window based descriptors.

The method to extract our facial activity descriptors is outlined in Fig. 1. On the frame level, we detect facial landmarks, estimate the head pose and extract facial expression features using both, landmarks and head pose (Sec. 2.1). To obtain feature signals (Sec. 2.2), frame-level features are gathered for a given time period, namely the length of the time window. This way we create a feature signal for each facial expression feature and each head pose parameter. Next, we condense each feature signal in a corresponding signal descriptor (Sec. 2.3) that describes a part of the facial activity. The signal descriptors are concatenated to one feature vector, which is subsequently used for classification (see Sec. 3). The described method is an advancement of our preliminary work [20], [26].

2.1 Frame-Level Processing

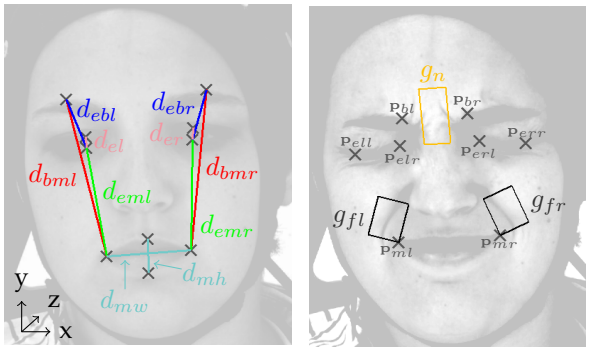
Essentially, the frame-level feature extraction is the same in our preliminary work [26], but we describe it in greater detail here.

Facial Landmarks: Our facial expression analysis is based on landmarks (see Fig. 2a). To detect and track them automatically, we apply the publicly available state-of-the-art software IntraFace [45]. It requires a face detection

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

(a) (b)

Fig. 2. Facial landmarks and head pose estimation. (a) Used subset of the IntraFace landmarks. (b) Measured 3D point cloud, model fitting residuals (cyan) and nose-tip coordinate system illustrating the determined pose.



(a) (b)

Fig. 3. Frame-level features of facial pain expression. (a) 3D distance features. (b) Gradient feature regions and their anchor points.

method, for which we use the Haar-like feature detector cascade by Lienhart et al. [46] (available within OpenCV). IntraFace provides facial feature points that are more accurate and robust than those used in our previous approach [20]. It accounts for less noise in the facial features, which leads to improved recognition rates (see Sec. 4.1.1).

Head Pose: To estimate the head pose, we utilize depth information. For a volume of interest the depth map is converted into a 3D point cloud with a pinhole camera model. Afterwards, a generic face model is registered with the measured point cloud using a variant of the Iterative Closest Point (ICP) algorithm as presented by Niese et al. [47]. It provides a 6D head pose vector including the 3D position and 3D orientation (see Fig. 2b). This information is used as a frame-level feature itself (as we observed several head motion patterns during pain), and to improve pose invariance of facial expression features.

Frame-Level Facial Expression Features: For each image frame we extract several distance and gradient features. They are selected to capture pain related facial actions that have been identified and validated in numerous previous studies, e. g. by Prkachin [11], [12], Craig et al. [48] or Kunz et al. [49]. These actions include lowering of the brows, tightening of the lids, closing of the eyes, raising of the cheeks and the upper lip, wrinkling of the nose, and stretching and opening of the mouth. To uncouple facial expression from head pose and get pose-invariant expression features, distances are calculated in 3D, as proposed by Niese et al. [50]. Using a pinhole camera model, the previously detected landmarks are projected onto the surface of the generic face model placed according to the current head pose. From the obtained 3D points we calculate the distances between brows and eyes, eyes and mouth, brows and mouth, as well as the width and height of the mouth as depicted in Fig. 3a. In contrast to [20], we measure the closing of the eye by the distance between the upper and the lower eye lid landmark.

Next to these distances, some facial changes are measured from the texture. Based on landmarks we define rectangular regions of interest (see Fig. 3b) and calculate the mean gradient magnitude for each of these regions. This way, we measure the nasolabial folds and the wrinkles at the nasal root and between the eyebrows similar to [20], [26], [27]. The regions are anchored by the landmarks given in Fig. 3b and are placed according to assumptions on the facial geometry derived from empirical investigations of our data. See supplemental material for more details.

Our feature set is designed using domain knowledge. Compared to generic features extraction methods like local binary patterns [28], [30] or Gabor filter banks [14], [15] our approach has two major benefits. First, a lower dimensionality which means less computational effort. Second, the meaning of the features is easier to interpret. This facilitates to draw inferences about the data through feature space analysis, as we do in Sec. 4.1.2 and 4.1.3.

2.2 Feature Signals

The previously described processing steps provide a 6D vector of pose parameters per depth frame and a 13D vector of facial expression features per color frame. For each of these features we consider the corresponding time series obtained from all frames during a time window. For the experiments with the BioVid database we use a time window length of 5.5 s, which is the length of the videos in the dataset.

Fig. 4 depicts the feature signals for an exemplary time window of a painful stimulus. Some of the features are positively correlated with pain, e. g. the nasal wrinkles g_n, some are negatively correlated, e. g. the brow to mouth distance d_{bml}, and for some there is no prior knowledge about correlation, e. g. the head pose parameters. To simplify interpretation and later feature extraction steps, the negatively correlated signals are negated, e. g. $d^*_{bml} = -d_{bml}(t)$, the others are kept unchanged. The comprehensive list of signal definitions is given in the legend of Fig. 4.

Fig. 4 illustrates some benefits of our feature extraction method. First, the frame-level features are easy to interpret, e. g. an increase in d^*_{el} means that the left eye has been further closed. Second, the feature signals facilitate the analysis of dynamics. E. g. it is easy to distinguish blinks and eye closure based on $d^*_{el/r}$ and it seems even possible to spot eye blinks while eyes are nearly closed (see peak at 3.5 s). Further, you can clearly see that the facial expression is composed of two apexes and that the second apex involves much more activity of the corrugator muscle, which shows up in g^*_n and $d^*_{ebl/r}$.

2.3 Signal Descriptors

We condense each of the feature signals x^* in a signal descriptor. First, the signals are temporally smoothed using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

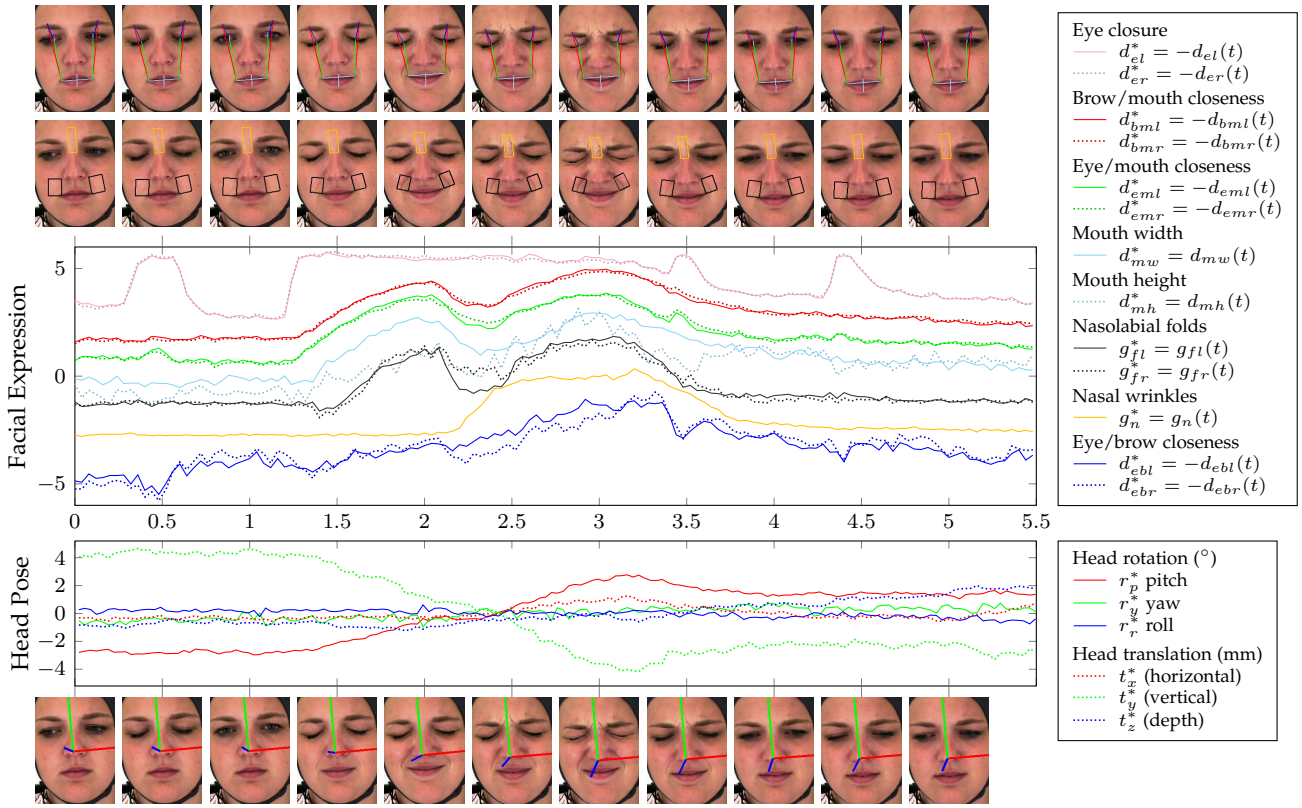


Fig. 4. Exemplary feature signals with corresponding video frames. To improve readability, head pose signals are centered and facial expression signals are standardized and shifted afterwards. Horizontal axis shows time t in seconds.

TABLE 1
Signal descriptor components.

Variable	Description	Domain
`mean`	mean value of signal	value
`median`	median value of signal	
`min`	minimum value of signal	
`max`	maximum value of signal	
`range`	range of signal	value variability
`SD`	standard deviation of signal	
`IQR`	inter-quartile range of signal	
`IDR`	inter-decile range of signal	
`MAD`	median absolute deviation of signal	
`tmax`	instant of time when signal is in its maximum	time
`TGM`	duration the signal is greater than mean	duration
`TGA`	duration the signal is greater than the average of mean and min	
`SGM`	number of segments where the signal is greater than mean	count
`SGA`	number of segments greater than the average of mean and min	
`area`	area between signal and its minimum	value × duration
`areaR`	quotient of `area` and area between `max` and `min`	

a first order Butterworth filter with cutoff 1 Hz. Next, we approximate the first and second temporal derivative of the smoothed signal by difference quotients. In analogy to position s, speed v and acceleration a in kinematics we denote the smoothed feature signal with $s(\cdot)$, its first derivative with $v(\cdot)$ and its second derivative with $a(\cdot)$, respectively. In the following we refer to $s(x^*)$ as state signal, to $v(x^*)$ as speed signal and to $a(x^*)$ as acceleration signal of feature signal x^*.

Subsequently, several features are extracted from each, the state, speed and acceleration signal. The features, which we call descriptor components, are listed in Table 1. They capture different aspects of the signal, such as variability and duration. Some of the variables are motivated by FACS parameters for describing facial actions, e. g., for the state signal `max` and `range` are strongly related to the intensity, `TGM` is related to the duration, and `tmax` corresponds to the time of the apex. Other variables capture additional information like the maximum speed and its timing (`max` and `tmax` of the speed signal). Compared to our preliminary work [20], [26] we use nine additional descriptor components. With 16 descriptor components, we have $3 \cdot 16 = 48$ descriptor variables that are extracted from each feature signal. So if we use the 19 frame-level features defined

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

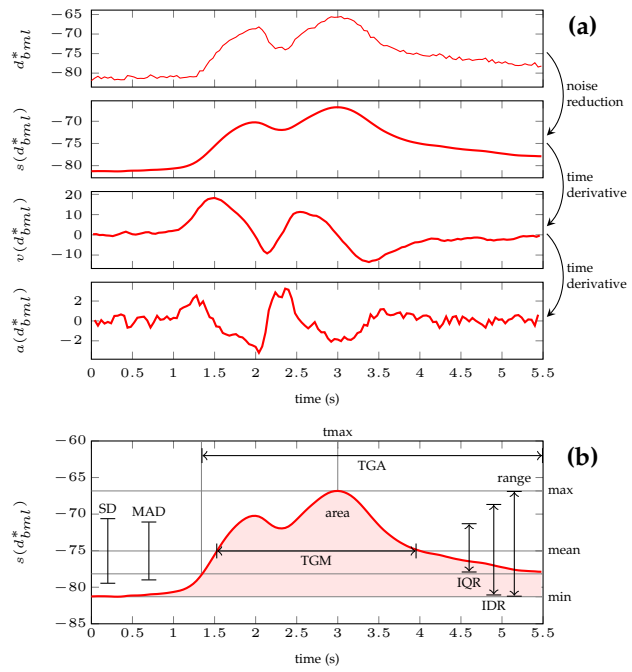


Fig. 5. Signal descriptor extraction example. (a) Raw feature signal (top), smoothed signal (below) and its first and second derivatives (two on bottom). Descriptor components are calculated for each of the latter three signals. (b) Illustration of some descriptor variables.

above, our final feature space has $19 \cdot 48 = 912$ dimensions. Fig. 5 illustrates the descriptor extraction for the feature signal d^*_{bml} as shown in Fig. 4.

3 Classification and Variable Importance

The facial activity descriptors serve as features for predicting the pain intensity. In our experiments we compare several classification methods, namely Random Forest and two variants of Support Vector Machines. This section shortly introduces the classifiers and the method to assess the importance of descriptor variables and feature signals.

3.1 Random Forest Classifier

A Random Forest (RF) classifier [51] is an ensemble of decision trees. It predicts the output for a given test pattern by the majority of the individual trees' decisions. Each tree is constructed on a randomly selected sample of the training set. For each node of a tree, only a randomly selected feature subset is considered to find the optimal split.

In a previous work we successfully applied Random Forests for pain recognition [26]. In follow-up experiments we found that the predictive performance benefits more from increasing the number of trees than from the parameter search done in [26]. For an ensemble we now create 1,000 trees, each with a maximum depth of 10 nodes. The number of features to consider for each split was set to the recommended default, i. e. the square root of the total feature count.

Random Forests facilitate to easily estimate the usefulness of each feature for the given classification task. Basically, the importance measure of a feature—which is usually called variable importance—is given by the decrease in predictive performance that occurs if we randomly permute the values of this feature across samples [51]. Strobl et al. pointed out that this measure is biased towards correlated features, and they proposed a conditional variable importance measure to avoid this bias [52]. Since our descriptors contain several correlated features, we use this conditional variable importance as it is implemented in the software RandomJungle [53] in our experiments.

3.2 Support Vector Machine Classifier

Support Vector Machines (SVM) are the most widely used classifiers in pain recognition. An SVM separates two classes by the hyperplane in the feature space that maximizes the margin between the classes. For non-linear classification the hyperplane is defined in a higher dimensional space, which is implicitly given by a kernel function. In our experiments we use both, the linear kernel and the radial basis kernel. In the following, the SVM with radial basis function (RBF) kernel is denoted as *RBF-SVM*. If not specified explicitly, the SVM is linear. For multi-class problems we used the one-vs-one strategy, i. e. one SVM is trained for each pair of classes; in testing, the final decision is made via voting.

The predictive performance of SVMs depends on the selection of parameters. The error penalty parameter C has to be chosen in all cases. For the RBF-SVM there is an additional kernel parameter γ. To select C and γ we apply a grid search with stratified cross validation [54]. We use 5x2-fold cross validation, as it outperforms the k-fold variant in model selection [55]. Note that this cross validation for parameter selection only uses training data. In our experiments it is nested in an outer cross validation that estimates the generalization performance of the method.

4 Experiments and Discussion

In this section we compare the proposed facial activity descriptors with several other temporal integration methods. Further, we compare the performance of different classifiers and classification problems, analyze the importance of the features, evaluate our descriptor with alternative frame-level features, and discuss the variation in recognition rates regarding different subjects and considered pain intensities.

We conducted experiments with the BioVid Heat Pain Database (see Sec. 4.1) and the UNBC-McMaster database (see Sec. 4.2). Our descriptor-based temporal integration outperforms previous state-of-the-art methods in sequence-based pain classification on both datasets. All reported performances were estimated by leave-one-subject-out cross validation. In this process the parameter selection (see Sec. 3.2) was done individually for each fold only using the respective training data. Sec. 4.3 discusses the Prkaching and Solomon Pain Intensity (PSPI) score, which is widely used as frame-level ground truth.

4.1 BioVid Heat Pain Database

We conduct our main experiments with the BioVid Heat Pain Database (Part A) [21]. Each of the 8,700 video samples is labeled with the respective pain stimulus intensity. So we distinguish five classes: no pain (level 0), low pain (level 1, pain threshold), severe pain (level 4, pain tolerance), and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

two intermediate intensities (level 2 and 3). To maintain comparability, the features are standardized per subject before fed into the classifier as in [26], i. e. the subject-specific mean is subtracted from each feature and the result is divided by the subject-specific standard deviation. Results without this subject-specific feature adaptation are reported in the supplemental material.

4.1.1 Predictive Performance

This section evaluates our proposed facial activity descriptors with different classifiers regarding generalization performance. Further, our technique is compared with other temporal integration methods and our preliminary work. Except for the latter, all methods are validated with the same frame-level features to focus on comparison of the temporal integration methods and classifiers.

The comparison includes the following temporal integration methods.

- Ashraf et al. [22] proposed to learn a frame-level SVM on sequence-level ground truth, i. e. frames are treated as samples and labeled with the sequence's label. A sequence is classified by evaluating the SVM on each frame and by averaging all the distances from the separating hyperplane (*mean score pooling*). To facilitate training with the huge number of samples, the frames are clustered to get 20 samples per sequence (but only for training).
- Similarly, we examine *max score pooling*, which is related to the ideas of Sikka et al. [25] and was used by Prkachin and Solomon in their pain intensity score validation [12]. Instead of averaging the distances to the separating hyperplane, the maximum is chosen and compared with a threshold that has been selected as the optimal point of the Receiver Operating Characteristic for the training data sequences.
- As proposed by Lucey et al. [23], we train a set of frame-level SVMs following the one-vs-all strategy and fit a sigmoid function to estimate the class probability for each SVM. For testing, each frame gets assigned to the class with the highest probability. The majority of frame classes finally decides on the class of the sequences (*mode of classes*). The proposed temporal smoothing worsened the results, so we skipped it. As BioVid is a much larger database than UNBC-McMaster, training was only manageable with clustered training samples (with 20 samples per sequence we still had 174,000 samples in the 5-class problem).
- The methods above drop any dynamics information. In contrast, Bartlett et al. [15] recently proposed an approach with a Bag of Temporal Features (BoTF) descriptor. It condenses filter responses of several feature signals into histogram descriptors. To optimally exploit the idea with the given features, we canceled out the mean of our feature signals. As proposed by Bartlett et al., an SVM with RBF kernel was used for classification at the sequence level. For comparison with our descriptor we also applied the random forest classifier as described in Sec. 3.1.

We optimized the SVM parameters for each method. The results are given in Table 2. The rows list the various temporal integration methods and classifiers; each column contains the accuracies obtained for an individual subset of classes. We show the results of pair-wise classification, as they strongly depend on the considered intensities, which is discussed later. The last column lists the accuracies for the all versus all intensity classification (but some methods can only be applied in the two-class case). We use the accuracy measure to report performances, as it is very intuitive and well suited for data with balanced classes, as the BioVid database. It is defined as the percentage of samples that are classified correctly. Permutation tests [56] (with 50,000 permutations) were applied to show the statistical significance of the improvements, as described in [26].

Generally, higher pain intensity yields better results, as there is more behavioral response. Further, a greater difference between intensities yields better results, as the classes are less similar. With pain level 2 or below the accuracies are hardly above chance, whereas the Random Forest (RF) classifier seems be able to cope better with the high class similarity than other classifiers.

Regarding the temporal integration methods the proposed descriptor performed best in nearly all considered class combinations. In general, the Random Forest is the best performing classifier, followed by SVM with Radial Basis Function kernel (RBF-SVM) and linear SVM. The proposed descriptor outperforms our preliminary works [20], [26] where a subset of the proposed descriptors was used (with some differences in feature signal extraction). When comparing the RFs with 100 and 1,000 trees, we see several significant improvements by changing the RF training (compared to [26]) as described in Sec. 3.1. Further, the proposed descriptor in general also performs better than the Bag of Temporal Features (BoTF) [15], a competitive descriptor for temporal integration, even if our 1,000 trees RF classifier is applied with it. The max pooling idea reaches the highest performance among the methods that are based on frame-level classifiers (and drop dynamics information), but it is also outperformed significantly by the proposed descriptor.

In line with the other results, the pain intensity classification ([all-all]) accuracy is best with the proposed descriptor and Random Forest. It is low (30.8%) but clearly above chance (20%). All results are very far from 100%. This indicates the challenges posed by the complex nature of pain, which we discuss later.

4.1.2 Analysis of Descriptor

To analyze the proposed signal descriptor and the importance of its variables we utilize the Random Forest conditional variable importance (see Sec. 3.1), which was calculated from the whole database with 5,000 trees.

Fig. 6 depicts the calculated variable importance scores, which were obtained with classifiers trained to distinguish highest pain intensity versus no pain ([4-0]). We grouped scores of the same descriptor variable by picking the maximum value among the feature signals. In the rows you find the descriptor variables sorted by their importance.

The most important descriptor variable is the `mean` speed, which encodes the tendency of the feature signals. As most of the signals are positively correlated to pain behavior, their value usually increases during pain. The second is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

TABLE 2
Predictive performance of temporal integration methods and classifiers (rows) on BioVid Heat Pain Database. Accuracy in percent for different classification tasks, i. e. combinations of considered pain intensities (columns).

	Considered classes										
Method / classifier	**[4-0]**	**[4-1]**	**[4-2]**	**[4-3]**	**[3-0]**	**[3-1]**	**[3-2]**	**[2-0]**	**[2-1]**	**[1-0]**	**[all-all]**
Chance					50.0*						20.0*
Frame-level classifier											
SVM + mean score pooling [22]	64.4*	63.3*	61.7*	58.1*	56.7*	55.5*	54.4*	52.9*	50.3*	50.9*	-
SVM + max score pooling	69.0*	68.6*	67.3*	59.5*	58.8*	56.6*	56.8	52.4*	51.2*	50.8*	-
SVM + mode of classes [23]	63.6*	62.6*	61.6*	58.2*	55.9*	53.9*	53.7*	52.0*	50.5*	50.8*	24.3*
BoTF descriptor											
RBF-SVM [15]	71.2*	68.7*	67.6*	57.1*	60.7*	56.2*	54.0*	51.1*	51.6*	49.5*	26.3*
RF (1000 trees)	70.4*	69.1*	68.6	60.6*	61.5*	60.4	**59.1**	52.2*	51.4*	51.9	29.1*
Prelim. work descriptor											
RBF-SVM [20]	66.3*				56.4*			51.7*		50.1*	
RF (100 trees) + better landmarks [26]	71.6*				62.9*			54.4*		51.2*	
Proposed descriptor											
SVM	68.1*	67.4*	69.3	59.4*	58.6*	52.4*	51.2*	50.3*	49.4*	51.0*	27.3*
RBF-SVM	72.1	**71.3**	68.2*	59.3*	62.7	58.1*	56.4*	53.4*	51.0*	51.3*	27.9*
RF (100 trees)	71.6*	70.7	69.4	**63.0**	63.5	60.4*	58.6	54.9*	52.3*	52.3*	29.9*
RF (1000 trees)	**72.4**	70.8	**69.5**	62.9	**64.0**	**61.1**	58.5	**56.0**	**53.4**	**53.3**	**30.8**

* Significantly worse ($p < 0.05$) than RF (1000 trees) on proposed descriptor
SVM: Support Vector Machine (linear) RBF-SVM: SVM with Radial Basis Function kernel BoTF: Bag of Temporal Features
RF: Random Forest

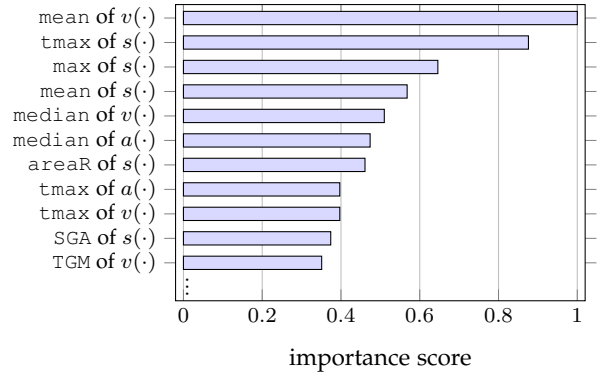


Fig. 6. Importance of descriptor variables. Compare with Fig. 5 and Table 1. For omitted variables see supplemental material.

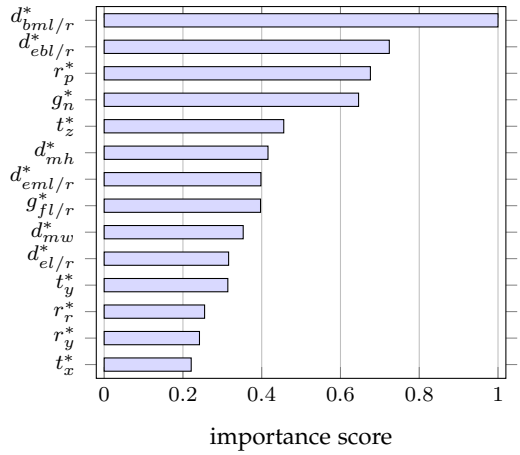


Fig. 7. Importance of feature signals. Compare with Fig. 4.

tmax of the state signal, i. e. the timing of the feature signals' maxima. The maximum of a feature signal usually corresponds to the apex of a facial action. It is reasonable that the temporal relationships of those maxima is valuable for recognition. Among the most important descriptor variables we also find the maximum and mean of the state signal. If there is an facial action, the max corresponds to the feature values during the apex. In contrast, the mean summarizes the whole signal, i. e. it is also influenced by the duration of the apex, speed of the onset etc. The median of the speed and acceleration signals encode the central tendency when excluding outliers (fast movement or high acceleration). They summarize the slower and more constant movements during the time window. Other highly ranked descriptor variables are areaR of the state signal (which estimates the mean intensity relative to the apex intensity if there is a facial action), the tmax of the speed and acceleration signals (which describe temporal relationships between the onsets), as well as SGA of the state and TGM of the speed signals. SGA estimates the number of facial actions if there are any. In the absence of facial actions it tends to be higher, as random noise will lead to more fluctuation in the signal which result in more segments. TGM of the speed estimates the duration of the onset if there is any facial action. Otherwise it tends to be near half of the time window. Further, if we consider symmetric signals like d^*_{bml} and d^*_{bmr}, TGM will be approximately the same in case of a facial action and differ more if there is none. This symmetry property also applies to areaR and several other descriptor variables and is exploited by the learned models.

4.1.3 Importance of Feature Signals

Similarly to the previous section, we analyze the importance of the feature signals (see Fig. 4) for classification with Random Forest conditional variable importance. The results are shown in Fig. 7. Left and right side feature signals are summarized in one score by selecting the maximum score of both. Similarly, each feature signal score is given by the maximum of the corresponding descriptor variable importance scores.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

TABLE 3
Predictive performance comparison of the modalities head pose, facial expression and their combination (accuracies in %).

Features	Considered classes [4-0]	[all-all]
Chance	50.0	20.0
Head pose only	65.5	27.5
Facial expression only	72.1	30.0
All	72.4	30.8

The most important feature signal is the brow-to-mouth closeness $d^*_{bml/r}$. This is reasonable, as this signal captures two important pain related facial actions at once: lowering of the brows and raising of the cheeks. The eye-to-brow closeness $d^*_{ebl/r}$, which measures lowering of the brows, is also rated very high. Compared to these two, the eye-to-mouth closeness $d^*_{eml/r}$ is considered unimportant, probably because it does not add much information if brow-to-mouth and eye-to-brow are already used. In contrast to $d^*_{bml/r}$, the signal $d^*_{eml/r}$ also suffers from blinking artifacts (see Fig. 4), which degrade the feature quality.

The nasal wrinkling g^*_n (in fourth place) is the most useful texture-based measure. The deepening of the nasolabial folds $g^*_{fl/r}$ is rated lower, which is consistent with our observation that out-of-plane rotations disturb this measure (occlusion by nose and overlap with background). Mouth width d^*_{mw} and height d^*_{mh} get a moderate importance rating, which is in line with findings that facial actions like lip pulling, lip stretching, and mouth opening can also be observed during pain [7], [12], [13], but less consistently than other actions. Eye closure is rated low $d^*_{el/r}$. This suggests that eye closure is less specific to pain than other pain related facial actions, which is in line to our observations that some subjects also close the eyes during no-pain videos.

Among the head pose signals, the pitch rotation r^*_p and the distance to the camera t^*_z get high scores. This is consistent with our observation that a person in pain often lowers the head and leans forward. As evident in Table 3, head movement behavior itself is a useful predictor for pain. If we only use head pose descriptors, we obtain accuracies of 65.5% ([4-0]) and 27.5% ([all-all]). Both are significantly above chance and better than some other methods when applied with facial expression features (see Table 2). Compared to facial expression, head pose may yield lower results, but the recognition rate can be improved slightly by combing both. Pain related head poses and movements will be analyzed in detail in one of our follow-up publications.

4.1.4 Comparison of Frame-Level Features

In this section we show that our signal descriptor can be combined with alternative frame-level features. We compare results obtained with different feature types and dimensionalities. Three types of generic features were extracted: 2D landmarks, local binary pattern (LBP) histograms, and 3D distances. LBP and 2D landmarks are widely used in the literature (see Sec. 1.2). Based on the fiducial points from IntraFace, an affine transform was applied to align the 2D landmarks and texture with a mean face to compensate for translation, scale, and in-plane rotation as done in [57] and many other works. From the resulting 180 x 180 pixel image (subdivided by a regular 6 x 6 grid), we extracted uniform LBP yielding 2,124 dimensions on frame level. The aligned fiducial points were used as 2D landmark features directly (98 dimensions). Further, to show that the 3D distance features proposed in Sec. 2.1 are well-selected, we generalized the concept by calculating the 3D distances between *all pairs* of 3D points that we get from projecting the 2D landmarks against our 3D head model (see Sec. 2.1). So with 49 landmarks we get 1,176 facial 3D distances (a superset of the 10 proposed 3D distance features).

TABLE 4
Classification accuracy with different frame-level features (RF classifier with 1000 trees). For all except the last three rows, the dimensionality was reduced through PCA. The feature space for classification has 48 times the dimensions of the frame-level space. The results in the last two rows are significantly better ($p < 0.05$) than all above.

Frame-level features	Considered classes [4-0]	[all-all]
Chance	50.0	20.0
2D landmarks (9 dim.)	69.2	28.9
2D landmarks (15 dim.)	69.5	29.5
2D landmarks (29 dim.)	68.4	29.1
LBP (8 dim.)	68.0	28.8
LBP (15 dim.)	67.8	28.4
LBP (26 dim.)	68.2	28.1
LBP (46 dim.)	67.9	29.3
3D distances (8 dim.)	69.5	28.8
3D distances (12 dim.)	68.8	28.4
3D distances (29 dim.)	68.6	28.5
2D landm.+LBP (15+8 dim.)	69.3	29.4
2D landm.+LBP (15+15 dim.)	69.4	28.9
2D landm.+LBP (15+26 dim.)	69.1	29.0
3D dist.+LBP (8+8 dim.)	69.5	28.7
3D dist.+LBP (8+26 dim.)	68.7	29.0
Proposed texture feat.	68.7	28.9
Proposed 3D distances	72.1	30.3
Proposed in Sec. 2.1 (all)	**72.4**	**30.8**

Compared to the frame level, calculating the signal descriptors increases the feature space dimensionality by factor 48. So we reduce the number of frame-level dimensions for each feature type using Principal Component Analysis (PCA) before extracting the signal descriptors, which is necessary to keep training time in a manageable magnitude. To analyze the influence of dimensionality, we repeat the performance evaluation with varying number of dimensions. For 2D landmarks and 3D distances we keep 90%, 95%, and 99% of the variance; for LBP we keep 40%, 50%, 60%, and 70%. As for the other results we report here on the BioVid database, we apply a subject-specific standardization on the features before they were fed into the classifier. The corresponding cross-validation results are summarized in Table 4. Results without this person-specific adaptation can be found in the supplemental material (they are qualitatively similar).

From Table 4 it is apparent that all generic feature types work well with the proposed descriptor. We get similar results with 2D landmark features, 3D distances, and LBP. However, most LBP results are slightly worse than those of the geometric features. We further tried to combine geometric features with LBP, as combining both improved the results in related works. The combinations yield a slight improvement for the non-subject-standardized results (see supplemental material), but there is no improvement over

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

the geometric features in Table 4. A reason may be that LBP does *not only* encode information that is relevant to the task at hand. Head pose or identity might cause more variance than subtle expressions, so PCA might drop some information on the latter. In general, using more dimensions from PCA does not seem to help. An improvement for future works might be to use supervised methods to reduce dimensionality of generic appearance features. Alternatively, one can exploit prior knowledge, as we did with the hand-engineered frame-level features suggested in Sec. 2.1. Although LBP encodes more information on more parts of the face, the proposed gradient-based texture features (only 3 dimensions) perform similarly (or even slightly better). Such a low-dimensional heuristically chosen feature vector might miss some relevant information (as every feature type), but it also avoids to encode a information that is not relevant to the task at hand. This way, the classifier gets less noise features to cope with and may find a better decision boundary than with high-dimensional features—in much shorter training time. Our proposed set of 3D distances (10 dim.) performs significantly better than each of the generic features ($p < 0.05$ with permutation test, see Sec. 4.1.1). Using all features suggested in Sec. 2.1 improves the results slightly further.

4.1.5 Variability regarding subjects and intensities

Each person has an individual facial geometry and appearance, which is a general problem in facial expression recognition [24], [28], [58]. However, this problem gets even worse if we try to infer a person's internal state, which is usually the ultimate goal of facial expression analysis. Plenty of factors influence the communication of the internal state. Many of them can neither be measured nor controlled easily, are not even fully understood yet and are an area of active research itself.

Pain expression does not directly reflect the pain experience, but is known to be influenced by personal factors and social context [5], [6], [7]. Further, pain experience does not directly reflect the intensity of the noxious stimulus, as it is modulated by genetics, personality, previous pain experiences, drugs and other aspects [5], [6], [7]. I. e. the same stimulus temperature may lead to different pain experiences, as pain sensitivity varies between persons; and the same pain experience may yield different pain expressions, as expressiveness varies.

In the BioVid database, the variation in pain sensitivity was compensated by adjusting the stimulated pain levels according to the subject-specific pain threshold and tolerance. However, the personal factors influencing the pain expression could not be controlled and cause huge differences in expressiveness. Some participants show facial expression for all pain intensities, some for none; some react very intensely, some very subtle. We tried to compensate for subject-related differences in facial expression intensities by subject-specific standardization, as targeted clinical applications focus on comparing intensities for a single patient, not between patients. However, this cannot compensate the lack of facial expression. Thus, recognition rate varies with expressiveness.

Let us consider the random forest classification between highest pain level and no pain (4 versus 0). The performance averages to 72.4%, but for individual subjects it ranges from 100% to less than 40%. A quarter of the subjects have performances better than 90%, but for 9% it is less than chance.

TABLE 5
Classification accuracy with different subject subsets (with RF). FACS based grouping of subjects in those with lower and higher expressiveness (low resp. high PSPI variance) supports the hypothesis that variation in recognition rates and expressiveness are related.

	Considered classes	
Subject subset	**[4-0]**	**[all-all]**
Low PSPI variance	60.7	24.3
High PSPI variance	85.0	36.9
All	72.4	30.8

We conducted an experiment to show the difference in recognition rate between highly and lowly expressive individuals. For this purpose we used the Prkachin and Solomon Pain Intensity (PSPI) [12], [18], a pain expression score based on the facial action coding system (FACS) [10]. It is widely used in facial expression based pain recognition [22], [23], [24], [25], [28], [30], [31], [32]. A certified FACS coder annotated 5% of the BioVid Heat Pain data used for classification (1 sequence per intensity and subject = 435 videos = 60,030 frames) with action unit intensities. Next, we selected the maximum PSPI for each video (5 per person) and calculated the variance of these PSPI scores across the videos for each person to estimate the expressiveness. The subjects were split into two groups: these with expressiveness score above median (PSPI with high variance) and those below median (PSPI with low variance). On each of the both subsets, we evaluated the proposed method with Leave-One-Subject-Out Cross Validation. Table 5 lists the results. The performances differ highly significantly between the more and less expressive subject sets. A two sample t-test of the null hypothesis that performance distributions have equal means yields $p \approx 10^{-10}$ for [4-0] and $p \approx 10^{-9}$ for [all-all]. Even though expressiveness is estimated with a very simple criterion and only few samples, the experiment shows that our approach performs much better on more expressive subjects. Thus, large parts of the performance gap to the desired 100% is not caused by technical weaknesses of the recognition system, but by challenges posed due to the complex nature of pain and its expression.

Recognition rates vary significantly with the considered stimulus intensities, as shown in Table 2. Fig. 8 plots the maximum PSPI scores for each of the 87 subjects as a function of the pain stimulus intensity. Each of the thin, colored lines connects the PSPI scores from one subject. We added some jitter to improve visibility of single subject graphs and to give an idea of the number of samples through the density of lines. The figure also shows the mean PSPI scores across all subjects (thick black line). The mean increases with pain intensity, which is consistent to previous findings [12]. Some authors conclude that PSPI is a highly reliable and valid measure of pain. But a closer look relativizes this conclusion, as the picture is much more diverse when considering the observations underlying the mean.

According to PSPI, no facial pain reaction could be observed in about 72% of the samples at the pain threshold

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

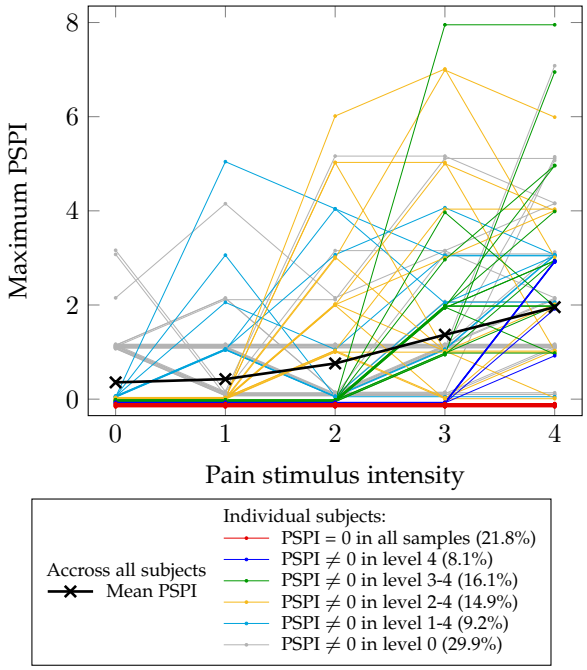


Fig. 8. Prkachin and Solomon Pain Intensity (PSPI) score across the pain intensities (1-4) and no pain (0). The mean increases with intensity, but the image is more diverse for individual subjects. Subjects are grouped as given in the legend (with percentage of subjects belonging to each group). Best viewed in color.

(level 1), in 67% of pain level 2, in 41% of pain level 3, and in 28% of the samples at pain level 4. These numbers are consistent with what we observe viewing the pain response videos. The lower the pain level, the more often there is no facial response at all. This explains the decrease of recognition rates with the considered pain intensity (see Table 2).

Among 87 trials is not a single subject sampling in which PSPI increases strictly with the stimulus level. Further, for many subjects PSPI is zero in a wide range of painful intensities. According to this, 22% of the subjects show no facial expression of pain across all intensities (red in Fig. 8). This is in line with other experimental studies which report a similar percentage of subjects that displayed no facial response to pain [5], [13]. A lack of facial activity was also observed for several subjects with permanently closed eyes (constant PSPI of one). For 8% of the subjects there is only facial response for pain level 4 (blue lines), 16% start to respond at level 3 (green), 15% at level 2 (yellow) and only 9% at the lowest pain level. These observations conform to the pain communication model by Prkachin and Craig [5], which involves that a pain experience must exceed a subject-specific severity to trigger facial response (also see [59]).

In conclusion, for several people it might be impossible to measure low pain intensities from facial expression only. Multi-modal pain recognition or person-specific adaptation can help to improve the results [26], but due to less pain response, recognition rate will always be lower for lower pain intensities. Fortunately, high pain intensities are more relevant in practice.

4.2 UNBC-McMaster pain database

In this section we apply the proposed signal descriptor on the UNBC-McMaster Shoulder Pain Expression Archive database [18]. The frame-level features are varied, similar to Sec. 4.1.4. We train and test classifiers to predict the observer-assessed pain intensity (OPI), i. e. to imitate an expert human observer. This is less challenging than predicting the stimulus (see Sec. 4.1), since it avoids the problem of differences in expressiveness. With OPI, a person that feels pain without showing it, is labeled with "no pain", as an observer bases his pain assessment on visible reactions only.

The database provides 66 tracked landmarks, which we use to calculate the similarity normalized shape features (SPTS) as done by Lucey et al. [23] and others. To reduce the dimensionality of these features, we apply Principal Component Analysis (PCA). We keep 99% of the variance, which reduces the frame-level feature space from 132 to 29 dimensions. For each frame-level feature in the transformed space, we extract the corresponding time series from the video sequence and calculate the proposed signal descriptor (see Sec. 2.3). Finally, the resulting signal descriptors are classified by a linear SVM. We use the same classifier as Lucey et al. [23] to focus the comparison on the temporal integration aspect.

Further, we extract the canonical normalized appearance features (CAPP) as suggested by Lucey et al. [23] and others. I. e., face appearance is normalized through piece-wise affine warping, which registers the 66 landmarks with the mean shape, and the gray-scale pixel intensities of the resulting image are used as features. We apply PCA and keep 99% of the variance, which reduces frame-level feature space from about 8,100 to 400 dimensions. Signal descriptors are extracted and classified as for SPTS. We also combine the SPTS and CAPP descriptors, as combining shape and appearance features often improves the performance.

In Table 6 we compare our results with those of Lucey et al. [23] and Sikka et al. [25]. Lucey et al. grouped the samples in three classes (OPI 0-1, 2-3 and 4-5); Sikka et al. used two classes (OPI 0 and 3-5). We conducted the experiment with both variants and different frame-level features. Next to the accuracy (number of correctly classified samples divided by total sample count) we also list the macro-averaged F1 measure [60]. It weights all classes equally, which is desirable in this context, since the sample counts are not balanced, but all classes are equally important.

The proposed descriptor outperforms the other methods due to its more sophisticated temporal integration. We use the same classifier as Lucey et al. and get better results, even with a subset of their frame-level features. Sikka et al. report an experiment in which SVM is outperformed by the MS-MIL classifier [25]. Nevertheless, our descriptor outperforms the MS-MIL approach although we use a simple linear SVM, which indicates the superiority of our temporal integration.

When comparing different frame-level features we observe that our descriptor performs better with shape that with appearance features. The reason may be the curse of dimensionality [61]. Generic appearance features are usually high dimensional. With our descriptor, each of the frame-level features adds 48 dimensions to the final feature space. So 400 dimensions of CAPP yield 19,200 descriptor fea-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2537327, IEEE Transactions on Affective Computing

TABLE 6
Predictive performance on the UNBC-McMaster Shoulder Pain Expression database.

Method	Frame-level features	Classifier	2-class problem Accuracy	2-class problem Macro-F1	3-class problem Accuracy	3-class problem Macro-F1
Lucey et al. [23]	SPTS+CAPP	SVM	-	-	0.610+	0.536+
Sikka et al. [25]	Dense SIFT BoW	MS-MIL	0.837	-	-	-
Proposed descriptor	SPTS (PCA, 29 dim.)	SVM	0.890	0.881	0.740	0.597
	CAPP (PCA, 400 dim.)		0.855	0.843	0.685	0.458
	SPTS+CAPP (PCA, 29+400 dim.)		0.862	0.852	0.725	0.490
	SPTS+CAPP (PCA, 29+10 dim.)		**0.917**	**0.911**	**0.755**	**0.600**

+ Calculated from published confusion matrix and sample counts.

tures, whereas there are only 200 samples for the sequence-level classification task, which might be too few to spot the relevant features among the noisy ones. As we see in Table 6, combining full SPTS and CAPP descriptors does not improve performance compared to SPTS. However, if we reduce the frame-level dimensionality of CAPP to the 10 PCA scores with the highest variance (instead of 400), the results also outperform SPTS.

To the best of our knowledge, our method defines new state-of-the-art in sequence-based pain recognition on the UNBC-McMaster database. The experiment also shows again that our signal descriptor can be successfully combined with other than the frame-level features proposed in Sec. 2 (also see Sec. 4.1.4). Further, it shows that the descriptor can successfully handle varying time window lengths, as the whole video was used as the time window.

We also saw that the number of frame-level features is an important factor for applying our descriptor. It is beneficial to use dimension reduction techniques at frame level (as done here and in Sec. 4.1.4) or to use prior knowledge (as done in Sec. 2.1) to get a low-dimensional frame-level representation, as this avoids the curse of dimensionality.

4.3 Comments on the Prkaching and Solomon Pain Intensity (PSPI) score

This section discusses the PSPI score, which is widely used as ground truth to approach pain recognition from frame-level. A more detailed version of this section (with more examples and images from 3 databases) can be found in the supplemental material. PSPI was proposed by Prkaching and Solomon [12], [18] and validated based on video-level ground truth and a maximum pooling temporal integration of video frames. The statistical analysis in their study was done on sequence level; so the conclusion that PSPI is valid and reliable *on frame-level* is not fully justified.

The PSPI of a single frame should not be confounded with the *feeling* of pain at this particular moment, as it only measures the facial expression of pain. Although both are strongly related, they should be distinguished carefully, as they are connected through a complex time-dependent process with a lot of influencing factors.

The PSPI may be zero although the person actually feels pain. There may be no facial reaction at all due to too low pain intensity and/or low pain expressiveness [5], [13], [20], [59]. Further, the feeling of pain may induce a facial response that is not part of the prototypic pattern underlying PSPI (AU4/6/7/9/10/43). A recent study by Kunz et al. [13] suggested that there are several "faces of pain". They found facial activity patterns that include the raising of eyebrows (AU1/2) or opening of the mouth (AU25/26/27), which are both not considered by PSPI. It is sometimes argued that PSPI offers a high temporal resolution, as it provides an independent score for each frame. This is true, but we should keep in mind that PSPI does not directly reflect the feeling of pain. As you can see in the supplemental material, PSPI can go up and down (with tension and relaxation in the facial expression) although the felt pain is steadily increasing. So if we are interested in the feeling of pain, the temporal resolution of PSPI might be misleading, especially if the pain persists for longer time.

The PSPI may be also non-zero although the observed subject does not feel pain. Most obvious, AU43 (closed eyes) is not specific to pain, e. g. it also occurs during sleep and relaxation. Further, several facial expressions of emotions share AUs with PSPI [62], [63], e. g. disgust (AU9 or 10), fear (AU4), sadness (AU4), or happiness (AU6). If PSPI is used in a wider context, many frames are labeled as painful by mistake, which could be easily avoided by using sequence-level ground truth.

As illustrated above, there are several shortcomings with PSPI. More research is needed to find better measures of pain or augment the knowledge on the existing measures, also regarding dynamics. Probably, a better frame-level measure can be constructed, e. g. by subtracting AUs that do not occur during pain and by considering multiple "faces of pain" in a non-linear combination. However, we believe that sequence-level ground truth is more promising. Self-report is the gold standard of pain assessment and can be acquired easily as ground truth. Sequence-level observer pain ratings (as provided in the UNBC-McMaster database) are also an option. Further, in experimental pain studies we have the opportunity to get ground truth with higher temporal resolution from the applied pain stimulation, e. g. the time-series of stimulation temperatures in a heat pain experiment. But as already shown in the literature, even ground truth with coarse temporal resolution can be exploited to learn a models for frame-level prediction [22], [25], [35]. So temporal resolution is no plausible argument against using sequence-level ground truth. But higher reliability and validity argue for it. That is why we suggest to build systems on sequence-level ground truth and further explore temporal integration methods. This also offers the opportunity to exploit facial expression aspects that go beyond PSPI and discover new knowledge about pain.

5 Conclusion

In this article we proposed a new feature set for describing facial actions and their dynamics, which we call facial activity descriptors. Each of these descriptors condenses a feature signal—a time series of a frame-level feature—into a signal descriptor by extracting features from the time series and its first and second derivative. We conducted experiments in pain detection and intensity estimation in which the proposed descriptors outperformed previous state-of-the-art methods in sequence-level classification on both, the BioVid Heat Pain and the UNBC-McMaster Shoulder Pain Expression database. The results highlight the relevance of a proper temporal integration method, which offers the opportunity to exploit information that goes beyond prior knowledge used to create expensive frame-level ground truth.

We also discussed challenges such as inter-individual differences in expressiveness or the lack of behavioral response to pain of low intensities. Inter-individual differences may be approached by systems that adapt to the individual, e. g. via transfer learning, or by broadening knowledge about the influential factors. The lack of behavioral response possibly can be compensated by biomedical information [26]. Another promising research direction is to look for temporal integration methods that can get as much as possible out of sequence-level ground truth, which can be acquired easily with various established methods. Finally, we should improve the systems' ability to cope with extreme head poses, partial occlusions and bad lighting, as this will be needed to take the step to clinical practice.

Acknowledgments

This work was funded by the German Research Foundation (DFG), project AL 638/3-1 and project AL 638/3-2.

References

[1] D. C. Turk and R. Melzack, *Handbook of pain assessment*. Guilford Press, 2011, ch. Preface.

[2] M. E. Lynch, K. D. Craig, and P. W. H. Peng, *Clinical Pain Management: A Practical Guide*. John Wiley & Sons, 2011.

[3] H. McQuay, A. Moore, and D. Justins, "Treating acute pain in hospital." *BMJ: British Medical Journal*, vol. 314, no. 7093, p. 1531, 1997.

[4] H. Merskey, "Editorial: The need of a taxonomy," *PAIN*, vol. 6, no. 3, pp. 247–252, 1979.

[5] K. M. Prkachin and K. D. Craig, "Expressing pain: The communication and interpretation of facial pain signals," *J Nonverbal Behav*, vol. 19, no. 4, pp. 191–205, 1995.

[6] K. D. Craig, "The social communication model of pain." *Canadian Psychology/Psychologie canadienne*, vol. 50, no. 1, p. 22, 2009.

[7] K. D. Craig, K. M. Prkachin, and R. E. Grunau, "The facial expression of pain," in *Handbook of Pain Assessment*, D. C. Turk and R. Melzack, Eds. Guilford Press, 2011.

[8] S. M. G. Zwakhalen, J. P. H. Hamers, H. H. Abu-Saad, and M. P. F. Berger, "Pain in elderly people with severe dementia: A systematic review of behavioural pain assessment tools," *BMC Geriatrics*, vol. 6, no. 1, p. 3, 2006.

[9] K. D. Craig, "The facial expression of pain Better than a thousand words?" *APS Journal*, vol. 1, no. 3, pp. 153–162, 1992.

[10] P. Ekman and W. V. Friesen, *Facial Action Coding System: A technique for the measurement of facial movements*. Palo Alto: Consulting Psychologist Press, 1978.

[11] K. M. Prkachin, "The consistency of facial expressions of pain: a comparison across modalities," *Pain*, vol. 51, no. 3, pp. 297–306, 1992.

[12] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *PAIN*, vol. 139, no. 2, pp. 267–274, 2008.

[13] M. Kunz and S. Lautenbacher, "The faces of pain: a cluster analysis of individual differences in facial activity patterns of pain," *Eur J Pain*, vol. 18, no. 6, pp. 813–823, 2014.

[14] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797–1803, 2009.

[15] M. Bartlett, G. Littlewort, M. Frank, and K. Lee, "Automatic Decoding of Facial Movements Reveals Deceptive Pain Expressions," *Current Biology*, vol. 24, no. 7, pp. 738–743, 2014.

[16] P. Werner, A. Al-Hamadi, and R. Niese, "Comparative learning applied to intensity rating of facial expressions of pain," *Int. J. Patt. Recogn. Artif. Intell.*, vol. 28, no. 05, p. 1451008, 2014.

[17] F. D. l. Torre and J. F. Cohn, "Facial Expression Analysis," in *Visual Analysis of Humans*, T. B. Moeslund, A. Hilton, V. Krger, and L. Sigal, Eds. Springer London, 2011, pp. 377–409.

[18] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *IEEE Int'l Conf. on Automatic Face & Gesture Recognition and Workshops (FG)*, 2011, pp. 57–64.

[19] S. Walter, P. Werner, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, A. Al-Hamadi, A. O. Andrade, G. Moreira da Silva, and S. Crawcour, "The BioVid Heat Pain Database: Data for the Advancement and Systematic Validation of an Automated Pain Recognition System," in *Cybernetics (CYBCONF), IEEE Int'l Conf. on*, 2013, pp. 128–131.

[20] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges," in *British Machine Vision Conf.* BMVA, 2013, pp. 119.1–119.13.

[21] BioVid Heat Pain Database website. [Online]. Available: http://www-e.ovgu.de/biovid/

[22] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face - Pain expression recognition using active appearance models," *Image and Vision Computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[23] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews, "Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database," *Image and Vision Computing*, vol. 30, no. 3, pp. 197–205, 2012.

[24] J. Chen, X. Liu, P. Tu, and A. Aragones, "Person-specific expression recognition with transfer learning," in *Image Processing (ICIP), IEEE Int'l Conf. on*, 2012, pp. 2621–2624.

[25] K. Sikka, A. Dhall, and M. S. Bartlett, "Classification and weakly supervised pain localization using multiple segment representation," *Image and Vision Computing*, vol. 32, no. 10, pp. 659–670, 2014.

[26] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic Pain Recognition from Video and Biomedical Signals," in *Pattern Recognition, Int'l Conf. on*, 2014, pp. 4582–4587.

[27] P. Werner, A. Al-Hamadi, and R. Niese, "Pain Recognition and Intensity Rating based on Comparative Learning," in *Image Processing (ICIP), IEEE Int'l Conf. on*, 2012, pp. 2313–2316.

[28] O. Rudovic, V. Pavlovic, and M. Pantic, "Automatic Pain Intensity Estimation with Heteroscedastic Conditional Ordinal Random Fields," in *Advances in Visual Computing - 9th Int'l Symposium, ISVC, Greece. Proceedings, Part II*, ser. LNCS, vol. 8034. Springer, 2013, pp. 234–243.

[29] ——, "Context-Sensitive Dynamic Ordinal Regression for Intensity Estimation of Facial Action Units," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 944–958, 2015.

[30] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous Pain Intensity Estimation from Facial Expressions," in *Advances in Visual Computing*, ser. LNCS. Springer, 2012, no. 7432, pp. 368–377.

[31] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in *Proceedings of the 14th ACM Int'l Conf. on Multimodal Interaction*. New York: ACM, 2012, pp. 47–52.

[32] C. Florea, L. Florea, and C. Vertan, "Learning Pain from Emotion: Transferred HoT Data Representation for Pain Intensity Estimation," in *ECCV workshop on ACVR*, Zurich, 2014.

[33] R. Niese, A. Al-Hamadi, A. Panning, D. Brammen, U. Ebmeyer, and B. Michaelis, "Towards Pain Recognition in Post-Operative Phases Using 3d-based Features From Video and Support Vector Machines," *JDCTA*, vol. 3, no. 4, pp. 21–33, 2009.

[34] Z. Hammal and M. Kunz, "Pain monitoring: A dynamic and context-sensitive system," *Pattern Recognition*, vol. 45, no. 4, pp. 1265–1280, 2012.

[35] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal Data Fusion for Person-Independent, Continuous Estimation of Pain Intensity," in *Engineering Applications of Neural Networks*, ser. Communications in Computer and Information Science, L. Iliadis and C. Jayne, Eds. Springer, 2015, no. 517, pp. 275–285.

[36] M. Kächele, P. Werner, A. Al-Hamadi, G. Palm, S. Walter, and F. Schwenker, "Bio-Visual Fusion for Person-Independent Recognition of Pain Intensity," in *Multiple Classifier Systems*, ser. LNCS, F. Schwenker, F. Roli, and J. Kittler, Eds. Springer, 2015, no. 9132, pp. 220–230.

[37] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang, "Automated Assessment of Children's Postoperative Pain Using Computer Vision," *Pediatrics*, pp. peds.2015–0029, May 2015.

[38] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H. C. Traue, P. Werner, A. Al-Hamadi, N. Diniz, G. M. da Silva, and A. O. Andrade, "Automatic pain quantification using autonomic parameters," *Psychology & Neuroscience*, vol. 7, no. 3, pp. 363 – 380, 2014.

[39] R. Treister, M. Kliger, G. Zuckerman, I. G. Aryeh, and E. Eisenberg, "Differentiating between heat pain intensities: The combined effect of multiple autonomic parameters," *Pain*, vol. 153, no. 9, pp. 1807–1814, 2012.

[40] N. Ben-Israel, M. Kliger, G. Zuckerman, Y. Katz, and R. Edry, "Monitoring the nociception level: a multi-parameter approach," *J Clin Monit Comput*, vol. 27, no. 6, pp. 659–668, 2013.

[41] S. Gruss, R. Treister, P. Werner, H. C. Traue, S. Crawcour, A. Andrade, and S. Walter, "Pain Intensity Recognition Rates via Biopotential Feature Patterns with Support Vector Machines," *PLoS ONE*, vol. 10, no. 10, p. e0140330, 2015.

[42] E. G. Krumhuber, A. Kappas, and A. S. Manstead, "Effects of dynamic aspects of facial expressions: a review," *Emotion Review*, vol. 5, no. 1, pp. 41–46, 2013.

[43] M. G. Frank, P. Ekman, and W. V. Friesen, "Behavioral markers and recognizability of the smile of enjoyment." *Journal of personality and social psychology*, vol. 64, no. 1, p. 83, 1993.

[44] E. G. Krumhuber and A. S. R. Manstead, "Can Duchenne smiles be feigned? New evidence on felt and false smiles," *Emotion*, vol. 9, no. 6, pp. 807–820, 2009.

[45] X. Xiong and F. De la Torre, "Supervised Descent Method and its Applications to Face Alignment," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conf. on*, 2013, pp. 532–539.

[46] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection," in *DAGM 25th Pattern Recognition Symposium*, 2003, pp. 297–304.

[47] R. Niese, P. Werner, and Ayoub Al-Hamadi, "Accurate, Fast and Robust Realtime Face Pose Estimation Using Kinect Camera," in *IEEE Int'l Conf. on Systems, Man, and Cybernetics*, 2013, pp. 487–490.

[48] K. D. Craig, S. A. Hyde, and C. J. Patrick, "Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain," *Pain*, vol. 46, no. 2, pp. 161–171, 1991.

[49] M. Kunz, S. Scharmann, U. Hemmeter, K. Schepelmann, and S. Lautenbacher, "The facial expression of pain in patients with dementia," *Pain*, vol. 133, no. 1-3, pp. 221–228, 2007.

[50] R. Niese, A. Al-Hamadi, A. Panning, and B. Michaelis, "Emotion Recognition based on 2d-3d Facial Feature Extraction from Color Image Sequences," *Journal of Multimedia*, vol. 5, pp. 488–500, 2010.

[51] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[52] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[53] D. F. Schwarz, I. R. Knig, and A. Ziegler, "On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data," *Bioinformatics*, vol. 26, no. 14, pp. 1752–1758, 2010.

[54] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003.

[55] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.

[56] D. S. Moore, G. P. McCabe, W. M. Duckworth, and S. L. Sclove, *The practice of business statistics: using data for decisions*. Wh Freeman, 2003.

[57] P. Werner, F. Saxen, and A. Al-Hamadi, "Handling Data Imbalance in Automatic Facial Action Intensity Estimation," in *British Machine Vision Conf. (BMVC)*. BMVA, 2015, pp. 124.1–124.12.

[58] W.-S. Chu, F. D. L. Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conf. on*. IEEE, 2013, pp. 3515–3522.

[59] M. Kunz, V. Mylius, K. Schepelmann, and S. Lautenbacher, "On the relationship between self-report and facial expression of pain," *The Journal of Pain*, vol. 5, no. 7, pp. 368–376, 2004.

[60] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[61] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[62] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4d-Spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[63] D. Matsumoto and P. Ekman, "Facial expression analysis," *Scholarpedia*, vol. 3, no. 5, p. 4237, 2008. [Online]. Available: http://www.scholarpedia.org/article/Facial_expression_analysis

Philipp Werner received his Masters degree (Dipl.-Ing.-Inf.) in computer science from the Otto-von-Guericke University Magdeburg, Germany, in 2011. Since then he has been working as a research assistant in the Neuro-Information Technology group. He is a PhD candidate and his current research focuses on automatic pain recognition, pattern recognition, and machine vision.

Prof. Dr. Ayoub Al-Hamadi received his PhD at the Otto-von-Guericke-University of Magdeburg, Germany in 2001. In 2003 he became Junior-Research-Group-Leader and in 2010 he received the Habilitation in Artificial Intelligence and the Venia Legendi in Pattern Recognition and Image Processing. He currently is an adjunct professor and Head of the Neuro-Information Technology group in the University of Magdeburg.

Dr. Kerstin Limbrecht-Ecklundt is a researcher and psychotherapist at the University of Hamburg, with a focus on chronic pain, pain management, acute pain, and placebo reseearch. She received her diploma at the CAU Kiel in 2009 and her PhD at the University of Ulm in 2012.

Dr. Steffen Walter is a researcher in the Medical Psychology at the University of Ulm, Germany. He received his PhD in human biology from the University of Ulm in 2008. His research focuses on Automated Pain Recognition, Affective Computing, and Psychotherapy Processes.

Dr. Sascha Gruss received his PhD from Department of Medical Psychology at the University of Ulm, Germany, in 2015. His research interests include affective and physiological computing, human-computer interaction, machine learning, pattern recognition, and pain stimulation.

Prof. Dr. Harald C. Traue is Head of Medical Psychology at Ulm University. Graduated as Computer Scientist he got his PhD in Human Biology. He was visiting and adjunct professor at the University of Calgary. His research areas are Cognitive Sciences and Emotions, Affective Computing in HCI, and Behavioural Medicine.